Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Právní aspekty budování a provozu datového skladu

DIPLOMOVÁ PRÁCE

Studijní program: Aplikovaná informatika

Studijní obor: Podniková informatika

Autor: Bc. Jan Janoušek

Vedoucí diplomové práce: Ing. Oleg Svatoš, Ph.D.

Praha, červen 2019

Prohlášení

Prohlašuji, že jsem diplomovou práci *Právní aspekty budování a provozu datového skladu* vypracoval samostatně za použití v práci uvedených pramenů a literatury.

V Praze dne 24. června 2019

Jan Janoušek

Poděkování

Rád bych na tomto místě poděkoval vedoucímu této diplomové práce za fundovaný a vstřícný přístup, věcné náměty a celkovou podporu, díky které jsem měl možnost naplnit stanovené cíle a své představy o přínosu této práce.

Jan Janoušek

Abstrakt

Tato diplomová práce se zaměřuje na právní aspekty budování datových skladů, konkrétně na ochranu osobních údajů. Ta je relativně novým tématem, které je v praxi v současné době intenzivně řešeno. Ačkoliv je ochrana osobních údajů zakotvena v legislativě již dlouhá desetiletí, masově začala být řešena teprve v roce 2018 v souvislosti se zavedením GDPR. Vysoké sankce za porušení ochrany osobních údajů vyvolaly zvýšený zájem veřejnosti.

Autor této práce v praxi identifikoval nedostatky v řešení této problematiky na úrovni datových skladů a rozhodl se využít této příležitosti a přispět k řešení ochrany osobních údajů v rámci datových skladů vlastními konkrétními návrhy.

Práce zahrnuje teoretický vhled do zpracování dat na úrovni informačních systémů, přehled legislativy GDPR v pojetí odpovídajícímu účelu této práce a dále vlastní návrhy úpravy procesů implementace a provozu datových skladů. V závěru práce autor demonstruje své návrhy na konkrétním praktickém příkladu.

Práce jako celek přispívá k osvětě v oblasti ochrany osobních údajů a vysvětluje základní principy a zásady legislativy GDPR tak, aby usnadnila její implementaci nejen na úrovni datového skladu.

Klíčová slova

Datový sklad, DWH, ochrana osobních údajů, obecné nařízení o ochraně osobních údajů, GDPR.

JEL klasifikace

K24 Cyber Law, M15 IT Management

Abstract

This diploma thesis is focused on an aspect of the development of data warehouses, more specifically on personal data protection. This topic is a relatively new one currently. Although personal data protection has been anchored in law for many, the 2018 launch of the GDPR brought it to light. High fines especially stimulated public interest.

The author of this thesis has identified shortcomings in data protection in data warehouse area and offer it´s solution.

This thesis includes a theoretical insight into data processing at the information systems level and the overview of GDPR law relevant to this thesis. The author argues his case using a practical case study at the end of the thesis.

The thesis makes a contribution to the field of data protection and explains principles of the GDPR law to make an implementation easier.

Keywords

Data warehouse, DWH, personal data protection, general data protection regulation, GDPR.

JEL Classification

K24 Cyber Law, M15 IT Management

Obsah

Úvod

Tato diplomová práce vznikla jako vyvrcholení autorova studia v oboru *Podniková informatika* na VŠE v Praze. Toto studium, a právě na této škole, si autor zvolil pro svůj zájem o byznys pohled na IT (Informační technologie). Výstupem studia je manažerský přístup k IT a nikoliv podrobná technologická znalost. V souladu s koncepcí studia vznikla i tato práce reflektující autorův konceptuální přístup k IT.

Téma práce

Zmíněnému přístupu odpovídá téma práce. Tím jsou *Právní aspekty budování a provozu datového skladu*. Téma navazuje na autorův osobní zájem o problematiku ochrany osobních údajů a etiku při zpracování dat. Aktuální dění v oblasti ochrany osobních údajů vnímá autor jako příležitost pro byznys inovace, nikoliv jako ohrožení byznysu a administrativní přítěž. V praxi je však legislativa vnímána spíše negativně, a proto jí není věnována dostatečná pozornost, respektive chuť se jí zabývat ze strany podniků či podnikatelů.

Cíle práce

Hlavním cílem této práce je přinést **obecný návrh** řešení problematiky **ochrany osobních údajů** na úrovni **datového skladu** a to jak ve fázi **implementace** datového skladu, tak i během jeho následného **provozu**.

Pro dosažení hlavního cíle práce si autor vytyčil **cíle vedlejší**, které pomohou celé řešení sestavit. Těmi jsou:

- představit **teoretické koncepce** zpracování dat a tvorby DWH,
- vytvořit základní **přehled legislativy** ve vazbě na DWH,
- navrhnout **koncept propojení** legislativního rámce s konkrétním technickým řešením implementace a provozu DWH,
- demonstrovat navržený postup ve formě **případové studie**.

Autor přináší **principy a přístupy** k řešené problematice tak, aby je bylo možno aplikovat v praxi. Jeho záměrem je **přispět k osvětě** v oblasti ochrany osobních údajů.

Použité metody

Závěrečnou kapitolu této práce tvoří případová studie, jejímž cílem je demonstrovat navržené řešení na konkrétním praktickém případě. Forma případové studie byla zvolena jako nejvhodnější metoda pro ověření stanovených hypotéz, protože je postavena na přímém vyzkoušení návrhů na praktickém příkladu. Oproti tomu by často používaná metodika průzkumů vycházela pouze z pocitů a představ dotazovaných a nepřinesla by opravdový důkaz o praktické použitelnosti návrhů.

Postup zpracování případové studie vychází z (Yin, 2009). Celá práce je pojata jako obecně platné řešení, které není specifické jen pro konkrétní podnik či konkrétní situaci. Případová studie pak není omezením právě na jeden případ, nýbrž demonstruje aplikaci obecného postupu na příkladu konkrétního podniku.

Předpoklady a omezení

Autor se snaží přijít zejména s principy přístupu k řešené problematice a snaží se o obecnější návrh, z něhož může praxe dále vyjít pro detailní technické řešení své konkrétní situace. Práce je zaměřena tak, aby poskytla obecně platné řešení.

Očekávaný rozsah této práce neumožňuje řešit všechny právní aspekty budování datového skladu. S ohledem na aktuálnost tématu se autor omezil na téma ochrany osobních údajů. A dále i v rámci tohoto zúžení se autor věnuje primárně pohledu komerční praxe. V problematice ochrany osobních údajů lze najít řadu nuancí např. ve vztahu k činnosti Policie ČR, soudů, orgánů státní správy apod. K těmto specifikám práce nepřihlíží. Snaha o detailní zachycení všech výjimek a specifik by vedla ke snížení přehlednosti celého textu, aniž by přinesla odpovídající přidanou hodnotu. Na text této práce je tedy třeba nahlížet optikou běžného podnikatelského subjektu.

V práci je využita technologie relačních databází. Není to jediné možné řešení datového skladu, nicméně odlišování této úrovně technického detailu by opět vedlo ke snížení přehlednosti práce jako celku.

Rešerše informačních zdrojů

Při zpracování této práce vycházel autor ze zjištění, že je zkoumaná problematika v praxi málo rozšířená a není tedy nijak významně podchycena ani v podobě informačních zdrojů. Ačkoliv samotná ochrana osobních údajů je předmětem legislativy již desítky let, stala se středem zájmu praxe až počátkem roku 2018 v souvislosti se zavedením specializované legislativy EU v podobě GDPR (General Data Protection Regulation, Obecné nařízení o ochraně osobních údajů).

Na VŠE byla na téma ochrany osobních údajů v kontextu datových skladů v minulosti zpracována diplomová práce (Tůma, 2017). Podle zkoumání autora této práce však tehdy

navržené řešení úplně nepokrývá potřeby vyplývající z legislativy, což odpovídá i tomu, že předmětná diplomová práce byla zpracovávána před vstupem legislativy v platnost. Její autor tak neměl možnost podchytit bouřlivé diskuse praxe a realitu po zavedení legislativy. Tyto informace již autor této práce má a navíc může zohlednit i reálné nasazení v praxi, neboť v současnosti již uběhl jeden rok od zavedení GDPR. Tato práce tak má ambici navázat na předešlé zmíněné dílo a rozšířit jej.

S ohledem na to, že je GDPR vlastně novinkou, zejména pokud jde o zaměření praxe, nenašel autor informační zdroje uceleně řešící právě kontext ochrany osobních údajů a datových skladů. Aby měl autor jistotu, že bude rešerše kompletní, využil konzultací a rešeršních služeb v rámci VŠE v Praze a Národní technické knihovny tamtéž. Z těchto zdrojů vyplynulo, že problematiku uceleně neřeší žádá literatura.

Informační zdroje lze rozdělit do několika skupin, které budou následně podrobněji rozebrány. Jde o literaturu zaměřenou na technické řešení samotných datových skladů a dále o zdroje související s problematikou ochrany osobních údajů. Jako poslední jsou pak uvedeny zdroje obecné, které pomohly autorovi propojit uvedené oblasti a navrhnout řešení v rámci této práce.

Pro návrh konkrétních řešení a pro získání představy, jak zvažovaná řešení reálně implementovat, absolvoval autor této práce v únoru 2019 on-line kurz Delivering a Relational Data Warehouse (Microsoft, 2019). Tento kurz využívá cloudové prostředí MS Azure a účastník projde kompletním procesem návrhu a implementace datového skladu. Inspirací byly autorovi i další méně související on-line kurzy absolvované paralelně při studiu VŠE v předešlých letech – Data Science Ethics (University of Michigan, 2018), Ethics and Law in Data and Analytics (Microsoft, 2018) a Computer Science for Business Professionals (Harvard University, 2018).

Dalším okruhem informačních zdrojů byly publikace týkající se datových skladů (Humpries, 2002), (Laberge, 2012), (Lacko, 2003) a (Lacko, 2009). Pro teoretickou část týkající se dat obecně byl využit (Gála, 2015). Tištěné publikace byly doplněny znalostmi získanými z řady internetových (např. (Zentut.com, nedatováno)).

Pro vypracování případové studie v závěru práce využil autor pojednání doktorandů z VŠE v Praze (Hrabě, 2016) a University Karlovy (Olecká, nedatováno), kdy oba rozebírají problematiku případové studie dle (Yin, 2009).

Relevance tématu vychází i z autorova osobního zájmu o problematiku GDPR jako takovou, kdy autor absolvoval komplexní zhruba 100 hodinové školení. Dále jsou bohatým zdrojem informací z této oblasti webové stránky správního úřadu pro problematiku GDPR v české republice (Úřad pro ochranu osobních údajů, 2019) a web Evropské unie (Evropská unie, 2019), kde je k nalezení jak nepřeberné množství komentářů a doplňujících informací k GDPR, tak i samotné znění předmětného nařízení. Jako další zdroje z oblasti GDPR využil autor publikace (Navrátil, 2018), (Nezmar, 2017) a (Žůrek, 2017), které poskytují více či méně komplexní přehled celé problematiky včetně praktických námětů na implementaci GDPR do praxe. Informace a názory autora jsou rovněž podepřeny nespočtem vyjádření odborníků na sociálních sítích, jako např. Mgr. Evy Škorničkové (Škorničková, 2019), která patří k významným názorovým leaderům v oblasti GDPR a kybernetické bezpečnosti v ČR.

Za informační zdroje je třeba rovněž označit kompendium znalostí autora získaných studiem na VŠE v rámci programu Podniková informatika, jehož součástí je i tato práce. Zdrojem pro tuto práci tak byly nejen samotné přednášky a výuka, ale i autorem zpracovávané semestrální práce jako např. práce z oblasti dobývání znalostí z databází (Janoušek, 2018) nebo bakalářská práce (Janoušek, 2017). V souvislosti se studiem se zároveň díky aktivitám VŠE v Praze otevírá prostor k účasti na řadě akcí, z nichž nejvýznamnější pro tuto práci byla autorova účast na KPMG Data festivalu (KPMG, 2019). Zde bylo možno nejen navštívit např. přednášku přímo související s DWH či na další relevantní témata, ale zároveň i s dodavateli z praxe diskutovat otázky související (nejen) s touto prací.

Naprosto zásadním informačním zdrojem, který autorovi vůbec umožnil uvažovat o řešeném tématu této práce je vlastní praxe autora (Janoušek, 2019), kdy autor v minulosti osobně navrhoval architekturu inovace firemního informačního systému silně vázanou na zákaznická data včetně návrhu datových struktur, čištění dat, datových analýz atp. Tento projekt bude využit jako podklad pro případovou studii v závěru této práce. V současné době se autor v roli HR konzultanta setkává s řadou klíčových pracovníků IT firem (od Team leaderů přes Scrum mastery, Product owners po CTO a CIO) a má tak možnost získat technický vhled do problematiky přímo od těch nejpovolanějších lidí z praxe. Navíc není odkázán jen na jednu firmu, ale má možnost nahlédnout do celého spektra firem a vnímat i rozdílnosti v přístupu. Zároveň je autor z druhé strany v úzkém kontaktu s kandidáty na pracovní pozice u zmíněných firem (vývojáři, systémoví administrátoři, QA inženýři, datoví analytici a konzultanti atp.), kde má opět prostor diskutovat témata z IT praxe nezávisle na konkrétních technologiích. Tuto praxi autor považuje za daleko cennější než např. konkrétní stáž na jednom pracovišti, protože dává autorovi přehled napříč technologiemi i v různém firemním prostředí (od malých firem po velké korporace). Právě takto získaných kontaktů mohl autor využít i ke konzultacím na téma této práce.

Samotné využití všech informačních zdrojů autor této práce pojal jako kompilaci získaných informací a formulování jednotlivých myšlenek souhrnně z těchto zdrojů. Autor považuje tento přístup za výhodnější oproti přímým citacím konkrétních úryvků textu proto, že tento způsob dává prostor demonstrovat autorovo pochopení problematiky v širším kontextu a schopnost jednotlivé informační zdroje kompilovat do jednoho celku a zároveň doplnit o vlastní myšlenky rozvíjející takto získané informace.

1 Úvod ke zpracování dat

Tato kapitola představuje vlastní úvod do zpracování dat. Jejím cílem je uvést čtenáře do problematiky dat jako takových a přinést představu o koncepci Datových skladů (Datawarehouse, DWH) a jejich využití.

Mezi základní informační zdroje, které byly zmíněny již v *Úvodu* této práce, a které tvoří podklad této kapitoly, patří (Gála, 2015), (Humpries, 2002), (Laberge, 2012), (Lacko, 2003), (Lacko, 2009), (Microsoft, 2019) a (Zentut.com, nedatováno). Zejména pro oblast práce s daty obecně pak autor čerpá i ze svých praktických zkušeností (Janoušek, 2019) a své předešlé Bakalářské práce (Janoušek, 2017).

1.1 Data, informace, znalosti

Máme-li se zabývat zpracováním osobních údajů, měli bychom si napřed ujasnit, co vlastně osobní údaj je. Definici právní nám poskytne *Podkapitola 2.1.3*. Ještě předtím se ale musíme dostat na nižší úroveň detailu, abychom mohli dále pracovat s propojením právního pohledu a informatiky, tedy konkrétně na zpracování dat.

Informatika se dle (Gála, 2015) zabývá vyjádřením, zpracováním a přenášením **informací** v určitém (informačním) systému. Informace je zde definována jako pojmenování pro obsah toho, co se vymění s vnějším světem, když se mu přizpůsobujeme a působíme na něj. Pro přenos informací mezi minimálně dvěma účastníky slouží komunikace. Ta pro přenos informací využívá systému znaků. Aby byly obě (všechny) strany komunikace schopné si vzájemně rozumět, využívá se kódu (strukturovaný systém znaků).

Pro porozumění kódu je pak klíčová sémantika (význam) a syntaxe (vztah mezi jednotlivými znaky navzájem). Převedeno na příklad běžného lidského jazyka, sémantika se zabývá významem textu, zatímco syntaxe řeší strukturu textu (tečka znamená konec věty, za tečkou se vkládá mezera a na začátku každé věty se používá velké písmeno). Sémantika a syntaxe nabývá na významu při snaze automatizovat zpracování informací pomocí počítačů, kde je právě formální správnost vyjádření naprosto nezbytnou podmínkou úspěchu.

Data, (jednotné číslo = údaj) a tedy i údaje, vyjadřují (Gála, 2015) informace v podobě formalizované soustavy symbolů (znaků), která umožňuje přenos, uchovávání, interpretaci či zpracování informací (ať už informačním systémem nebo i člověkem).

Znalostí pak (Gála, 2015) rozumíme soubor informací v určitém kontextu či vzájemné vazbě. Lidé získávají znalosti studiem či zkušeností a využívají je následně k řešení problémů či rozhodování.

Uveďme si jednoduchý příklad. Máme *údaj* „200°C“. Bez kontextu nám to neřekne nic. Za *informaci* pak můžeme považovat to, že „trouba v kuchyni je rozpálená na 200°C“.
Znalostí pak je to, že „položíme-li ruku na tuto troubu, spálíme se“.

Převedeno do oblasti zpracování osobních údajů pak máme např. *údaje* o našem klientovi: ID = 1; e-mail: jan.janousek@vse.cz. Na první pohled vidíme *informaci*, že náš klient označený interním číslem 1 má e-mailovou adresu v uvedeném tvaru. *Znalostí* pak je, že můžeme tuto adresu použít k zaslání e-mailu klientovi, anebo v tomto případě, že se daný klient zjevně jmenuje Jan Janoušek a zřejmě studuje či pracuje na VŠE v Praze[1].

Rozlišení pojmů výše nám pomůže lépe porozumět tématu zpracování osobních údajů. Riziko při zpracování osobních údajů, které jsou často samy o sobě „neškodné", spočívá právě v jejich možné kombinaci a získávání informací o jednotlivých osobách. Případné znalosti získané z porovnávání informací o jednotlivých osobách v porovnání s ostatními pak může vést až k mylné představě o daném člověku. Pokud např. zjistíme, že webové stránky podporující rasismus navštěvují uživatelé s obdobným smýšlením, snadno se nám stane, že za takto smýšlejícího označíme i uživatele, který je zásadně proti a třeba jen hledá související fakta a zkoumá opačné názorové proudy. A čím více se snažíme uživatele posuzovat automatizovaně, tím více hrozí vznik předsudků či nepřesností, ačkoli základní premisa využívání počítačů je, že počítače předsudky nemají a nebudou své závěry nijak citově zabarvovat a promítat do nich např. vlastní pocity.

Práci s informacemi v rámci podniku zajišťuje **podnikový informační systém**. Ten se (Gála, 2015) skládá z **neformální** (výměna a zpracování informací lidmi s využitím jazyka, mimiky či gest) a **formální** části (informační toky na základě formalizovaných politik, cílů, strategií apod.). Zároveň je zde velký prostor pro využití počítačů pro účely automatizace a zefektivnění informačních toků a zpracování dat.

Zúžíme-li svůj pohled na podnikovou informatiku jako klíčového reprezentanta práce s podnikovými informacemi, dojdeme k široce využívanému pojmu **ICT** (informační a komunikační technologie). Ty sestávají (Gála, 2015) z počítačů a jejich periferií, komunikačních zařízení, spotřební elektroniky, software a dále zahrnují i související služby (IT poradenství, telekomunikační služby, pronájmy ICT zařízení).

Fyzická zařízení (**HW**, hardware) jsou uživateli zpřístupněna prostřednictvím programového vybavení (**SW**, software), které rozdělujeme (Gála, 2015) na aplikační software (**ASW**), základní software (**ZSW**) a software podporující rozvoj IS, vývoj ASW a zajišťující podporu řízení provozu IS.

Využití ASW pak přináší uživatelům možnost pracovat s **podnikovými daty**. Rozlišujeme (Gála, 2015) na **klíčová referenční data** (údaje o zdrojích podniku a jeho schopnostech i omezeních – např. organizační struktura, konfigurace strojů a zařízení, geografické údaje o provozovnách apod.), **kmenová data** (charakteristiky zákazníků, produktů, materiálů a komponent apod.), **podmíněná kmenová data** (byznys pravidla jako např. cenová

[1] Protože víme, že doména vse.cz patří zmíněné škole a protože je to škola, je zde evidentní buďto studijní nebo zaměstnanecký vztah. Dokonce bychom mohli díky znalostem jít ještě dále, protože víme, že studentům jsou generovány tzv. xname, tedy text před zavináčem je složen z prvních tří znaků příjmení, jednoho znaku křestního jména a pořadového čísla této kombinace v systému obvykle doplněný znakem x na začátku. V textu uvedená adresa tak spíše odpovídá vyučujícímu, který již dosáhl jisté úrovně (protože i vyučující s titulem PhD. obyčejně nadále využívají adresu získanou ještě za statusu studenta. E-mailová adresa je samozřejmě smyšlená.

politika ve vztahu k typu zákazníka, podmínky ovlivňující pracovní tok – tzv. workflow, apod.), **transakční data** (realizace jednotlivých instancí byznys procesů) a **reporty** (data o stavu, data pro výstupní sestavy, analýzu a rozhodování).

Z perspektivy našeho tématu jsou zajímavá kmenová data (např. informace o zákaznících či zaměstnancích), transakční data (modifikace kmenových dat a jejich obohacování o další souvislosti) a reporty (tento typ dat je typickou ukázkou využití DWH).

Práce s podnikovými daty probíhá buďto automatizovaně na základě byznys pravidel, anebo ji provádí pracovníci podniku s využitím ASW. Nejtypičtějším zástupcem jsou počítačové aplikace **ERP** (enterprise resource planning), které mají za úkol realizovat workflow klíčových firemních procesů a zajišťovat hlavní datové transakce podniku. Z naší perspektivy jsou pak rovněž významné aplikace **CRM** (customer relations management) pro řízení vztahů se zákazníky a aplikace **SCM** (supply chain management) pro řízení vztahů s dodavateli. Velmi důležitým zástupcem ASW z našeho pohledu jsou i aplikace pro podporu rozhodování postavené na bázi **BI** (business intelligence). Ty jsou realizovány prostřednictvím procesů transformace dat **ETL** (Extract, transform, load) a fungují tak, že přenášejí data z produkčních databází využívaných při běžném provozu (např. ERP, CRM) do specializovaných úložišť určených pro skladování takových dat v datových skladech (**DWH**, data warehouse) či datových tržištích (**DMA**, data mart) a tvoří následný reporting. Právě této problematice se budou hlouběji věnovat následující podkapitoly.

Podkladem pro fungování aplikací, které musí s daty pracovat, jsou **datové modely**. Tyto modely jsou reprezentovány v podobě schémat, na základě kterých jsou pak vytvářeny konkrétní databáze. Při modelování využíváme (Gála, 2015) **konceptuální schéma** (jde o pohled uživatele bez znalosti technických souvislostí), **logické schéma** (transformuje koncept do konkrétní podoby datových modelů s ohledem na podmínky jazyka a formy datových prvků) a **fyzické schéma** (transformuje logické schéma do konkrétní technologie – např. MS SQL server jako úložiště dat).

Při využívání **relačních databází** a jazyka SQL, což je případ naší práce, budeme při modelování zohledňovat specifika těchto technologií. Principem relační databáze je soustava tabulek, které jsou vzájemně provázané prostřednictvím tzv. **klíčů**. Klíče jednoznačně identifikují daný záznam v tabulce a umožňují tak vytváření vazeb (**relací**) mezi jednotlivými tabulkami. Představme si např. tabulku se seznamem zákazníků, kde bude tzv. **primárním klíčem** (jde tedy o jednoznačnou identifikaci každého jednotlivého záznamu v tabulce) položka ID zákazníka. Dále máme jinou tabulku se seznamem produktů, kde primárním klíčem bude ID produktu. Následně připravíme další tabulku, která bude obsahovat seznam objednávek. Sama o sobě bude mít jako primární klíč ID objednávky, ale v každém záznamu zároveň v podobě tzv. **cizích klíčů** využije ID zákazníka, který objednávku učinil, a ID produktu, který zákazník objednal. Právě zde využijeme primární klíče z dalších tabulek a vzniknou nám požadované vazby.

Důvodem, proč takto tvoříme vazby mezi tabulkami, je využití základních vlastností relačních databází. Tím je zajištění **integrity dat**. Díky použití cizích klíčů, které odkazují na primární klíče v jiných tabulkách, máme jistotu, že se do tabulky nedostane záznam, který by odkazoval na neexistující hodnotu (tzv. **referenční** integrita). Jinými slovy např.

nemůže uživatel zadat do systému objednávku pod ID zákazníka, který není v systému zaveden, popř. do objednávky vložit neexistující zboží.

Rovněž je důležité (díky primárním klíčům) zajištění tzv. **entitní** integrity dat, kdy není možné v systému vytvářet duplicitní záznamy, které by vedly k nesprávnému fungování systému. Např. tak není možné zadat do systému stejného zákazníka dvakrát[2].

Posledním typem integrity dat je **doménová** integrita. Ta zajišťuje, že jako hodnotu atributu budou zadány vždy jen přípustné hodnoty. Díky tomu např. v objednávce není možné zadat množství objednávaného zboží jinak než v podobě celého čísla nebo třeba celého nenulového a nezáporného[3].

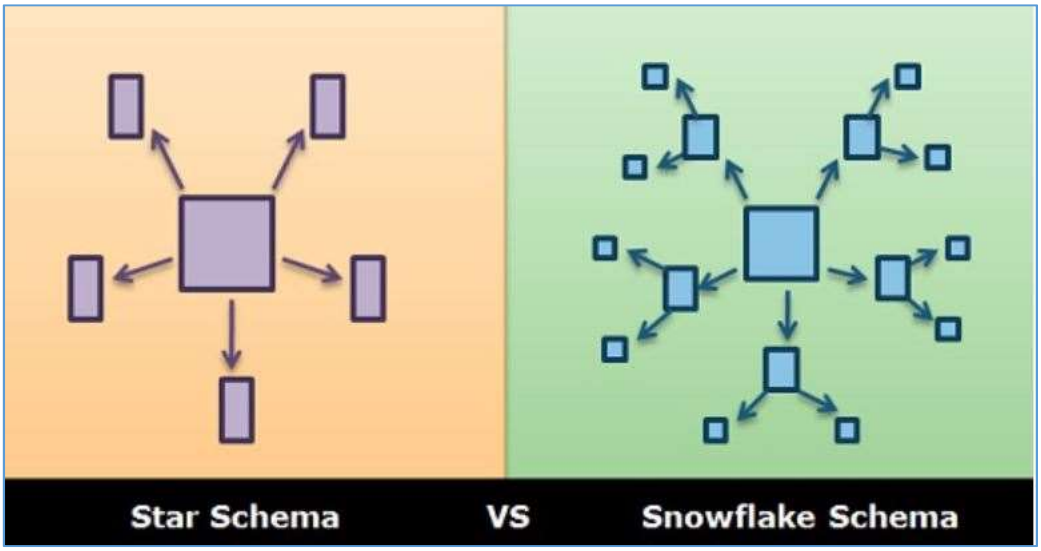
Uspořádání tabulek a jejich relací pak musí zohledňovat hlediska přehlednosti a výkonu při práci s tabulkami. Nabízí se zadání všech dat do jedné tabulky. Tím bychom dosáhli zjednodušení ve smyslu eliminace relací. Naproti tomu bychom však systém při práci s daty neustále nutili pracovat se všemi daty a tím by se celý výkon značně zbrzdil. Proto jsou nejčastěji tabulky rozdělené na tzv. **tabulky faktů** a **tabulky dimenzí**.

Faktové tabulky typicky obsahují samotná data o transakcích. Zjednodušeně data, která jsou proměnlivá a velmi rychle se mění nebo přibývají. Příkladem může být třeba tabulka objednávek zákazníků. Naproti tomu dimenzemi rozumíme relativně stálá data. Např. seznam výrobků.

Podle toho, zda jsou dimenze v jedné tabulce nebo jsou ještě rozdělené do dalších podtabulek, rozlišujeme datové schéma na schéma typu hvězda (tzv. **star**) nebo sněhová vločka (tzv. **snowflake**). Struktura star je taková, kdy jsou na jednu faktovou tabulku samostatně navázány jednotlivé dimenzionální tabulky, které se již dále nerozvíjejí. Snowflake schéma pak má dimenzionální tabulky, na které dále navazují další dimenzionální tabulky obsahující větší hloubku detailu dané dimenze. Vizuálně tyto situace vykresluje *Obrázek 1.* Představme si třeba situaci, kdy dimenzí bude produkt. Ten má nějaký název (úroveň 0) a spadá do nějaké subkategorie (úroveň 1) produktů, ta spadá do kategorie (úroveň 2) atp. Záleží pak na konkrétní situaci, zda zmíněnou strukturu vložíme do jedné dimenzionální tabulky, anebo vytvoříme pro každou úroveň (produkt, podkategorie, kategorie) zvláštní tabulky a ty pak navážeme postupně na sebe. Oba postupy mají svá pozitiva i negativa a záleží vždy na konkrétní situaci, který z nich zvolíme.

[2] Pochopitelně za předpokladu, že je systém korektně nastaven.
[3] Přesné zadání závisí na byznys pravidlech v daném případě.

Obrázek 1: Porovnání Star a Snowflake schéma



Zdroj: (TechDifferences.com, 2017)

Pro práci s daty je rovněž důležitý pojem **metadata**. Ta jsou vnímána jako tzv. data o datech. Jde o data, která popisují jiná data namísto toho, aby nesla nějakou konkrétní informaci. Díky nim jsme schopni s daty pracovat, protože dokážeme popsat, co která data vlastně znamenají. Máme-li tabulku se seznamem výrobků, pak metadata najdeme v záhlaví sloupců. Nevyjadřují tedy konkrétní hodnoty, ale definují, jaká data v daném sloupci najdeme. Pokud budeme tedy v případě osobních údajů umět např. říci, že jméno a příjmení považujeme za osobní údaj, pak jsou to právě metadata, která nám pomohou v databázi identifikovat, kde se dané osobní údaje nachází.

1.2 Teoretický koncept datového skladu

V této podkapitole budeme čerpat z (Lacko, 2003), (Lacko, 2009) a (Gála, 2015). Podkapitola zasadí DWH do souvislostí celého BI konceptu. Porozumění souvislostem je v práci s daty obecně velmi důležité a stejně tak důležité je i tuto práci neomezovat jen na objasnění koncepce DWH bez širších souvislostí.

Základní **principy** řešení **BI** spočívají ve správném výběru a organizaci dat, v ukládání dat na odpovídající úrovni detailu (tzv. *granularity*) a zároveň ukládání dat agregovaných, ve využití multidimenzionálního pohledu na data včetně časové dimenze, ve vysokých nárocích na kvalitu dat.

Databáze určené pro transakční aplikace (např. ERP a CRM) jsou optimalizované pro pořizování transakcí a naopak nejsou vhodné pro účely reportování. Navíc analytické úlohy nad zdrojovými databázemi mohou vést ke zpomalení až nepoužitelnosti transakčních aplikací. Z toho důvodu vznikly analytické databáze, do nichž jsou data kopírována. **Výběr a organizace** dat probíhá prostřednictvím tzv. **ETL** procesu (zkratka z označení jednotlivých fází *extract*, *transform* a *load*). Dochází tak k výběru potřebných dat, jejich načtení do tzv. DSA (data staging area), transformaci do nových datových modelů

vhodných pro analytické úlohy a dále fyzické uložení do připravených struktur. Součástí tohoto procesu je i čištění dat, kdy dochází k opravě chybných údajů či odstranění duplicit apod. v datech tak, aby nebyly narušeny následné analytické úlohy. Kromě změny uspořádání dat může dojít i k transformaci jejich obsahu (např. převod kategoriálních proměnných na číselné pro využití neuronových sítí v rámci úloh dobývání dat z databází). Rovněž může docházet k výpočtu nových hodnot a k tvorbě agregovaných hodnot, které se ukládají přímo do analytické databáze, aby jejich následné výpočty nezpomalovaly průběh úloh.

Granularitou dat rozumíme míru úrovně detailu dat. Obecně platí, že se do analytických databází ukládají data v maximální míře granularity pro všechny **dimenze** a současně jednotlivé agregace na každé úrovni, aby bylo možno využívat tzv. drill-down a drill-up přístupu, kdy ze zobrazené úrovně dat můžeme postupně získávat větší detail, anebo naopak globálnější pohled. Nicméně i s granularitou je třeba zacházet opatrně. Ne vždy je nezbytné ukládat data v úplném detailu. Např. v případě časové dimenze máme každou transakci zaznamenánu s přesností na vteřiny. Ne vždy je však třeba mít takto detailní úroveň. Pokud v systému běžně zaznamenáváme jednu transakci za řádově minuty, je dimenze vteřin zbytečná a jen naroste databáze, kterou bude úroveň vteřiny pouze zatěžovat, protože bude obsahovat tutéž informaci jako dimenze minuty, ale vynutí si zároveň ještě totožná „agregovaná“ data. Namísto jednoho řádku tak databáze spotřebuje čtyři. Rovněž se právě zde může odrazit velmi efektivní řešení práce s osobními údaji. Uvažme, zda pro analytické úlohy potřebujeme v případě dimenze zákazníků detail až na jednotlivého zákazníka, a tedy zda je opravdu nutné do analytických databází kopírovat jeho osobní údaje, abychom jej uměli rozlišit.

Pokud jsme zmínili dimenze v pojetí relačních databází, uveďme okrajově i pojetí dimenzí v prostředí **OLAP** (on-line analytical processing) technologií. Přestože toto prostředí není v této práci využito, krátce si jej popíšeme. Databáze OLAP totiž nepracují na principu vzájemně propojených tabulek. Jejich pojetí je takové, že jsou dimenze přidány jako de facto další rozměr v tabulce. Budeme-li např. mít tabulku prodeje určitého výrobku v čase, pak další dimenze mohou tvořit např. organizační či geografické jednotky. Tím se dostáváme k typickému využití tzv. OLAP kostky. Díky tomu lze na stejná data nahlížet z různých úhlů, jako kdybychom kostkou otáčeli v prostoru. To umožňuje velkou rychlost zobrazení vybraného pohledu na data, protože nedochází k žádným výpočtům (u relační databáze musíme při změně pohledu přepočítat vazby a načíst dané agregované hodnoty na dané úrovni).

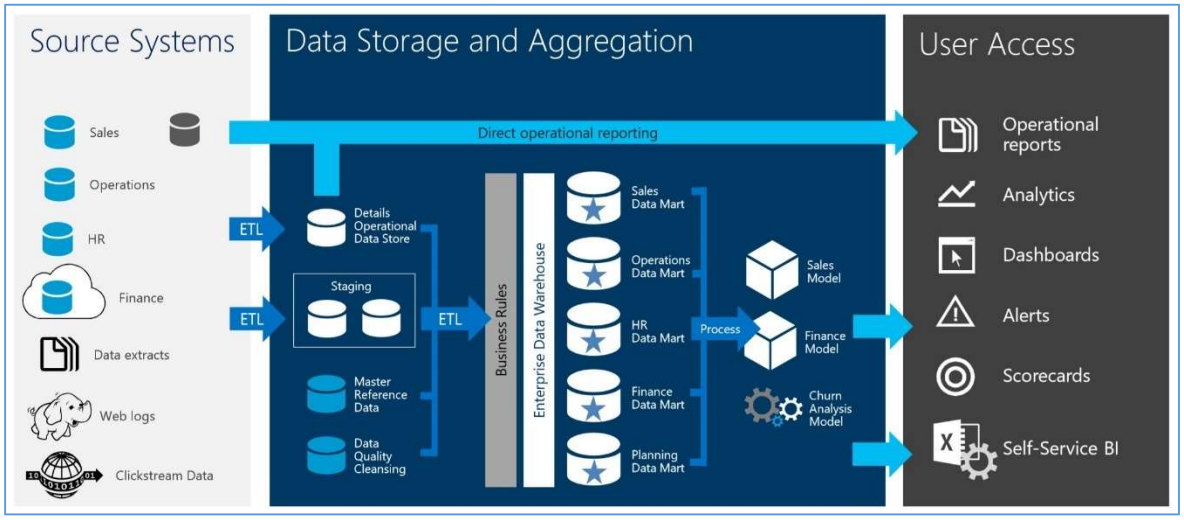
Pro uložení dat pro analytické účely na závěr procesu ETL bývají využívány právě DWH nebo DM. K architektuře podnikového DWH jsou dva základní přístupy nazvané podle svých autorů Ralpha **Kimballa** a Billa **Inmona** (Zentut.com, nedatováno). Zatímco Kimball preferuje vytvoření jednoho dimenzionálního DWH, které obsahuje relacemi provázané tabulky uspořádané uvnitř do struktury DM, Inmon preferuje vytvoření enterprise DWH, na který jsou navázány samostatné DM pro jednotlivá oddělení či jiné organizační struktury (např. marketing, sales, finance apod.). Stejně jako při volbě mezi *star* a *snowflake* schématem databáze ani zde není jedno z řešení správné či špatné. Záleží opět na konkrétní situaci a důležité je spíše pracovat systematicky a raději se držet jednoho z přístupů. Význam to má proto, že jsou dané přístupy vyzkoušené a ověřené praxí.

Posledním z uvedených principů BI je řešení **datové kvality**. Ta se posuzuje z hlediska dostupnosti dat (**availability** - dostupnost v požadovaném čase, místě, struktuře a formátu), přesnost dat (**accuracy** – data jsou ve správném kontextu – např. PSČ odpovídá adrese), úplnost dat (**completness** – např. úplnost adresy) a konzistence (**consistency** – nedochází k narušení vazby mezi daty). Není náhodou, že se obdobné požadavky objevují i v souvislosti s GDPR a řízením rizik.

1.3 Implementace DWH dle Microsoft

Jak jsme si upřesnili výše, DWH je součástí procesu BI. Jako zdroj reálné simulace tohoto procesu pro potřeby této práce posloužil on-line kurz společnosti Microsoft (Microsoft, 2019) zaměřující se na implementaci DWH v prostředí MS Azure. Celý proces je přehledně představen v modulu 1 kurzu a v této podkapitole bude volně parafrázován. Východiskem pro další řádky bude schéma dle *Obrázek 2*.

Obrázek 2: Schéma procesu BI dle Microsoft



Zdroj: (Microsoft, 2019)

Základem pro veškerou práci s daty jsou zdrojové systémy (Source systems). Ty mohou být nejrůznějšího druhu. Ať už jde o podnikový ASW typu ERP, CRM apod. provozovaný on-site (na vlastním HW) či prostřednictvím cloudu (např. formou SaaS – Software as a Service), anebo různé jiné zdroje (např. data získaná z internetu – tzv. big data). Tato data mohou být replikována do samostatné databáze na straně zdrojových systémů a může k ní být vytvořen přímý přístup pro uživatele pro účely operativního reportingu (Direct Operational Reporting). Takový přístup je však vhodný pouze pro výkonnostně méně náročné úlohy a pro jednoduché úlohy, popř. pro účely získání informace s okamžitou platností (např. aktuální stav zboží na skladě). Naopak se tento postup nehodí pro úlohy analytické, kdy potřebujeme pracovat s daty agregovanými, rozsáhlými a z pohledu různých dimenzí. Takové úlohy jsou kromě výkonnosti omezené i citlivostí na kvalitu dat.

Za tímto účelem vznikla právě úložiště typu DWH, DM a celý proces ETL, který slouží k naplnění těchto úložišť daty tak, jak je to znázorněno ve schématu v sekci *Data Storage*

and Aggregation. Jak je z obrázku patrné, je využita struktura dle Inmona (viz *Podkapitola 1.2*), kdy výstupem ETL procesu je Enterprise Data Warehouse vytvořený na základě *Business Rules* (byznys pravidel). Na DWH pak navazují jednotlivé Data Marty pro specifická oddělení. Z nich jsou následně produkovány datové modely určené pro jednotlivé analytické úlohy.

Samotný ETL proces je rozdělen na dvě fáze, mezi nimiž se nachází specializované procesy *Master Data Management* a *Data Quality Management*, které jsou součástí procesu *Enterprise Information Management*. Právě toto „vyčlenění" poukazuje na specifické postavení těchto procesů a jejich souvislosti s vyšším celkem. A to je právě ukázkou toho, že data jsou v rámci podniku vnímána jako strategická hodnota.

Master Data Management je o kontrole dat na úrovni celého podniku a je tím místem, kde lze nejefektivněji pracovat s metadaty na úrovni podniku tak, aby byly jednotlivé entity sjednoceny. Příkladem může být používání jednotného seznamu produktů napříč celým podnikem, kdy každé oddělení využívá svůj úhel pohledu na daná data, a může tak často vznikat situace, že každé oddělení eviduje jiná data o týchž entitách. Jiný pohled budou mít účetní, jiný logistika a jiný marketing. V praxi se pak často stávalo, že každé oddělení evidovalo data o produktech odděleně ve svých specializovaných informačních systémech a s různými atributy. Je nasnadě, že takové sjednocení vede k efektivitě nejen z hlediska samotné práce s daty, ale především z hlediska rozhodování a analytických úloh.

1.4 Shrnutí kapitoly

Kapitola 1 nás provedla úvodem do problematiky zpracování dat. Byly vysvětleny základní pojmy jako jsou data, informace a znalosti, a dále jsme zde našli stručný přehled podnikových informačních systémů tak, abychom si mohli začlenit zpracování dat do celého systému. Problematika byla následně zúžena přímo na zpracování dat v databázích a na získání představy o jejich modelování. Dále jsme si udělali obrázek o procesu BI a o tom, jak do tohoto procesu zapadá využívání DWH. A následně jsme zrekapitulovali, jak se DWH tvoří a v závěru kapitoly, jaký koncept pro proces BI využívá společnost Microsoft, jejíž řešení se stalo podkladem pro tuto práci.

Tím jsme získali vhled do problematika zpracování dat (a tedy i osobních údajů) z pohledu informatiky. Následující kapitola přinese vhled z pohledu právního. Na základě těchto dvou pohledů pak bude následovat konkrétní návrh autora na řešení problematiky ochrany osobních údajů v rámci DWH.

2 Právní pohled na osobní údaje

Jak bylo zmíněno již v *Úvodu* této práce, následující kapitola vychází ze samotné legislativy a doprovodných informací EU (Evropská unie, 2019), webu relevantního správního úřadu v ČR (Úřad pro ochranu osobních údajů, 2019), vyjádření odborníků na danou problematiku (Škorničková, 2019), školení absolvovaného autorem a praxe, v níž se téma ochrany osobních údajů v posledním roce do značné míry objevuje (Janoušek, 2019). Dalšími podklady byly publikace k problematice GDPR (Navrátil, 2018), (Nezmar, 2017) a (Žůrek, 2017).

Tato kapitola neslouží jako obecný a detailní průvodce ochranou osobních údajů. Jejím cílem je problematiku objasnit jen v míře dostatečné pro daný kontext a vyzdvihnout zejména to, co je relevantní k tématu DWH. Budou zmíněny skutečnosti, na jejichž základě autor staví svůj návrh řešení.

2.1 Ochrana osobních údajů

S účinností od 25. 5. 2018 nabylo v platnost obecné nařízení Evropské unie č. 2016/679, které se vžilo pod zkratkou **GDPR** (General Data Protection Regulation, dále jen „Nařízení" nebo „GDPR"). Smyslem tohoto nařízení je ochrana fyzických osob v souvislosti se zpracováním osobních údajů. Ačkoliv bylo nařízení schváleno již v roce 2016 a všechny státy EU měly 2 roky na jeho implementaci, stalo se ústředním tématem veřejných diskusí až počátkem roku 2018 a ostatně i adaptační zákon byl v ČR schválen až v roce 2019. Přestože legislativa samotná není revolucí v ochraně osobních údajů, diskuse eskalovaly zejména z titulu maximální výše stanovených sankcí. Ty jsou nastaveny tak, aby mohly znamenat citelný zásah do rozpočtu i pro globální společnosti na světovém trhu. Vznikla však jistá forma paniky, kdy právě tyto sankce vyvolaly strach u zástupců malých a středních podniků či v oblasti veřejné správy a samosprávy.

Uvedený strach vycházel ze samostatného pojetí legislativy, která pracuje na bázi stanovení zásad a apeluje na používání zdravého selského rozumu. Vyžaduje přístup „*privacy by design*", čímž je míněno, že ochrana soukromí by měla být zakotvena již na úrovni designu byznys procesů. Tedy že již samotný byznys je postaven na záměru chránit zpracovávaná data, čímž dojde k maximalizaci této ochrany.

Právě díky tomuto pojetí považuje autor této práce legislativu za inovativní a zajímavou, protože již jen nestačí splnit nebo nesplnit nějaký konkrétní pokyn a dostat za to či nedostat konkrétní sankci. Pojetí GDPR přináší **etický rozměr** a právě etika je dle názoru autora této práce východiskem úspěšné implementace nařízení do praxe. Pokud bychom totiž s osobními údaji nakládali eticky a s čistými úmysly, nemusela legislativa vzniknout. A právě implicitní **záměr chránit** svěřené **osobní údaje** a nakládat s nimi jen v rozsahu nezbytně nutném a ve prospěch subjektu těchto údajů je cesta, jak snadno dosáhnout

souznění s GDPR. Problematika změny přístupu k osobním údajům by však vydala na samostatnou práci.

Dalším principem, na kterém legislativa stojí, je **otevřenost a informovanost**. Zjednodušeně řečeno, každý by měl být srozumitelně informován o zpracovávání svých osobních údajů včetně informací o jejich ohrožení (např. neprodlená informace při jejich úniku).

V následujících řádcích se budeme blíže věnovat jednotlivým detailům tak, aby nám kapitola poskytla dostatečný přehled problematiky pro potřeby kontextu DWH.

2.1.1 Prameny práva

Abychom si formálně zasadili obecné nařízení do legislativního rámce, uvedeme si zde stručný přehled vývoje ochrany osobních údajů.

Ochrana osobních údajů je jednou ze skutečností, která nabyla obrovského významu v souvislosti s prudkým rozvojem informačních technologií v posledních desetiletích. A právě rychlost tohoto rozvoje vedla k tomu, že legislativa nebyla schopna fungovat jako prevence a předejít potenciálním problémům. Velmi podobná situace se v dnešní době odehrává v souvislosti s rozvojem AI (Artificial Intelligence; Umělá inteligence). V tomto případě naštěstí již ve fázi počínajícího rozmachu vnímá odborná veřejnost hrozbu jejího zneužití a celosvětově se objevuje celá řada aktivit podporující etický a opatrný přístup k AI.

Aktuálně platné **nařízení** EU č. **2106/679** navazuje na předešlou **směrnici** EU č. **95/46/ES**. Jak vyplývá z pojetí legislativy EU (Evropská unie, 2019), směrnice EU nemá přímou účinnost v členských zemích a zakládá jejich povinnost zapracovat její ustanovení do místní legislativy. V ČR tak platil zákon č. **101/2000 Sb.** o ochraně osobních údajů. Jeho platnost byla technicky vzato ukončena právě účinností *Nařízení* (to na rozdíl od směrnice platí v členských státech automaticky), formálně byl v naší legislativě nahrazen až zákonem č. **110/2019 Sb.**, který je vnímán jako adaptační zákon k *Nařízení*.

Jako správní orgán pro realizaci *Nařízení* v ČR byl ustanoven Úřad pro ochranu osobních údajů, který danou problematiku zastřešoval již v předešlém období.

V širším kontextu pak problematika ochrany osobních údajů vychází jak z Listiny základních práv EU, tak i v rámci naší legislativy z Listiny základních práv a svobod.

2.1.2 Koho se GDPR týká

Předmětem ochrany jsou osobní údaje živých fyzických osob (tzv. **Subjekt údajů**) na území EU (*Kapitola I Nařízení*). Legislativa chrání ta zpracování osobních údajů, která probíhají na území EU bez ohledu na to, zda je *Subjekt údajů* občanem EU. Zároveň tak legislativa nechrání občana EU při zpracování osobních údajů mimo EU. Tím však není dotčena ochrana při předávání osobních údajů z EU mimo její území.

Dochází-li tedy ke zpracování osobních údajů v souvislosti s poskytováním služby na území EU, nezáleží na sídle poskytovatele služby. Např. vyhledáváme-li v ČR údaje na webu

google.com (tedy využíváme konkrétní službu), musí s našimi osobními údaji společnost Google nakládat v souladu s GDPR, ačkoliv její sídlo není na území ČR, a to bez ohledu na to, zda jsou data zpracovávána mimo EU. Zároveň v takové situaci platí, že takové zpracování podléhá legislativě bez ohledu na to, zda službu využil občan EU či kdokoliv jiný.

Na druhou stranu, pokud budeme využívat nějakou službu provozovanou v ČR, vztahuje se na ni daná legislativa bez ohledu na to, že jsme se ke službě připojili mimo území EU a opět bez ohledu na to, zda jsme či nejsme občany EU.

Poslední variantou pak je služba provozovaná mimo EU a využitá zákazníkem mimo EU. Pokud osobní údaje nejdou přes území EU, legislativa se na ně nevztahuje. A to ani v případě, že by takovou službu poskytovala společnost se sídlem v EU (aniž by ovšem data putovala do EU) občanovi EU, který se v dané chvíli nachází mimo EU.

Na druhé straně zpracování osobních údajů je pak tzv. **Správce osobních údajů** (dále jen „Správce"). To je subjekt, který s osobními údaji nakládá. Podstatné je, že daný subjekt nese svou zodpovědnost za zpracování osobních údajů bez ohledu na to, zda samotným zpracováním pověří třetí stranu (např. oustourcing). Zodpovědnosti se *Správce* nemůže zbavit ani formou ujednání se *Subjektem údajů*. *Správce osobních údajů* je ten, kdo osobní údaje získává a rozhoduje o tom, jak budou zpracovány. Může jím být jak fyzická, tak i právnická osoba, bez ohledu na to, zda jde o komerční subjekt či např. úřad, neziskovou organizaci, školu apod. Nerozhoduje tedy právní forma, nýbrž to, zda daný subjekt provádí zpracování osobních údajů či nikoliv.

Zpracovatel osobních údajů je subjekt, který osobní údaje reálně zpracovává, ať už na základě pověření *Správcem*, nebo přímo v roli *Správce*.

2.1.3 Co je osobní údaj

Poměrně zásadním počinem GDPR je definice osobního údaje (*Článek 4 Nařízení*). Zatímco dříve šlo o explicitně vyjádřené údaje, nyní platí, že jde o každou informaci o identifikované či identifikovatelné osobě, kdy identifikovatelnou osobou se rozumí osoba, kterou lze přímo či nepřímo identifikovat odkazem na určitý identifikátor. Z uvedené definice vyplývá, že toto obecné pojetí dává obrovský prostor ke spekulacím a nejasnostem. Vždy je třeba přihlížet k danému kontextu a k míře rizika narušení práv *Subjektu údajů*.

Některé údaje budeme za osobní považovat zcela automaticky (jméno, poštovní či e-mailová adresa, telefonní číslo apod.). Jiné vyplývají z praxe či upřesňujících vyjádření EU k legislativě (např. IP adresa uživatele v síti internet, GPS lokalizace). Další informace pak mohou být osobním údajem v určitém kontextu (lokalizace služebního vozu ve firemní garáži v případě, kdy vůz má přidělen konkrétní zaměstnanec, a je tak tedy možno jej lokalizovat, oproti situaci, kdy jeden vůz sdílí více pracovníků a odstavení ve firemní garáži tak nemůže vést k lokalizaci žádného z nich).

Jak bylo naznačeno v definici *Subjektu údajů*, je osobním údajem informace vztažená k živé fyzické osobě. *Subjektem údajů* tedy není zesnulá fyzická osoba (ačkoliv dle názoru autora této práce v tomto směru dojde k úpravě, neboť např. informace o majetkových poměrech

či zdravotním stavu právě zesnulé osoby mohou mít zásadní dopad na práva pozůstalých osob).

V rámci osobních údajů GDPR rozlišuje (*Článek 9 Nařízení*) tzv. **zvláštní kategorii** osobních údajů. Jde o totéž, co jsme dříve znali jako tzv. citlivé údaje. Ani tyto údaje nejsou taxativně vyjmenovány, nicméně legislativa určuje, o které druhy údajů se jedná. Jde o informace o rasovém či etnickém původu, politických názorech, náboženském vyznání či filosofickém přesvědčení, členství v odborech, zdravotním stavu či o sexuálním životě nebo sexuální orientaci. Rovněž sem spadají biometrické a genetické informace[4]. Smyslem existence této skupiny osobních údajů je její přísnější ochrana, kdy je zjevné, že únikem takovýchto informací dojde k významnějšímu narušení práv fyzické osoby oproti údajům jako je např. adresa bydliště.

Z definice rovněž vyplývá, že jde o informace o fyzických osobách. Na údaje o právnických osobách se tedy legislativa nevztahuje. Nicméně je třeba si uvědomit, že existence právnických osob není možná bez angažmá osob fyzických, jejichž osobní údaje již pod ochranu spadají, byť v kontextu činnosti pro právnickou osobu. Typickým příkladem může být firemní e-mail ve tvaru jmeno.prijmeni@nazevfirmy.cz. Někdy se za osobní údaj považuje i e-mail v obecném tvaru, jestliže jej spravuje pouze jedna konkrétní osoba (viz např. (Evropská unie, 2019)). V souvislosti s právnickými osobami pak může docházet k dalším složitostem při definici toho, co je osobním údajem. Např. název pracovní pozice nijak osobně nezní, nicméně srovnejme pozice administrativní pracovník úseku investic a generální ředitel. Je zjevné, že ve druhém případě to k identifikaci konkrétní fyzické osoby minimálně v prostředí dané firmy stačí. A v kombinaci s dalším „neosobním" údajem *obchodní název* už to může k identifikaci osoby postačovat i v širším okolí.

Uvedené komplikace s definicí osobního údaje je pak třeba vždy brát v souvislosti s dalším kontextem. Např. informace, že generální ředitel vydal prohlášení nijak neohrožuje práva dané fyzické osoby. Naproti tomu informace, že generální ředitel trpí určitou nemocí už do práv zasahuje velmi citelně, ačkoliv obsahem sdělení byly jen neosobní informace typu název pozice a název nemoci.

2.1.4 Co je zpracování osobních údajů

Za zpracování osobních údajů je (*Článek 4 Nařízení*) považována **jakákoliv operace** s nimi (ať už automatizovaná či bez využití IT) – tedy jejich 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.

[4] Mezi tyto údaje však nepatří rodné číslo, které bývá laickou veřejností mylně za citlivý údaj považováno.

Ke zpracování dochází pouze tehdy, kdy *Správce* zpracovává osobní údaje **systematicky** a nejde tak o nahodilý jev. Zároveň do působnosti GDPR nespadá zpracování pro osobní účely a pro určité účely související např. s činností policie apod.

2.1.5 Právní důvody zpracování

Jedním z klíčových prvků legislativy GDPR je to, že každé zpracování osobních údajů musí probíhat pouze z některého z taxativně vymezených právních důvodů (*Článek 6 Nařízení*). Těmi jsou:

Souhlas se zpracováním, který musí být svobodný, konkrétní, informovaný a jednoznačný. Je možné jej kdykoliv odvolat a v praxi by měl být využíván v situaci, kdy nelze použít jiný právní důvod. V případě odvolání souhlasu dojde ke smazání osobních údajů. Pokud jsou tytéž údaje zpracovávány zároveň z jiného důvodu, nemusí k jejich výmazu dojít.

Příkladem takové skutečnosti může být souhlas se zpracováním za účelem zasílání marketingové komunikace současně s uzavřenou smlouvou o poskytování služby (právním důvodem je zde **plnění smlouvy)**. Pokud klient odvolá souhlas, musí být vyřazen z databáze pro zasílání marketingových materiálů, ale jeho osobní údaje (v odpovídajícím rozsahu) zůstanou zpracovávány pro potřeby plnění smluvního vztahu. Je proto třeba umět pro osobní údaje rozlišit jednotlivé účely a doby zpracování ke každému účelu samostatně tak, aby bylo možno při ukončení platnosti jednoho právního důvodu vymazat třeba i část osobních údajů, které pro nadále platný účel nejsou potřeba.

Dalšími právními důvody je naplnění **právní povinnosti**, **ochrana životně důležitých zájmů** *Subjektu údajů* a při plnění **veřejného zájmu**.

Posledním právním důvodem, který spolu se souhlasem vyvolává nejasnosti při aplikaci v praxi, je **oprávněný zájem Správce**. Posouzení, co je a co není oprávněný zájem, je pak na zdravém selském rozumu a úvaze, zda je v daném případě silnější ochrana práv *Subjektu údajů* či právě zájem *Správce*. Příkladem může být často diskutované využití průmyslové kamery k zabezpečení majetku na vnějším plášti budovy částečně monitorující veřejné prostranství (viz např. (Úřad pro ochranu osobních údajů, 2019)).

2.1.6 Práva Subjektu údajů

Legislativa rovněž (*Kapitola III, Oddíl 3 Nařízení*) přesně specifikuje jednotlivá práva *Subjektu údajů* ve vztahu ke zpracování jejich osobních údajů. Tato práva se mohou významně dotýkat provozu DWH.

Jde o:

- právo na **přístup** k osobním údajům,
- právo na **opravu** či doplnění,
- právo na **výmaz**,
- právo na **omezení zpracování**,
- právo na **přenositelnost** údajů,
- právo vznést **námitku** ke zpracování,

- právo **nebýt předmětem** automatizovaného individuálního rozhodování s právními či obdobnými účinky včetně tzv. profilování.

Obsah práv na přístup, opravu a výmaz osobních údajů vyplývá již z názvu. Právo na omezení zpracování se týká dočasného pozastavení zpracování na dobu nezbytně nutnou např. v případě využití práva na vznesení námitky ke zpracování. Jinými slovy tedy *Subjekt* může namítnout, že *Správce* jeho osobní údaje zpracovává nezákonně a do doby, než bude námitka vyřízena, má právo na pozastavení takového zpracování[5]. Přenositelnost údajů znamená možnost získat údaje v elektronické standardizované podobě (např. .xml) tak, aby mohl *Subjekt údajů* tato data předat třetí straně (typicky např. při přechodu od jednoho provozovatele služby k druhému). Právo nebýt předmětem automatizovaného zpracování v podstatě ukládá povinnost *Správce* o takové skutečnosti *Subjekt údajů* informovat a na jeho žádost jej ze zpracování vyjmout. Nemůže to ovšem mít za následek významnější ohrožení např. uzavření smluvního vztahu. Na druhou stranu jsou situace, kdy to jinak není možné. Např. po poskytovateli hypotečního úvěru nelze spravedlivě požadovat, aby jednoho z žadatelů vyjmul z automatizovaného zpracování a způsobil si tak nepřiměřené riziko zpracováním manuálním.

2.1.7 Povinnosti správce

Z hlediska povinností *Správce* zde nebude uveden přehled všech povinností (detailně o nich pojednává *Kapitola IV Nařízení*). *Správce* je především povinen **reflektovat** uvedená **práva** *Subjektu údajů*. Z našeho pohledu tedy být schopen zajistit přístup k osobním údajům (tedy poskytnout přehled o tom, jaké údaje, z jakého právního důvodu, za jakým účelem a po jakou dobu zpracováváme), umožnit opravu, doplnění, výmaz či přenositelnost osobních údajů.

Námitky, přenositelnost údajů a omezení zpracování se bude spíše odehrávat již na úrovni primárních informačních systémů. Samotné **omezení zpracování** může mít dopad do DWH, ale zde bude dle názoru autora této práce převažovat zájem *Správce*, kdy technické řešení by přineslo neúměrně vysoké náklady a komplikace, zatímco přínos pro *Subjekt údajů* je zanedbatelný[6].

Významnou povinností je **ohlašovací** povinnost v případě porušení zabezpečení osobních údajů. Týká se to situací, kdy dojde k ohrožení práv *Subjektů údajů* (např. dojde ke krádeži dat z databáze). *Správce* je povinen[7] podat bez zbytečného odkladu, zpravidla do 72 hodin (pokud dojde k ohlášení později, je nutné řádně odůvodnit toto prodlení), oznámení

[5] V duchu GDPR se i zde uplatní zdravý selský rozum.
[6] Jde o to, že omezení zpracování je krátkodobé (v ideálním případě by měly být námitky vyřešeny do 1 měsíce). Navíc s osobními údaji se v DWH pracuje v pasivním módu (jsou v podstatně jen uloženy) a výstupy bývají agregované. Nad daty z DWH ani zpravidla nedochází k automatizovanému rozhodování, to je opět záležitostí primárních informačních systémů, kdy do DWH se již dostávají výsledky takového rozhodování.
[7] V závislosti na rozsahu a míře rizika pro *Subjekty údajů*.

o takové události správnímu úřadu a zároveň musí *Správce* informovat jednotlivé *Subjekty údajů* v případě, že jsou přímo ohrožena jejich práva.

Rovněž důležité je uvést, že jakékoliv předávání osobních údajů **mimo území EU** podléhá dalším specifickým podmínkám. V praxi je tedy třeba sledovat fyzické umístění dat a uvědomit si tyto souvislosti např. při zálohování či výběru fyzického umístění DWH samotného.

V souvislosti s povinnostmi *Správce* je třeba zmínit sankce v případě porušení povinností. Výše těchto sankcí byla dle názoru autora této práce primární příčinou toho, že se v době implementace GDPR zvedla velká vlna diskusí napříč celou společností, která vnímala hrozbu likvidačních sankcí např. pro neziskové organizace či malé podniky. Záměrem legislativy je však dát dostatečný prostor v uložení citelné sankce globálním korporacím zneužívajícím svého postavení. Zároveň je zjevný záměr využívat sankce až jako poslední krok v procesu řešení vzniklých problémů (po různých upomenutích, projednávání, lhůtách k odstranění nedostatků apod. – blíže viz (Evropská unie, 2019) či (Úřad pro ochranu osobních údajů, 2019)) a navíc ve výši přiměřené danému *Správci* a dané situaci. Jinak bude správní orgán nahlížet na subjekt, který vědomě a okázale ignoruje potřebu chránit osobní údaje a jinak se bude správní orgán dívat na organizaci, která pro ochranu osobních údajů udělala vše, co mohla s ohledem na své odborné a finanční možnosti. Navíc zde bude hodnocen i reálný dopad na práva *Subjektů údajů* a na jejich objem.

2.1.8 Zásady a principy

Poté, co jsme si objasnili termíny jako je osobní údaj a zpracování a ujasnili jsme si, jaké jsou právní důvody zpracování, práva *Subjektu údajů* a povinnosti *Správce osobních údajů*, můžeme se dostat k základu GDPR, kterým jsou zásady zpracování osobních údajů (*Kapitola II Nařízení*).

Zásada zákonnosti, korektnosti a transparentnosti poukazuje na to, že pro zpracování je třeba mít vždy konkrétní právní důvod a že zpracování musí být vůči *Subjektu údajů* korektní a transparentní. V praxi to znamená, že je *Subjekt údajů* vždy srozumitelně, jasně a kompletně informován o zpracování, popř. že jím poskytovaný souhlas je srozumitelný a konkrétně vymezený.

Zásada omezení účelu nad rámec právních důvodů stanoví potřebu zpracovávat údaje za konkrétním a legitimním účelem, o kterém je *Subjekt* dle předchozí zásady informován. Účely upřesňují jednotlivé právní důvody. Např. zpracování z důvodu plnění smlouvy za účelem poskytování služeb v souladu se smlouvou o poskytování telekomunikačních služeb. Takové zpracování nezakládá možnost využít osobní údaje např. k marketingovým účelům.

Zásada minimalizace údajů omezuje zpracování pouze na ty údaje, které jsou k danému účelu relevantní a nezbytné. Spolu se **zásadou omezení uložení** (tedy omezení zpracování jen na dobu nezbytně nutnou pro daný účel) tak dochází k jasné reakci na dosavadní praxi, kdy si *Správci údajů* ponechávali osobní údaje např. po ukončení smluvních vztahů, což nebylo vůči *Subjektům údajů* korektní a zvyšovalo to riziko poškození práv *Subjektu údajů*.

Zásada přesnosti vyžaduje, aby byly osobní údaje správné. Tedy nestačí jen právo *Subjektu* na opravu, ale je potřeba, aby *Správce* přiměřeně usiloval o zachování přesnosti údajů např. formou jejich pravidelné aktualizace.

Zásada integrity a důvěrnosti pak vyžaduje zajištění technických a organizačních opatření k ochraně osobních údajů. Podstatné je, aby opatření byla v rozumné míře a odpovídala míře rizika pro *Subjekt údajů*. Vždy je třeba v úvahu brát konkrétní kontext a schopnost obhájit přiměřenost opatření. Vezmeme-li příklad šifrování dat, je vhodné je použít na služebním notebooku, s nímž uživatel cestuje a využívá neznámé sítě pro připojení k internetu, zatímco je nadbytečné takové šifrování aplikovat na pevné pracovní stanice uvnitř podnikové sítě, která je chráněna řadou bezpečnostních prvků.[8]

Celkově je legislativa GDPR postavena na dvou základních principech. Tím prvním je **princip odpovědnosti Správce**, kdy je stanovena zodpovědnost *Správce* za zpracování v souladu s uvedenými zásadami a zároveň schopnost tento soulad doložit. K tomu slouží primárně záznamy o zpracování, kodexy a další dokumentace. Rovněž je klíčové, že zodpovědnosti se *Správce* nemůže zbavit přenesením na jiný subjekt (např. na *Zpracovatele* v rámci outsourcingu).

Druhým základním principem je **přístup založený na riziku**, který bývá označován výrazem *privacy by design*. Tím je míněno, že *Správce* již od samotného začátku tvorby koncepce ochrany osobních údajů musí brát v úvahu účel, kontext, povahu a rozsah zpracování a zohlednit pravděpodobnost a míru rizika pro práva *Subjektů údajů* a s odpovídající přiměřeností přistupovat k eliminaci takových rizik. Zjednodušeně by se to dalo označit povinností zohledňovat vždy práva *Subjektu údajů* a dokázat objektivně upřednostnit zájmy *Subjektů údajů* před vlastními (např. komerčními) zájmy, pakliže je to objektivně vzato správnější. V tomto směru vnímá autor této práce jako východisko etické chování, které samo o sobě může vést k naplnění uvedeného přístupu. Jestliže bude mít *Správce* sám o sobě implicitní zájem nepoškodit zájmy *Subjektu údajů* a zajistit jeho práva, nemůže se prakticky stát, že by nebyl v souladu s GDPR.

2.2 Shrnutí kapitoly

V této kapitole jsme se seznámili s jednotlivými prvky legislativy GDPR a získali jsme tak celkový přehled o jejím pojetí, a to do omezené míry detailu s ohledem na cíle a téma této práce.

Pro další kapitoly této práce, které již přinášejí konkrétní návrh aplikace GDPR do DWH, si z této kapitoly pro potřeby implementace a provozu DWH odnášíme následující:

Zodpovědnost za naplnění legislativních požadavků je vždy na *Správci osobních údajů* a nemůže ji přenést na další subjekt. Pokud tedy DWH dodává externí dodavatel, nemůže

[8] Samozřejmě je však třeba zároveň vzít v potaz, že data nejsou uvnitř firmy v bezpečí, protože velmi významným rizikem jsou vlastní zaměstnanci. To už ale nevyřeší šifrování.

převzít právní odpovědnost za porušení zpracování osobních údajů. Na druhou stranu je pravděpodobné, že součástí SLA mohou být ustanovení vázaná na tuto problematiku, kde mohou být sjednány smluvní sankce ve vazbě na porušení zásad GDPR. Externí dodavatel tedy teoreticky nemusí GDPR řešit, pakliže to nebude součástí zadání *Správce*. V praxi je však nasnadě, že zde bude zájem ze strany dodavatele DWH dodat takové řešení, které je schopné těmto požadavkům dostát. A naopak, které bude proaktivně propojeno s celopodnikovým řešením.

Má-li dojít k naplnění uvedených zásad, musí být v rámci DWH možné zobrazit zpracovávané osobní údaje, umožnit jejich aktualizaci či doplnění. Zároveň je třeba umožnit identifikaci účelu a doby zpracování daného osobního údaje a umožnit jeho smazání. Tyto skutečnosti budou nejspíše navázány na primární informační systémy, protože samotná data v DWH nedokáží determinovat účel zpracování.

Současně musí DWH poskytnut odpovídající zabezpečení dat. To je však v praxi již existující požadavek ze strany byznysu, protože jde o klíčová podniková data bez ohledu na osobní údaje. K tomu je třeba zajistit funkční proces identifikace bezpečnostních incidentů a jejich vyhodnocení, resp. spíše naplnění oznamovací povinnosti.

Pro DWH je pak rovněž důležitý závěr, že za zpracování osobních údajů je považováno i jejich samotné uložení. Takže nelze argumentovat např. tím, že data jsou uložena, ale výstupy jsou agregovány a tudíž ke zpracování nedochází.

Pravděpodobně nejtěžším úkolem bude identifikace toho, která data jsou osobním údajem a která nikoli. Zde musí jednoznačně napomoci již zdrojové informační systémy. Ovšem s ohledem na to, že za osobní údaje jsou považovány i takové informace, které vedou k identifikaci *Subjektu údajů* jen nepřímo, např. v kombinaci s jinými údaji. V tom je poměrně zásadní výzva při tvorbě DWH, protože je reálné, že z jednotlivých primárních informačních systémů přitečou data, která nebudou označena jako osobní údaje, ale osobními údaji se uvnitř DWH stanou v kombinaci s jinými „neškodnými“ daty z dalších primárních informačních systémů.

Příklad takové situace demonstruje *Tabulka 1* a *Tabulka 2*. Představme si situaci, kdy každá z těchto evidencí je zcela oddělená. Evidentně neobsahují žádné osobní údaje. V první tabulce je GPS evidence jízd referentských služebních vozů (tzn. těch, které nejsou přiděleny jednomu konkrétnímu pracovníkovi). Další tabulka ukazuje záznam z evidence využití firemní platební karty opět bez osobních údajů (vazba je přes osobní číslo zaměstnance, aniž by tato evidence obsahovala např. jméno). Pokud si představíme situaci, že tyto informace importujeme do DWH např. spolu se seznamem zaměstnanců (osobní číslo, jméno, datum nástupu apod.), nejen že dokážeme identifikovat konkrétní platební transakci ve vztahu ke konkrétnímu zaměstnanci, ale zároveň jsme schopni z kombinace údajů zjistit, kde se daný zaměstnanec v určitém okamžiku nacházel.[9]

[9] Uvedený příklad je opravdu jen ilustrační. Reálně by nemusela být jednoznačná vazba před datum, popř. by již sama evidence platebních transakcí obsahovala místo.

Tabulka 1: GPS evidence firemních vozů

RZ	Datum	Odjezd	Příjezd	Trasa
DWH 12345	10. 5. 2019	11.00	12.00	Praha - Plzeň

Zdroj: Autor

Tabulka 2: Evidence používání firemních platebních karet

Osobní číslo	Číslo karty	Datum	Účel	Částka
6789	1234 5678 9012 3456	10. 5. 2019	Tankování PHM	1.648,50

Zdroj: Autor

Uvedený příklad demonstruje to, že (nejen) při realizaci DWH je třeba uvažovat o datech v širších souvislostech. Na druhou stranu je však nutno podotknout, že je třeba k dané problematice přistupovat přiměřeně a použít zdravý selský rozum. Za prvé s ohledem na to, že nelze vše zcela domyslet, a za druhé proto, že je v souladu s principy GDPR potřeba jednat přiměřeně míře rizika a nevytvářet neúměrně komplikovaná řešení.

Uvedené závěry, které autor této práce učinil na základě interpretace informační zdrojů pro oblast GDPR specifikovaných v úvodu této kapitoly, vedou autora k obecnému návrhu při implementaci a provozu DHW, kterými se budou zabývat následující kapitoly.

3 Implementace DWH

Tato kapitola je autorovým návrhem, jak propojit ochranu osobních údajů s implementací DWH a navazuje tím na znalosti získané v online kurzu (Microsoft, 2019) na téma dodávky relačního DWH. Zároveň je inspirována obecnými principy dle (Gála, 2015), (Humpries, 2002), (Laberge, 2012) a (Lacko, 2003). Kapitola přinese pohled na způsob, jak je možné postupovat při implementaci DWH s ohledem na potřeby ochrany osobních údajů. Právě implementace je klíčovým momentem, který rozhoduje o možnosti čelit výzvám spojeným s ochranou osobních údajů.

Pokud bude této fázi věnován dostatečný prostor a pokud bude provedena s dostatečnou péčí a vhledem do problematiky ochrany osobních údajů, umožní to naplnění cílů této ochrany při budoucím provozu. Preventivní opatření jsou v tomto případě významně efektivnější než ta reaktivní. Bude-li tedy již samotná implementace DWH vedena snahou tyto cíle naplnit, může být v tomto směru považována za úspěšnou.

V této kapitole budou zmíněny jen relevantní části procesu implementace DWH. Cílem tedy není podat komplexní přehled o implementaci DWH jako takové, ale jen o těch částech, kde je ochrana osobních údajů odlišuje od zavedené praxe bez jejího zohlednění.

Obecně by měla být ochrana osobních údajů řešena na úrovni organizace jako celku. Některé z navržených procesů by tak mohly být realizovány z vně DWH v rámci procesů *Data governance* (strategické řízení dat n úrovni podniku). S ohledem na užší vymezení řešené problematiky konkrétně na DWH a vzhledem k tomu, že v praxi často podniková řešení DHW neberou v potaz, bude řešení v této práci pojato jako samostatné komplexní nezávislé řešení problematiky v rámci DWH. Cílem je ukázat takové řešení, které je plně funkční samo o sobě bez ohledu na své okolí, a které zároveň může být snadno provázáno s rozsáhlejším podnikovým řešením.

Toto pojetí principiálně odpovídá v současné době populárnímu trendu ve vývoji podnikových aplikací tzv. *Microservices architecture*, kdy se namísto jednoho monolitického řešení využívá více autonomních prvků, které jsou vzájemně propojeny. Toto pojetí navazuje na již rozšířený koncept *SOA* (Service Oriented Architecture). Blíže např. viz (Red Hat, 2019).

3.1 Vazba implementace DWH na GDPR

V této podkapitole se zaměříme na provázání teoretických základů a implementace DWH tak, jak bude následně rozebrána v následujících podkapitolách. Je zjevné, že některá ustanovení GDPR se týkají spíše implementace a některá ovlivňují hlavně provoz. Samozřejmě již ve fázi návrhu DWH je třeba vzít je všechna v úvahu, a proto si je zde nyní stručně zrekapitulujeme a vytvoříme si přehled, jak se vážou na jednotlivé následující podkapitoly týkající se implementace a provozu DWH.

Z GDPR jsme zmínili jednotlivé zásady a principy (*Podkapitola 2.1.8*), které celou problematiku zastřešují jako obecně platná pravidla. Pro celý životní cyklus DWH si musíme uvědomit, koho se GDPR týká (*Podkapitola 2.1.2*) a co je zpracování osobních údajů (*Podkapitola 2.1.4*). Následně se zejména implementace DWH dotýká definice osobního údaje (*Podkapitola 2.1.3*) a právních důvodů zpracování (*Podkapitola 2.1.5*). Do provozu DWH se nejvíce promítají práva *Subjektu údajů* (*Podkapitola 2.1.6*) a povinnosti *Správce* (*Podkapitola 2.1.7*).

Jelikož se jednotlivé zásady, principy a práva *Subjektu údajů* prolínají všemi procesy a fázemi implementace i provozu DWH, není vhodné jednoznačně stylizovat následující text do podkapitol exaktně vázaných na jednotlivá témata GDPR. Je naopak vhodnější přinést ucelený přístup k jednotlivým tématům. V *Kapitolách 3* a *4* tak budou jednotlivé vazby na dotčenou právní úpravu vždy konkrétněji specifikovány v dané podkapitole.

3.2 Porozumění datům

Za zásadní prvek při realizaci DWH je třeba považovat **porozumění datům** (Gála, 2015), (Lacko, 2009), (Microsoft, 2019). Bez toho, abychom chápali obsah dat, jejich zdroje a byznys pohled na ně, není možné navrhnout vhodnou architekturu DWH. Jak si ukážeme později, nebylo by možno bez vhledu do obsahu dat determinovat správně fakta a dimenze. A v neposlední řadě by nebylo možno zajistit korektní výstupy. Navíc ICT v moderním pojetí je o generování byznysu. Práce s daty tedy musí přinášet nové a netušené obzory a přinést výsledky daleko nad očekáváním byznysové strany.

Na počátku tedy stojí práce se zdroji dat. Musíme si udělat představu, **jaká data** máme k dispozici, a to co do **obsahu a původu**. Původ dat nám poslouží jako vodítko pro posuzování kvality, a tedy i důvěryhodnosti dat (viz *Podkapitola 2.1.8* – zásada integrity a důvěrnosti). To má přímý dopad jak na procesy transformace dat v rámci ETL, tak i celý proces *Quality data managementu*. Navíc to má jasnou souvislost s ochranou osobních údajů. Máme-li být schopni v rámci DWH identifikovat osobní údaje, je třeba je znát již na vstupu. A máme-li být schopni zajistit správnost a aktuálnost těchto osobních údajů (viz *Podkapitola 2.1.8* – zásada přesnosti), musíme znát jejich zdroj. Poznání zdrojů dat je rovněž důležité po technické stránce, aby bylo možno korektně nastavit datovou integraci a umožnit tak průběh procesů ETL.

Kromě zdroje dat je třeba hned na začátku **pochopit vazby** mezi nimi. To je směrodatné pro tvorbu datového modelu DWH (Microsoft, 2019) a architekturu databáze (rozlišení faktů a dimenzí). Vazby nám rovněž pomohou správně nastavit architekturu dimenzionálních tabulek. V souvislosti s ochranou osobních údajů je pak znalost vazeb naprosto nezbytným předpokladem pro identifikaci toho, co osobním údajem je a co není (viz *Podkapitola 2.1.3* – definice osobního údaje). Jak již bylo zmíněno dříve, určitá data mohou být osobním údajem třeba až ve spojení s jinými daty.

Z praxe i z literatury (např. (Lacko, 2009) či (Microsoft, 2019)) vyplývá, že podíl této fáze implementace je **nejnáročnější** jak z hlediska času, tak i samotné důležitosti. Před samotným zahájením implementace by tak měla proběhnout důsledná byznys analýza

stávajícího stavu, ze které bude tým realizující DWH vycházet. A zároveň je potřeba, aby samotný realizační tým (ať už interní nebo externí) věnoval tomuto kroku potřebné úsilí. Jen díky tomu může být realizace DWH provedena s ohledem na byznys požadavky a naplnit očekávání byznysu. Zároveň pochopením byznys pohledu dojde k lepšímu nastavení škálování a potenciálního rozvoje a vývoje DWH v čase tak, aby architektura odrážela nejen aktuální požadavky, ale dokázala i reagovat na nové požadavky byznysu během provozu.

Ukázkou takového přístupu může být časová dimenze. Má-li byznys aktuálně požadavek na rozpad dat na úroveň měsíců, je vhodné v této fázi odladit potenciální požadavek do budoucna. Samotné rozšíření např. na úroveň let by bylo jednoduše realizovatelné v podobě dimenze či agregace. Pokud bychom ovšem architekturu ETL procesu nastavili tak, že při transformaci dat dochází k sumarizaci faktových dat z denních záznamů na měsíční, bude nesmírně náročné v budoucnu řešit požadavek rozpadu dimenze z měsíce např. na týdny.

Řešením tohoto problému se na první pohled zdá import dat na té nejnižší úrovni, kterou máme k dispozici. To by však mohlo vést k enormnímu nárůstu objemu dat a omezovat výkon řešení. Proto je potřeba na místě v této fázi odladit např. konsolidaci dat na úroveň dní. Byznys aktuálně bude mít sumovaná data za měsíc, ale je možné snadno přejít na týdenní či denní bázi. Naproti tomu nepřetěžujeme systém daty např. na úrovni vteřin, která by v daném případě neměla reálné opodstatnění.

Uvedený příklad je samozřejmě odvislý od konkrétní situace a toho, zda konsolidace dat vůbec dává smysl. V případě bankovních transakcí na běžném účtu je základním elementem samotná transakce a nedává smysl je např. v rámci jednoho dne konsolidovat. Naproti tomu vyhodnocení prodeje osobních automobilů v autosalonu pravděpodobně bude stačit na denní bázi a nebude třeba sledovat, kolik se kterých automobilů prodalo ve které minutě.

3.3 Architektura a koncepce DWH

Základní otázkou při tvorbě architektury DWH v kontextu ochrany osobních údajů je to, zda je opravdu nutné osobní údaje do DWH ukládat (viz *Podkapitola 2.1.8* – zásada minimalizace). Tato otázka by samozřejmě měla být vyřešena na vyšší úrovni, ale s ohledem na nedostatečná řešení v praxi je vodné, aby si ji kladl i datový architekt v projektu implementace DWH. Odpověď na tuto otázku by měla vycházet z pochopení byznys požadavků a očekávání a měla by být byznysem kriticky přehodnocena.

Obvyklý přístup tvůrců DWH (Microsoft, 2019) je spíše načíst a uložit všechna data, která máme k dispozici s tím, že je třeba i pro aktuální očekávané výstupy nebudeme potřebovat. Je však na místě připomenout zásadu minimalizace údajů (*Podkapitola 2.1.8*). Dle legislativy je tedy třeba postupovat tak, abychom **zpracovávali** (a tedy v našem případě do DWH importovali) nezbytné **minimum údajů**.

Budeme-li tedy řešit např. projekt DWH pro prodej v oblasti B2C, je k diskusi, zda je potřeba do DWH spolu s daty o nákupech jednotlivých zákazníků importovat kromě zákaznického ID i další osobní údaje (např. jméno či ulice). Pokud bychom v datové analýze v tomto

případě šli až do detailu jednotlivého zákazníka, pořád pro úlohy tohoto typu může stačit výstup v podobě ID. Samotné ID je za určitých okolností osobním údajem (protože vede k identifikaci *Subjektu údajů*), ale pokud bude v rámci DWH jediným osobním údajem, významně tím napomáháme dodržení zásady minimalizace. Navíc v případě smazání ostatních osobních údajů ztrácí tuto vlastnost a přitom pomáhá zachovat konzistenci dat.

Pokud ovšem je z pohledu byznysu nutné, aby se do DWH dostávaly další osobní údaje, je nadmíru vhodné pracovat s touto skutečností již ve fázi návrhu datové struktury. Je tak možné optimalizovat mnohdy protichůdné snahy o maximalizaci výkonu (tedy co nejjednodušší datové struktury) a o přehled o osobních údajích. Tento přehled totiž s sebou nese složitější datové struktury, aby bylo možno následně s těmito osobními údaji efektivně (anebo přinejmenším vůbec nějak) pracovat.

Jednoduchým příkladem může být přehled nákupu zboží zákazníky v segmentu B2C, kdy bychom při běžném přístupu vytvořili faktovou tabulku s atributy *datum*, *jméno zákazníka*, *název zboží*, *množství*, *cena* apod. K *datumu* bychom mohli navázat dimenzionální tabulku času, k *názvu zboží* pak tabulku dimenzí typu katalog produktů. *Jméno zákazníka* by ale nedávalo smysl propojovat s dimenzionálním pohledem.

Z pohledu zpracování osobních údajů by ale v takovém případě bylo vhodné do zmíněné faktové tabulky místo *jména zákazníka* vložit *ID zákazníka*. Pokud by to pro následné datové analýzy stačilo, bylo by to nejjednodušší řešení. Pokud bychom však v DWH opravdu potřebovali i jméno zákazníka, pak by mohlo být vhodné do původní faktové tabulky zadat *ID zákazníka* a k tomu provázat další faktovou tabulku se *jménem zákazníka*.

Takové řešení se samozřejmě zdá být na první pohled nesmyslné, protože zbytečně zvyšuje nároky na výkon. Zpracování osobních údajů ale přináší potřebu nejen identifikovat, které atributy jsou osobním údajem, ale rovněž, po jakou dobu je můžeme zpracovávat. Takto vytvořená faktová tabulka s osobními údaji tak umožní doplnění o atribut *platnost do* a zároveň usnadní realizaci procesů typu náhled na osobní údaje či jejich mazání (viz *Podkapitola 2.1.6* – právo na přístup k osobním údajům a právo na výmaz). Pokud bychom totiž *platnost do* přidali do původní faktové tabulky, museli bychom takto zohledňovat všechny atributy mající charakter osobních údajů. U složitější tabulky obsahující např. i adresu, datum narození atp. by tak tabulka narůstala nejen o fakta samotná, ale i o jejich platnost. Stala by se tak nepřehlednou a její zpracování by se pak na výkonu mohlo projevit mnohem negativněji než v případě vyčlenění jednotlivých osobních údajů do samostatných faktových tabulek.

Vliv opatření pro ochranu osobních údajů na výkonnost systému je pochopitelně vždy třeba detailně uvážit pro každý daný případ. Konkrétní řešení nezáleží jen na množství takových atributů, ale i na způsobu práce s daty a očekávání ze strany byznysu.

3.4 Logické schéma DWH

Jak vyplývá z příkladu uvedeného v předchozí podkapitole, bude se logické schéma DWH zohledňující ochranu osobních údajů odlišovat od řešení, které by to v úvahu nebralo.

Rozhodujícím faktorem zde nejsou jen samotné osobní údaje, ale i **doba** jejich **platnosti**. A co více, doba platnosti může být odlišná podle důvodu zpracování (viz *Podkapitola 2.1.5* – přehled právních důvodů zpracování). Např. jméno zákazníka může mít jinou dobu platnosti z titulu souhlasu se zasíláním marketingové komunikace, jinou z titulu uzavřené kupní smlouvy a jinou z titulu oprávněného zájmu.

Návrh DWH tedy musí být schopen podchytit všechny tyto možnosti a varianty, které mohou v daném kontextu nastat. Rovněž zde hraje velkou roli to, jakým způsobem se s osobními údaji pracuje v kontextu celé organizace. Má to totiž vliv i na koncepci nad úrovní jednotlivého DWH, kdy je možno jednotlivé DHW či DM dedikovat jednotlivým právním důvodům zpracování. Vždy je však nutno si uvědomit, že potřeby byznysu musí být prioritou a teprve následně je možno optimalizovat řešení zpracování dat. Nemělo by to být obráceně, aby pohodlí při zpracování dat omezovalo byznys požadavky a jeho očekávání. Zde se však pohybujeme daleko nad rámec této práce.

Cílem návrhu takto relativně složité datové struktury vlivem ochrany osobních údajů je možnost následné **automatizace procesů** z této ochrany vyplývajících (viz *Podkapitola 2.1.6* – práva *Subjektu údajů*). A dále by řešení mělo být otevřené dalším potenciálním úpravám či změnám regulace a rovněž i odlišnému přístupu v rámci odlišné legislativy mimo EU tak, aby bylo řešení univerzálně použitelné.

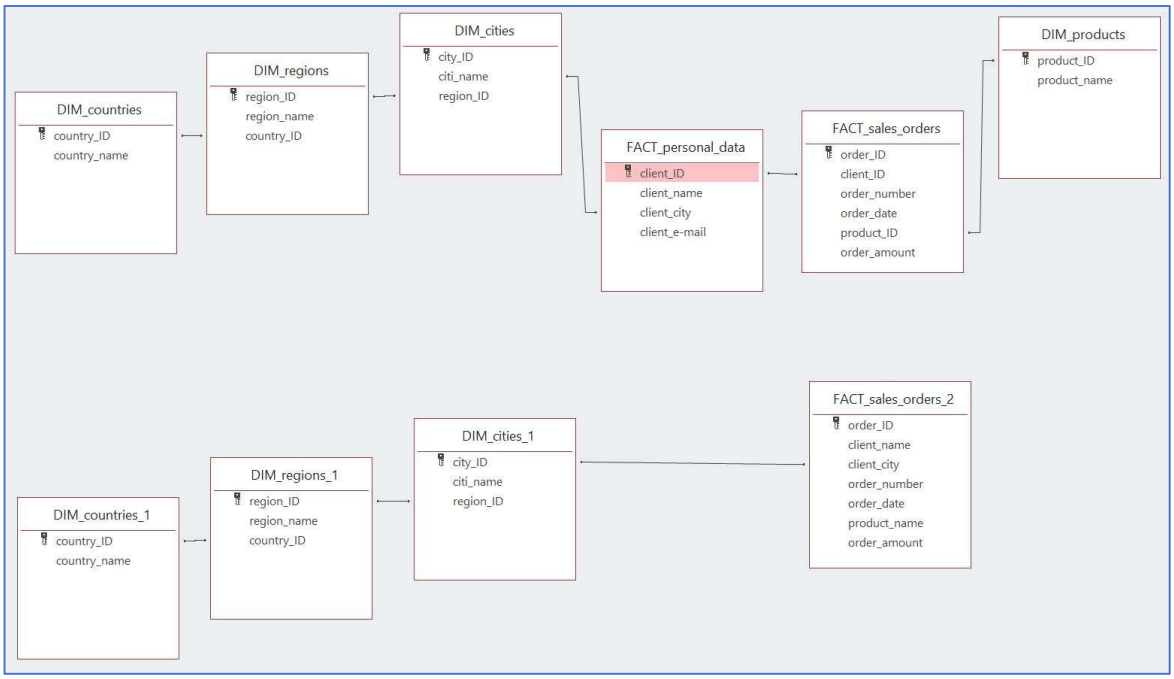
Stručnou ukázku oddělení osobních dat v rámci návrhu DWH architektury ukazuje *Obrázek 3*. Dolní část ukazuje jednu faktovou tabulku zahrnující i osobní data klientů (a zároveň s ponecháním produktů jako součásti tabulky faktů) a s náznakem řešení dimenzí (na příkladu dimenze lokality). V horní části je pak vidět rozdělení faktové tabulky do dvou částí, kdy jsou osobní údaje zahrnuty do samostatné tabulky a zároveň je demonstrováno pojetí seznamu produktů jako dimenze[10].

Je zřejmé, že u takto jednoduchého příkladu by nebyl zásadní problém uřídit správu osobních údajů zahrnutých v jedné faktové tabulce s ostatními daty. Je však nezbytné si uvědomit, že reálná praxe přináší mnohem komplikovanější situace, kde se osamostatnění osobních údajů následně ukáže jako praktické právě z důvodu další práce s nimi v čase. Hlavním důvodem pro takové řešení je totiž to, že faktové tabulky obvykle nedoznávají změn, resp. bývají v pravidelných intervalech přírůstkově doplňovány o nová fakta, zatímco osobní údaje mohou podléhat změnám i u záznamů, které již DWH historicky obsahuje. Za dosavadních podmínek tedy nebylo příliš třeba hledat mechanismy, jak měnit starší data. V kontextu ochrany osobních údajů je to však nutné (viz *Podkapitola 2.1.6* – právo na aktualizaci údajů a *Podkapitola 2.1.8* – zásada přesnosti).

Práce s těmito daty je ještě navíc zkomplikována tím, že ne vždy je žádoucí změnit všechna data. Např. v případě změny adresy je třeba umět adresu změnit jen pro určitou dobu, od kdy změna platí. S ohledem na právo *Subjektu údajů* na správnost dat si zde nevystačíme s tím, že změna nastane automaticky změnou v primárních informačních systémech.

[10] Vyčlenění produktů tímto způsobem je závislé na množství produktů a frekvenci jejich změny.

Obrázek 3: Ukázka oddělení osobních dat z hlavní faktové tabulky



Zdroj: autor

Naproti tomu v případě výmazu dat (viz *Podkapitola 2.1.6* – právo na výmaz), který bude realizován formou anonymizace dat (vysvětlíme si později v *Podkapitole 4.4.4*), musí dojít k výmazu napříč celou historií.

Další složitost pak přináší fakt, že je třeba rozlišit mezi **právními důvody** zpracování (viz *Podkapitola 2.1.5*) a dodržet při mazání dat platnost jednotlivých osobních údajů. Ne vždy totiž bude možné realizovat DWH samostatně ve vazbě na jednotlivé důvody zpracování. A ostatně takové omezení by mohlo být v praxi spíše ke škodě.

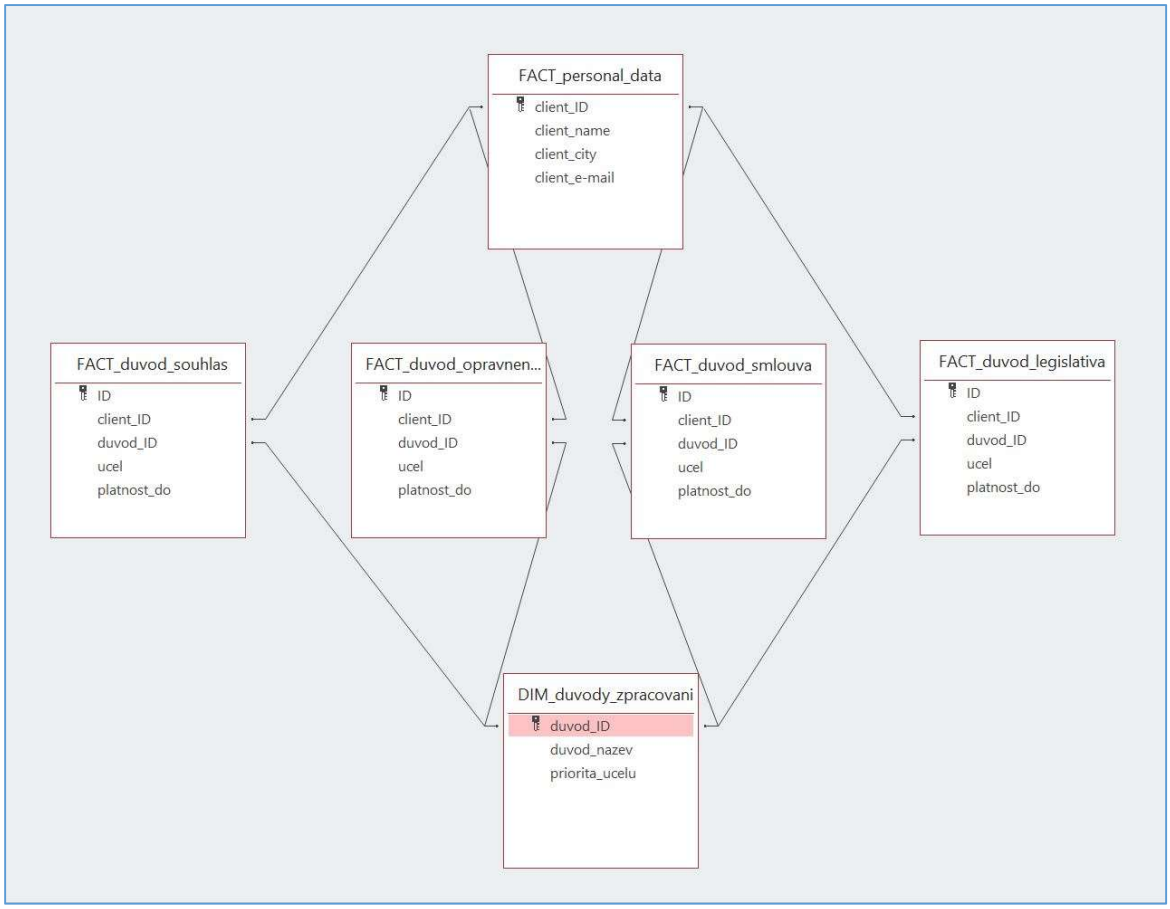
Návrh takového řešení schematicky znázorňuje *Obrázek 4*. Již dříve zmíněná faktová tabulka obsahující osobní údaje *FACT_personal_data* je doplněna o další faktové tabulky s přehledem jednotlivých důvodů zpracování a jejich platností. Vazba je realizována přes atribut *client_ID*. Faktová tabulka může obsahovat více záznamů pro každé *client_ID*, protože je možné zpracovávat osobní údaje z více důvodů zároveň. Např. může existovat situace, kdy proběhne dodávka zboží dle kupní smlouvy a její platnost tím zanikne. Na základě této skutečnosti však může dojít k dalšímu zpracování z titulu oprávněného zájmu, kdy mohou být zpracovávány osobní údaje např. po dobu možné reklamace zboží. A současně po celou dobu můžeme zpracovávat údaje třeba z titulu smlouvy o dodávce služeb. Každá z uvedených skutečností má své datum platnosti a zároveň se mohou vyskytovat současně.

Jednotlivé právní důvody zpracování vychází z dimenzionální tabulky DIM_*duvody_zpracovani*, která obsahuje i jejich prioritizaci. Pokud např. zákazník odvolá svůj souhlas se zpracováním osobních údajů, může ve stejné chvíli současně existovat silnější důvod zpracování, pro který není možno výmaz údajů provést (např. požadavky legislativy). Tyto situace může řešit právě prioritizace, která stanoví, který právní důvod má

vyšší váhu. Dále je zde i konkrétní účel zpracování. Platí, že pro jeden právní důvod můžeme mít více účelů zpracování.

Celá struktura je zároveň postavena tak, že v případě potřeby přidání nového, dosud neexistujícího právního důvodu, bude tento přidán do tabulky dimenzí a může být jednoduše přidána a implementována faktová tabulka daného důvodu. Zároveň je systém připraven na realizaci práv *Subjektu údajů* (jako např. informaci o důvodu a doby zpracování dat, výmaz apod.)

Obrázek 4: Faktové tabulky důvodů zpracování



Zdroj: autor

3.5 Master data management

Při implementaci DWH je obecně velmi užitečné pečlivě myslet dopředu a brát v úvahu nejen samotnou implementaci, ale i následné využívání a rozvoj DWH tak, aby tyto budoucí kroky byly co nejsnazší a nejefektivnější. Obdobně je to i se zohledněním ochrany osobních údajů ve fázi tvorby architektury DWH.

V obou věcech může pomoci využívání principů **MDM** (Master data management) tak, jak jej navrhuje společnost Microsoft ve svém cloudovém řešení v prostředí MS Azure (Microsoft, 2019). Stejné principy jsou v různé míře vlastní i dalším řešením od jiných

společností. Nejde zde o vyzdvihnutí jednoho konkrétního řešení, nýbrž o využití tohoto řešení jako příkladu vhodné implementace DWH s ohledem na další provoz.

V podání MS Azure je možno napojit DWH na prostředí MS Excel a umožnit tak oprávněným uživatelům přímo odtud spravovat metadata a dimenzionální tabulky. Kromě toho je však MDM poměrně rozsáhlým procesem umožňujícím práci s dimenzemi a metadaty jako takovými a umožňuje řídit DWH jako celek, ovlivňovat ETL procesy atd.

Právě pro svůj „nadhled" je tento proces **ideálním místem**, kam začlenit situace související s ochranou osobních údajů. Ta totiž vyžaduje specifické procesy, jak ostatně již vyplynulo z předchozích podkapitol. Jsou jimi správa právních důvodů zpracování (viz *Podkapitola 2.1.5*), doby zpracování (viz *Podkapitola 2.1.8* – zásada omezení uložení) a zajištění realizace práv *Subjektů údajů* (viz *Podkapitola 2.1.6* - informace o zpracování, náhled na osobní údaje, jejich aktualizace a výmaz).

Z praktického pohledu je to díky tomu, že MDM umožňuje **aplikování** různých **byznys pravidel a automatizace**. Je zde tedy možné nastavit automatizované mazání osobních údajů z titulu ukončení platnosti účelu jejich zpracování (viz *Podkapitola 2.1.8* – zásada omezení uložení). Touto aktivitou by měl systém disponovat na bázi automaticky spouštěné aktivity v relevantním časovém intervalu. Denní spouštění bude ve většině případů zbytečně frekventované a přineslo by to neúměrnou zátěž na systém. Nejvhodnější pak bude pravděpodobně měsíční periodicita. Nižší frekvence (úspora zátěže systému) už pak může být nepřiměřeným zásahem do práv *Subjektu údajů*.

V rámci byznys pravidel v procesu MDM je možné rovněž realizovat **kontrolní mechanismy** (ověřování správnosti dat) a zajistit aktualizaci údajů (viz *Podkapitola 2.1.8* – zásada přesnosti). V primárních informačních systémech by to bylo možno realizovat interakcí se *Subjektem osobních údajů*. Zde to bude spíše na bázi interakce s primárními informačními systémy. Je totiž rozdíl mezi tím, zda dochází k opravě chybně zadaného údaje, anebo ke změně v daném údaji. Jednoduchým případem je příjmení. Pokud bylo do systému zadáno chybně (překlep), má *Subjekt údajů* právo na jeho opravu napříč všemi daty včetně historie. Pokud však dochází ke změně příjmení (např. z titulu sňatku), dojde k aktualizaci tohoto údaje od daného okamžiku s tím, že historická data se upravovat nebudou. Právě to je důvodem, proč uvažovat o řešení této problematiky na úrovni procesu MDM (tedy až nad importovanými daty), namísto fáze ETL procesů nebo kdykoliv, kdy řešíme kvalitu dat. Nicméně i v rámci ETL procesů je možné tuto problematiku řešit a nejspíše nebude jedna z variant „správným" řešením. Spíše půjde o konkrétní situaci, které bude třeba tyto skutečnosti přizpůsobit.

3.6 ETL procesy

Následující podkapitoly zabývající se ETL (Extract – Transform – Load) procesy budou problematiku řešit optikou ochrany osobních údajů. Nepůjde tedy o komplexní vysvětlení problematiky, nýbrž jen o specifika kontextu této práce.

Zkratka ETL označuje (viz např. (Laberge, 2012)) procesy potřebné k naplnění DWH daty a to s akcentem na jednotlivé logické celky. Fáze *Extract* označuje činnosti spojené s napojením na zdrojové datové systémy, fáze *Transform* se zabývá automatizací převodu dat z původní struktury do podoby, jak je navržen datový sklad, včetně procesů čištění a úprav dat (tedy dříve naznačená datová kvalita). Fáze *Load* pak řeší samotný import dat, automatické spouštění aktualizací DWH ze zdrojových dat a řešení přírůstků a případně změn v datech.

3.6.1 Extrakce (Extract)

V kontextu ochrany osobních údajů bude v této fázi klíčové **provázat** datovou strukturu DWH navrženou s ohledem na ochranu osobních údajů, jak byla popsána výše, a datové struktury zdrojových databází. Bude zde klíčové identifikovat správně osobní údaje a ve zdrojových datech primární zdroje těchto dat. Je zřejmé, že atributy typu jméno klienta se budou vyskytovat ve více systémech. V lepším případě budou tyto systémy provázány a budou mít jeden zdroj dat. V horším případě půjde o datové duplicity (např. evidence zákazníků v CRM systému nebude provázána s evidencí v ERP a kromě duplicity může docházet i k rozdílům mezi daty např. vinou jejich aktualizace jen v jednom ze systémů).

3.6.2 Transformace (Transform)

Uvedenou problematiku pomohou vyřešit procesy *Transformace* dat. Právě zde je možno automatizovaně řešit duplicity či naopak rozdílný obsah „stejných“ dat (např. různá e-mailová adresa v různých systémech pro toho samého klienta).

V rámci těchto procesů dochází i k **čištění dat**, tzn. řešení chybně zadaných dat. Typicky jde o atributy s danou vstupní maskou jako je telefonní číslo (v různých systémech může být zadáno různě – bez mezinárodní předvolby, s mezerami apod.). Pokud přidáme mezinárodní předvolbu, musí zbytek vždy obsahovat jen číslice a musí jich být právě 9. Nebo např. e-mailová adresa musí vždy obsahovat znak @ a zároveň na konci tečku s doménou prvního řádu. Čísla účtu či rodná čísla by měla být dělitelná 11. Zadání adres lze kontrolovat vůči ověřeným databázím adres, názvy firem vůči patřičným rejstříkům. A tak podobně[11]. Čím více automatizovatelných úkonů, tím vyšší kvality dat můžeme dosáhnout. Cílem čištění dat je zajištění konzistence a funkcionality následných systémů.

[11] Některá z uvedených pravidel platí jen pro prostředí ČR.

Správnost dat je důležitým požadavkem z oblasti ochrany osobních údajů (viz *Podkapitola 2.1.8* – zásada přesnosti). Odhalení a odstranění chyb je tak v souladu se zájmy *Subjektu osobních údajů*. Z tohoto procesu tak těží obě strany.

3.6.3 Načtení dat (Load)

Posledním krokem jsou procesy fáze *Load* - samotné načtení dat do DWH. Z pohledu ochrany osobních údajů je možné zde řešit problematiku **aktualizace** či **oprav** již existujících **dat** (viz *Podkapitola 2.1.6* – právo na opravu či doplnění), jak bylo uvedeno výše. Bude tedy potřeba identifikovat pro každý záznam, zda již v DWH existuje, zda je totožný a pokud ne, zda se má aktualizovat.

3.7 Shrnutí kapitoly

Kapitola 3 nás provedla procesem implementace DWH z pohledu ochrany osobních údajů. Přinesla konkrétní návrhy, jak koncipovat architekturu DWH se zohledněním této problematiky. Zároveň však nevnucuje jedno konkrétní řešení, nýbrž nabádá čtenáře k zamyšlení a hledání vlastních konkrétních řešení.

Bylo zmíněno, že je v kontextu ochrany osobních údajů důležité věnovat dostatečný prostor počátečním fázím porozumění datům a architektuře DWH. Bylo ukázáno, jak přistoupit k návrhu datové struktury a byly zmíněny odlišnosti u standardních procesů ETL. Zároveň byl doplněn příklad využití procesu MDM v prostředí MS Azure (Microsoft, 2019).

V kapitole byly průběžně řešeny jak zásady a principy (viz *Podkapitola 2.1.8*), tak i práva *Subjektu údajů* (viz *Podkapitola 2.1.6*).

4 Provoz DWH

Poté, co jsme úspěšně zvládli implementaci DWH, došlo na fázi provozu. Díky zohlednění ochrany osobních údajů ve fázi přípravy a spuštění DWH bude jeho provoz s ohledem na ochranu osobních údajů výrazně snazší. I přesto jsou i zde určitě odlišnosti oproti standardním přístupům (viz např. (Humpries, 2002), (Laberge, 2012) či (Lacko, 2003)) a nové výzvy, jímž musíme kvůli ochraně osobních údajů čelit. Cílem této kapitoly je postihnout právě tyto odlišnosti od zažitých postupů. Proto obsahuje jen vybrané procesy provozu DWH a řeší je právě z úhlu pohledu ochrany osobních údajů. Stejně jako předchozí kapitola, i tato přináší autorův návrh řešení ochrany osobních údajů v DWH jako nadstavbu na znalosti získané z online kurzu o dodání DWH (Microsoft, 2019) a uvedené literatury.

4.1 Vazba provozu DWH na GDPR

Jak již bylo uvedeno v *Podkapitole 3.1*, bude i v této kapitole konkrétní provazba na právní úpravu jednotlivých skutečností uváděna průběžně v textu. Nicméně v rámci provozu DWH již nacházíme konkrétnější vazby na jednotlivá ustanovení. Jde zejména o jednotlivé procesy provozu specifikované v *Podkapitole 4.4* vztažené k jednotlivým právům *Subjektu údajů* v *Podkapitole 2.1.6*.

4.2 Tok dat do a z DWH

Jakkoliv jsme v rámci implementace automatizovali ETL procesy a nastavili byznys pravidla pro identifikaci a korektní zpracování osobních údajů, není tento proces prakticky nikdy u konce. Pokud obecně dojde ke změně struktury či obsahu dat na straně primárních informačních systémů, je nutné to reflektovat patřičnou úpravou ETL procesů. Je třeba identifikovat příčinu a obsah změn a zohlednit přitom i souvislosti s osobními údaji. Pokud tedy vznikne např. potřeba zahrnout do DWH nové atributy, zapracujeme je do patřičných faktových tabulek s ohledem na architekturu navrženého řešení. Pokud jsme realizovali řešení dle *Podkapitoly 3.4*, zapracujeme nové osobní údaje do faktové tabulky osobních údajů a nadále pracujeme v logice zvoleného řešení.

Důvodů k takovému postupu je více. Především jde o to, že na stávající logiku jsou navázány procesy práce s osobními údaji, jak si je popíšeme v dalších podkapitolách. Dalším důležitým aspektem důsledného dodržování nastavené logiky řešení je to, že jakoukoliv nesystémovou úpravou se řešení stává méně efektivní a jeho následná údržba a rozvoj je komplikovanější. V případě více takových zásahů může řešení zcela ztratit na své použitelnosti a v konečném důsledku tak zásadně ohrozit návratnost investice do DWH.

Zcela zásadní je rovněž potřeba mít **kontrolu nad datovými výstupy** z DWH. Jeho účelem totiž bývá základ pro BI a reporting. Z pohledu ochrany osobních údajů je tak nezbytné vědět, kdy, kdo a jak osobní údaje využívá (pokud vůbec). Zda taková data mohou

být pouze zobrazena uživatelem nebo zda je mu umožněn export dat. Regulace zpracování dat exportovaných z DWH již sice nemůže být předmětem procesů provozu DWH, nicméně je důležité umět poskytnout informace nejen o zpracování osobních údajů uvnitř DWH, ale i o jejich pohybu směrem k dalším informačním systémům.

4.3 Pseudonymizace a anonymizace dat

Uvedené termíny úzce souvisí s potřebou mazání osobních údajů (viz *Podkapitola 2.1.6* – právo na výmaz). Jak bylo uvedeno výše, za určitých okolností je nutné z databáze odstranit osobní údaje. Když se ale zamyslíme nad tím, jak jsou tabulky vzájemně provázané, prosté smazání řádku v seznamu klientů by nám způsobilo nemalé technické potíže.

V první řadě by mohlo dojít k narušení konzistence dat. Jednoduchým příkladem může být náš datový model, který naznačuje *Obrázek 3*. Máme samostatnou faktovou tabulku se seznamem klientů a na ní navázanou tabulku se seznamem objednávek. Vazba mezi tabulkami je realizována prostřednictvím atributu *client_ID*. Pokud bychom smazali řádek s klientem v seznamu klientů, jeho ID by přestalo existovat. V tu chvíli by ale vznikl problém v tabulce objednávek, která by odkazovala na neexistující *client_ID*. V horším případě by navíc došlo ke smazání řádků v tabulce objednávek odpovídajících danému klientovi. To by sice zachovalo integritu dat, ale přineslo by to chybné datové výstupy (např. počty a objemy objednávek by jednoduše nebyly správné).

Uvedené problémy jsou řešeny právě pseudonymizací či anonymizací dat. Tyto pojmy jsou v praxi mnohdy vnímány jako totožné, ale je mezi nimi zásadní rozdíl, který na svých rozměrech nabývá právě v kontextu zpracování osobních údajů.

Pseudonymizaci jsme vlastně již využili v našem příkladu, kdy jsme ve faktové tabulce objednávek namísto jména klienta použili jeho ID. A pokud bychom v DHW neměli tabulku s osobními údaji (výstupy bychom nechtěli až na detail jednotlivého klienta), mohli bychom mít pocit, že je vše v pořádku. Z pohledu ochrany osobních údajů tomu tak však není, protože stále existuje působ, jak ID dešifrovat a konkrétní osobu tak k objednávkám přiřadit. V našem případě je to proto, že existuje seznam klientských ID s dalšími osobními údaji.

A nevystačíme si ani s takovým postupem, kdy nastavíme algoritmus, který osobní údaj zašifruje do neidentifikovatelné podoby. Přestože nebude existovat seznam klientů, protože jejich jména budou přímo ve faktové tabulce a budou pseudonymizována algoritmem, je stále možné je s použitím daného algoritmu dešifrovat zpět a získat původní údaje. Tento způsob jistě zvyšuje bezpečí osobních údajů, protože je činí bez znalosti algoritmu nečitelné, ale z pohledu ochrany osobních údajů je to nedostatečné.

Za všeobecně uznávaný (Evropská unie, 2019) dostatečný nástroj se považuje právě **anonymizace**. Ačkoliv i zde se k úpravě osobních údajů používá určitý algoritmus, je navržen tak, aby nebylo možné zpětně osobní údaje dešifrovat. Můžeme používat náhodné generování, anebo jednoduše jméno klienta nahradíme textem určeným pro tento případ

(např. „deleted“). Jednotný řetězec pak může být výhodnější pro další zpracování, protože umožňuje např. odfiltrovat smazané uživatele.

V praxi se však ukazuje, že ani anonymizace osobních údajů nemusí být dostatečným nástrojem. Ukážeme si to na jednoduchém příkladu zaměstnance, který již ve společnosti nepracuje, a proto byl jeho záznam v evidenci parkovacích karet anonymizován (viz *Tabulka 3*) . V tabulce byly anonymizovány ty údaje, které považujeme za osobní. Když si ovšem představíme, že by takový případ skutečně nastal, nejspíše by všichni zaměstnanci společnosti věděli, o koho se jedná.

Tabulka 3: Příklad nedostatečné anonymizace osobních údajů

ID	*Jméno*	*Bydliště*	*Funkce*	*Oddělení*	*Vozidlo*	*RZ*
147	Deleted	Deleted	Deleted	TOP Management	Porsche 911	Deleted

Zdroj: Autor

Celý problém zde spočívá v kontextu. Často totiž známe více informací než jen data uvedená v předmětné databázi. A z daného kontextu jsme pak schopni *Subjekt údajů* identifikovat, i když přímo nemáme k dispozici jeho základní osobní údaje.

Velmi zajímavý příklad z praxe popsal autor této práce ve své semestrální práci k předmětu Dobývání znalostí z databází (Janoušek, 2018). Stručně shrnuto, jde o případ, kdy společnost Netflix zveřejnila (důvod není pro tento účel podstatný) korektně anonymizovanou databázi svých uživatelů a jejich hodnocení filmů. Následným srovnáním s veřejně dostupným hodnocením filmů na serveru www.imdb.com došlo k jednoznačnému spárování veřejně známých uživatelů IMDB a anonymizovaných uživatelů Netflixu a tím k odhalení jejich identity. Opět zde hrál roli kontext, ve kterém bylo možno i z anonymizovaných dat identifikovat konkrétní osoby.

Celá tato problematika tak získává zcela nový etický rozměr a není náhodou, že se etickým pohledem na zpracování dat zabývají přední světové komerční společnosti či univerzity (viz např. (KPMG, 2019), (University of Michigan, 2018), (Microsoft, 2018) apod.).

4.4 Procesy související s legislativou

V následujících podkapitolách budou podrobněji rozebrány procesy specificky vázané na ochranu osobních údajů. Vyplývají z jednotlivých práv *Subjektu údajů* (viz *Podkapitola 2.1.6*).

4.4.1 Informace o zpracování

Jednou z prvních záležitostí v souvislosti se zpracováním osobních údajů je mít přehled o jejich zpracování (viz *Podkapitola 2.1.8* – zásada transparentnosti). Není to však pouze informace o tom, jaké údaje zpracováváme, ale i proč je zpracováváme a jakou dobu je zpracovávat budeme. Sám realizátor DWH tak musí mít přehled o těchto skutečnostech

a musí být schopen tyto informace zároveň sdělit *Subjektu údajů*. Právě z tohoto důvodu je v návrhu implementace kladen důraz na separaci osobních údajů do samostatné faktové tabulky včetně informací o důvodu, účelu a době zpracování.

Navržené řešení pak umožňuje na základě jednoduchého skriptu všechny tyto informace snadno získat. Informací, jaké osobní údaje zpracováváme se v daném momentu rozumí přehled relevantních metadat. Nejlépe to vystihne ukázka takového výpisu viz *Tabulka 4*. Právní důvod vždy vyplývá z definovaných důvodů dle legislativy a účel pak konkretizuje a zpřesňuje tento důvod.

Tabulka 4: Příklad výpisu informací o zpracování

Osobní údaj	*Právní důvod*	*Účel*	*Zpracování do*
e-mail	souhlas	marketing	12. 6. 2021
e-mail	plnění smlouvy	smlouva č. 864	konec platnosti smlouvy
jméno a příjmení	plnění smlouvy	smlouva č. 864	konec platnosti smlouvy

Zdroj: Autor

V této fázi bychom rovněž měli znát i důvody a místa, kam putují osobní údaje z DWH, protože i tato informace je podstatná jak pro řízení zpracování osobních údajů, tak i pro informování *Subjektu údajů*. Typicky se bude jednat o výstupy v podobě reportingu, kde se mění způsob zpracování z uložení dat na čtení dat. Případné exporty dat je pak nutno rovněž postihnout.

Dalším rovněž běžným případem bude zálohování. Smyslem zálohování je vytvořit kopii dat mimo zálohovaný systém. Detailněji se tímto procesem budeme zabývat v *Podkapitole 4.5*.

4.4.2 Náhled na osobní údaje

Možnost náhledu na zpracovávané osobní údaje navazuje na předchozí proces. Oba tyto procesy bývají v praxi spojovány a *Subjektu údajů* bývá současně poskytnuta informace o zpracovávaných údajích včetně jejich konkrétního výpisu (viz *Podkapitola 2.1.6* – právo na přístup k osobním údajům).

Z pohledu DWH je jen třeba si uvědomit, že v tomto případě se již nejedná o metadata, nýbrž o konkrétní osobní údaje. Výstupem dotazu na zpracovávané údaje daného *Subjektu* tedy nebude informace typu „jméno, příjmení, e-mail…“, nýbrž informace typu „Jan, Janoušek, *****@gmail.com…“.

Rozdíl mezi oběma výstupy tedy jasně ukazuje, že pro získání informací je třeba použít odlišný skript. I kdyby výstupem byly společně obě informace (tedy metadata i data samotná), získáme je odlišným způsobem.

4.4.3 Správnost osobních údajů

Jedním z důležitých práv *Subjektu údajů* je právo na opravu či doplnění zpracovávaných údajů (viz *Podkapitola 2.1.6*). V případě DWH jsme již vytvořili možnost získání informací a náhledu na zpracovávané údaje. V praxi však nebude účelné umožnit opravu osobních údajů přímo v DWH. Zachycení těchto změn bude nejspíše muset zůstat na primárních informačních systémech s tím, že opravovaná data se do DWH dostanou formou nastavených ETL procesů (viz *Podkapitola 3.6.3)*.

Ačkoliv se tak zdrží oprava a mohlo by se zdát, že to tím pádem bude zásah do práv *Subjektu údajů*, je zde oprávněný zájem *Správce údajů* silnější. V prvé řadě proto, přímé zápisy do DWH neúměrně zvyšují riziko narušení konzistence a bezpečnosti dat. Zároveň samotný chybný údaj v DWH zpravidla nezpůsobuje příliš velké riziko vůči právům *Subjektu údajů*.

Lépe si to představíme na konkrétním příkladu. *Subjekt údajů* např. zjistí, že evidujeme jeho příjmení zkomolené, tato zkomolenina je zesměšňující a *Subjekt* žádá opravu. Samotná existence této chyby v DWH však nepřináší přílišné riziko, protože jsou údaje primárně využívány k vnitřním účelům (reporting apod.) a nedostávají se mimo společnost. Jiné by to např. bylo v případě primárního informačního systému, kdy bychom zkomoleninu použili pro zaslání zásilky klientovi.

Pro tento případ tedy postačí okamžitá oprava v primárních informačních systémech s tím, že se tato oprava v rámci ETL dostane do DWH, kde dojde k opravě i všech již existujících záznamů, protože oprava je platná i pro všechna historická data.

Dalším pohledem na správnost osobních údajů je jejich proměna a aktualizace v čase. Tím jsou typicky změny adres, kontaktních údajů či příjmení atp. Zde je třeba důsledně rozlišovat časové hledisko a do DWH je korektně zahrnout. Tento typ úpravy je zcela běžně v rámci implementací DWH řešen a souvislost s ochranou osobních údajů zde nepřináší potřebu změny daného řešení.

Aktualizace se řeší tak, že záznamy v databázi mají vyznačenou vlastní platnost pro záznam jako celek = řádek v tabulce (analogicky jako máme platnost jednotlivých důvodů zpracování dat). Při načtení upravených údajů se vytvoří duplicitní řádek s tím, že u původního se ukončí jeho platnost a systém tak při požadavku na aktuální data načítá novější verzi a při požadavku na historická data načítá verzi odpovídající požadovanému času. Podrobněji viz např. (Microsoft, 2019).

4.4.4 Smazání osobních údajů

Procesu mazání osobních údajů (viz *Podkapitola 2.1.6* – právo na výmaz) jsme se již dotkli v *Podkapitole 4.3*, kdy jsme identifikovali anonymizaci dat jako aktuálně nejpoužívanější způsob mazání dat.

Oproti aktualizaci dat však v tomto případě musíme naopak upravit data přímo v DWH, protože je na místě mazat i historické záznamy. Zároveň je třeba ošetřit fázi *Load* v rámci ETL (viz *Podkapitola 3.6.3*), abychom si anonymizovaná data nepřepsali původními, pokud by nebyla smazána i v primárních informačních systémech.

Proces mazání dat tak musíme doplnit o tabulku změn, kterou bude proces ETL zohledňovat. Obsah této tabulky pak bude zachovávat historii po dobu nezbytně nutnou. Tedy dokud daná změna nezačne být zohledněna i v primárních informačních systémech a zároveň s ohledem na frekvence zálohování – blíže viz *Podkapitola 4.5*.

4.5 Problematika zálohování DWH

Zálohování DWH obecně vyvolává dojem rozporu se zásadami ochrany osobních údajů, protože zde dochází ke kopírování osobních údajů mimo daný informační systém. Na první pohled se tedy zdá, že zde dochází ke zpracování např. v rozporu s obdrženým souhlasem. A zároveň, že osobní údaje jsou tímto způsobem nadále zpracovávány i poté, co měly a byly smazány na základě žádosti *Subjektu údajů*.

Je zde třeba si uvědomit několik faktů. *Správce údajů* je povinen zajistit realizaci zásady integrity a důvěrnosti dat (viz *Podkapitola 2.1.8*). To v praxi znamená, že musí chránit osobní údaje nejen před neoprávněným přístupem, ale i před smazáním. Zálohování je neoddiskutovatelně nejlepší způsob ochrany dat (a tedy i osobních údajů) před ztrátou či smazáním. Zálohu je zároveň nutné z důvodu eliminace rizik přenést mimo stávající systém (fyzické umístění). Z těchto důvodů je zjevné, že **zálohování** je naprosto **legitimní účel zpracování** dat s právním důvodem legislativního požadavku. Sama regulace ochrany osobních údajů totiž vyžaduje přiměřená opatření k zajištění práv *Subjektů údajů*.

Dalším diskutabilním bodem bývá **doba uchovávání** takové zálohy. Jistě neobstojí uchovávání záloh na dobu neurčenou. Je nezbytné stanovit lhůtu, po níž bude záloha smazána a tedy budou smazány i obsažené osobní údaje. Délka takového zpracování (doba uchovávání záloh) bude vyplývat primárně z povahy konkrétního informačního systému a daného byznysu. Jednoznačně zde převládají požadavky legislativy (tedy např. informace z mzdové agendy či o bankovních transakcích budou včetně záloh uchovávány poměrně dlouhou dobu). Naproti tomu informace obchodní povahy, které nevyžaduje legislativa, může společnost uchovávat dle svého uvážení. Pokud budou tyto informace obsahovat osobní údaje, pak by uchovávání záloh mělo být na minimální dobu s ohledem na zajištění konzistence dat. Pokud by se zde zájmy *Subjektu údajů* a *Správce* rozcházely, pak by bylo vhodné zálohovat osobní údaje odděleně a po dobu nezbytně nutnou. K tomu opět může dopomoci architektura DWH, jak byla navržena výše, kdy jsou osobní údaje vyčleněné do samostatné faktové tabulky.

Podle povahy konkrétního byznysu může být např. zálohování osobních údajů korektní v situaci, kdy se na denní bázi vytváří přírůstková záloha a na týdenní bázi záloha kompletní. Pak by se dalo považovat za dostatečné ponechat si vždy dvě poslední kompletní zálohy a přírůstkové zálohy za dny mezi nimi. Zachovávání kompletních záloh např. za dobu jednoho roku by se zdálo nepřiměřené. Nicméně i zde je možno diskutovat podle konkrétní situace daného *Správce*.

Aby byla ochrana práv *Subjektu údajů* v oblasti zálohování dotažena téměř k dokonalosti, bude vhodné zavést **logování změn**. Toto logování by mělo obsahovat informace o všech provedených změnách v mezidobí mezi jednotlivými zálohami tak, aby v případě nutnosti

obnovy dat ze zálohy bylo možno rekonstruovat všechny změny za danou dobu. Logování by zachovávalo data pouze pro daný časový úsek, data starší (zahrnutá do zálohy) by se z logu mazala.

Zálohování může probíhat různou formou (např. i paralelním provozem dvou totožných databází) a v různé frekvenci. Vždy záleží na konkrétní povaze technického řešení informačního systému a daného byznysu. V obecné rovině je však zálohování korektním nástrojem zpracování osobních údajů a přítomnost osobních údajů i po jejich smazání ve zdrojových informačních systémech je v souladu se zájmy *Subjektu údajů*. Pokud je doba zachování záloh stanovena po dobu nezbytně nutnou pro daný případ, jedná se o oprávněný zájem *Správce údajů*, který je ještě navíc motivován zajištěním práva *Subjektu údajů* na korektnost osobních údajů.

4.6 Shrnutí kapitoly

Kapitola o provozu DWH navazuje na předešlou kapitolu o jeho implementaci. V rámci provozu DWH jsme se seznámili se specifiky provozu vázanými na ochranu osobních údajů. Zmínili jsme, jak je důležité zabývat se nejen samotnými daty uvnitř DWH, ale i daty, která přitékají dovnitř, a zároveň daty, která proudí z DWH ven. Máme-li totiž na úrovni DWH přijmout zodpovědnost za správu osobních údajů, musíme mít i tyto toky pod kontrolou.

Z hlediska práce s osobními údaji (a zejména v souvislosti s jejich mazáním) jsme se dotkli tématu pseudonymizace a anonymizace dat. Tyto pojmy bývají v praxi často zaměňovány. Navíc tato záměna často vede k nedostatečnému přístupu k mazání dat.

Následně nás tato kapitola provedla jednotlivými procesy úzce spjatými se samotným zpracováním osobních údajů, a to informací o zpracování osobních údajů, náhledem na zpracovávané údaje, dále opravou osobních údajů a na závěr i jejich výmazem (viz *Podkapitola 2.1.8* – zásady a principy a dále viz *Podkapitola 2.1.6*).

Závěr kapitoly pak patřil problematice zálohování, která v praxi v souvislosti s ochranou osobních údajů vyvolává nejasnosti.

5 Případová studie

Cílem této kapitoly je navržená řešení ověřit na praktickém případu. Teoreticky je postup případové studie inspirován prací (Yin, 2009), která byla detailněji rozebrána autory z VŠE v Praze (Hrabě, 2016) či University Karlovy (Olecká, nedatováno). Tyto interpretace posloužily autorovi jako podklad pro postup této studie.

Z pohledu praktického využívá autor reálnou situaci, kdy použije skutečná data ze svého dřívějšího zaměstnání. Projektem, který autor realizoval (šlo přibližně o rok 2010), bylo vytvoření byznys architektury řešení vývoje podnikového IS na straně zákazníka. Součástí projektu byl reengineering podnikových procesů a převod dat ze stávající struktury do nové podoby dle nově vytvořených datových modelů. Autor nyní využije tohoto reálného projektu a navrhne datové modely a struktury pro implementaci do podoby DWH s ohledem na ochranu osobních údajů, který v daném reálném projektu realizován nebyl. Výsledkem případové studie tak bude praktická aplikace navržených řešení na konkrétním příkladu.

5.1 Popis výchozího stavu

Předmětem činnosti popisovaného podniku (dále jen „podnik“ či „firma“) je distribuce potravinových doplňků prostřednictvím sítě externích dealerů, kteří jsou v drtivé většině fyzickými osobami a většinou zároveň i přímými spotřebiteli produktů (od registrace právnických osob abstrahujeme). Z tohoto důvodu jsou evidováni všichni v jedné databázi jako tzv. členové, kdy jsou odlišeni atributem typu „spotřebitel“, „distributor“. Toto rozlišení je podstatné pro identifikaci cenové hladiny pro nákup produktů a pro vyplácení provizí z prodeje. Pro výpočty provizí je pak rovněž zásadní struktura této sítě, protože jednotliví distributoři jsou odměňováni dle realizovaných nákupů v rámci jejich vlastní subsítě. Každý člen je proto v síti evidován i s informací, který stávající člen je ve struktuře přímo nad ním.

Podnik má 6 zaměstnanců a veškerý provoz včetně skladových prostor je soustředěn do jedné budovy ve vlastním majetku. Pro správu IS není dedikován konkrétní pracovník a správa serveru probíhá na základě ad hoc požadavků prostřednictvím místního živnostníka. IS je provozován na vlastním fyzickém serveru umístěném v provozovně podniku. Server není nijak samostatně fyzicky zabezpečen, je součástí provozovny, která je mimo pracovní dobu sledována bezpečnostními kamerami se záznamem a zabezpečovacím zařízením s pohybovými senzory.

Jak vyplývá z popisu situace v podniku, základním informačním aktivem jsou data o členech sítě. Tyto informace jsou do podniku přijímány formou papírových dokumentů – smluv. Tyto smlouvy definují pravidla členství. Smlouva může být ukončena jednostranně ze strany podniku na základě nedostatečných nákupů člena, anebo výpovědí člena. Smlouvy jsou po ukončení dále uloženy v podniku.

Po obdržení papírové smlouvy pracovníci podniku zadají údaje ze smlouvy do IS. V IS jsou nadále osobní údaje evidovány a ad hoc na základě požadavků členů upravovány (opravy, změny adres, jmen apod.). Po ukončení platnosti smlouvy je v adresáři IS daný člen označen za vyřazeného ze sítě a veškerá data zůstávají beze změny uložena v systému.

IS je postaven na relační databázi, která obsahuje několik faktových tabulek (*t_adr* = adresář členů[12], *t_vydhlav* = přehled prodeje produktů členům – hlavičky výdejek, *t_vyd* = přehled položek jednotlivých výdejek, *t_dob* = přehled zaslaných zásilek se zbožím, *t_adrOBD* = přehled provizí za aktuální období, apod.) a tabulek dimenzí (např. *t_kz* = seznam produktů, *t_okres* = okres, *t_kraj* = kraj, apod.)

Databáze obsahuje řádově 16.000 členů sítě, z toho přibližně 5.500 aktivních. Jinými slovy, databáze obsahuje i členy, jejichž členství již vypršelo. Jsou označeni jako neaktivní, nicméně všechna jejich data v systému zůstávají. Identifikace členů a vazeb mezi nimi je realizována prostřednictvím atributu *členské číslo* (*id_adr*), který je v tabulce *t_adr* primárním klíčem.

Tabulka *t_vyd* obsahuje řádově 430.000 záznamů, zatímco hlaviček *t_vydhlav* (jednotlivých výdejek) je řádově 195.000. Z toho plyne průměrně cca 27 prodejů na jednoho člena a cca 4 produkty na jednu výdejku. V rámci datových analýz by samozřejmě bylo třeba získat další statistické hodnoty a jiné souvislosti mezi daty.

Uvedený objem dat přímo nevyzývá k řešení formou DWH, protože by bylo možno BI řešení efektivněji postavit rovnou na produkčních datech. Nicméně pro potřeby této případové studie nám to postačí. Navíc struktura dat přímo vybízí k datovým analýzám (hodnocení prodejů v čase, v místě, trendy atp.).

5.2 Přístup dle GDPR

Následující podkapitoly nás provedou klíčovými body, jak byly definovány v *Kapitole 2*.

5.2.1 Identifikace osobních údajů

Přehled osobních údajů můžeme rozdělit podle času jejich prvního výskytu v podniku:

- prvopočátkem je příjem papírové smlouvy, která obsahuje tyto údaje: rodné číslo, jméno, příjmení, titul, adresa (sídlo), adresa (korespondenční), telefon, e-mail, bankovní spojení,
- při zadání uvedených údajů do IS vznikají další charakteristiky, které mohou mít vlastnosti osobních údajů: členské číslo, členské číslo nadřazeného člena,
- během evidence se objevují další osobní údaje: výše provize.

[12] Adresář by mohl být vnímán jako dimenze. V tomto konkrétním případě však dochází k jeho změně velmi často (denně noví členové či měsíčně vyřazování členů) a navíc obsahuje na dimenze velké množství záznamů. Jako vhodnější řešení je tak zvoleno pojetí v podobě faktové tabulky.

Jak bylo uvedeno v *Podkapitole 2.1.3*, osobní údaje jsou i údaje, které mohou vést k identifikaci fyzické osoby v kombinaci s jinými údaji. S ohledem na to se můžeme zamyslet např. nad tím, proč by výše provize měla být osobním údajem. Jistě, že ve většině situací z ní identitu konkrétního člena neurčíme. Ale co třeba v situaci, kdy se ve firemním zpravodaji objeví článek o tom, že „naše síť distributorů roste a máme mezi sebou již prvního distributora, jehož příjem přesáhl 100.000 Kč"? V tomto okamžiku je již zjevné, že jakákoliv číslovka nad touto hranicí povede k jednoznačné identifikaci daného člena. A v tomto okamžiku by této identifikace mohlo být schopno všech 5.500 členů, kteří daný zpravodaj obdrželi.

Navíc jistě cítíme, že obecně jakékoliv propojení konkrétní osoby s výší jejího příjmu je velmi citlivé téma a zveřejnění příjmu by nebylo etické. Rozhodně bude výhodnější zařadit tento údaj do režimu osobních údajů a nakládat s ním obezřetně nežli riskovat poškození práv *Subjektu údajů*, které by mohlo být poměrně citelné.

Dalším příkladem k zamyšlení může být datum uzavření a případně ukončení smlouvy. Zde se zdá, že na základě data uzavření smlouvy by bylo možno identifikovat maximálně prvního člena v síti, protože jeho datum by bylo „nejstarší". Identifikaci fyzické osoby by tedy mohli opět provést všichni členové, protože jsou obeznámeni se strukturou členů v síti nad sebou. Nicméně díky tomu stejně osobní údaje znají a identifikace dle data uzavření smlouvy by vlastně nic nového neodhalila. A navíc zjištění, že daný člen je členem, jej mezi ostatními členy nijak neohrožuje a riziko ohrožení jeho práv jako *Subjektu údajů* je tak minimální. Z tohoto důvodu, a i proto, že může identifikovat prakticky jen jednoho člena sítě, není v našem případě potřeba brát datum uzavření smlouvy jako osobní údaj, ačkoli může vést k identifikaci konkrétní fyzické osoby.

V obdobném duchu bychom mohli spekulovat o řadě dalších informací. Podstatné je posoudit míru ohrožení práv *Subjektu údajů* a rozhodovat se přiměřeně (nikoliv však na úkor *Subjektu údajů*).

5.2.2 Determinace zpracování osobních údajů včetně účelů

Pro optimální realizaci je dále potřeba identifikovat, jaké zpracování (viz *Podkapitola 2.1.4*) probíhá. Resp. zda vůbec dochází ke zpracování osobních údajů.

V našem případě jsou osobní údaje **evidovány** v listinné a elektronické podobě a dále jsou v elektronické podobě **čteny** a **editovány**. Dále jsou **využívány** pro potřeby zasílání vyúčtování provizí a pro zasílání zásilek se zbožím. Dále jsou **zobrazovány** jiným členům v rámci podřízení sítě. K jejich výmazu prakticky nedochází. Ke všem uvedeným činnostem navíc dochází záměrně a **systematicky**.

Účely zpracování nebyly v rámci *Kapitoly 2* definovány. Pro zákonnost zpracování jsou nezbytné, nicméně jde o účely dané konkrétními byznys procesy podniku.

V našem případě jde o tyto účely:

- evidence v rámci členské sítě,
- výpočet a zasílání provizí,

- zasílání zásilek se zbožím,
- účetní evidence dle platné legislativy,
- zasílání marketingových materiálů.

5.2.3 Právní důvody zpracování

Na rozdíl od účelů jsou právní důvody taxativně definovány legislativou (viz *Podkapitola 2.1.5*). Pro daný účel a daný právní důvod můžeme vždy evidovat různou skupinu osobních údajů. Současně pro každý údaj platí určitá konkrétní doba zpracování, jak ukazuje na příkladu části osobních údajů *Tabulka 5*.

Tabulka 5: Příklad evidence účelů a důvodů

Osobní údaj	*Právní důvod*	*Účel*	*Zpracování do*
Číslo účtu	Plnění smlouvy	Výpočet a zasílání provizí	Ukončení platnosti smlouvy + 1 kalendářní měsíc (z důvodu vypořádání provizí)
E-mail	Plnění smlouvy	Evidence v rámci členské sítě	Ukončení platnosti smlouvy
E-mail	Plnění smlouvy	Výpočet a zasílání provizí	Ukončení platnosti smlouvy + 1 měsíc (z důvodu vypořádání provizí)
E-mail	Plnění smlouvy	Zasílání zásilek se zbožím	Ukončení platnosti smlouvy + 1 týden (z důvodu případného doručení zásilky ke konci smlouvy)
E-mail	Souhlas	Zasílání marketingových materiálů	Ukončení členství nebo odvolání souhlasu (co nastane dříve)

Zdroj: Autor

Obdobným způsobem vydefinujeme všechny kombinace jednotlivých osobních údajů, právních důvodů a účelů. To nám pomůže si vyjasnit, zda opravdu pro každý osobní údaj existuje legitimní účel a právní důvod zpracování, anebo zda údaje neevidujeme bezúčelně. Vyjasnění doby zpracování je rovněž důležité jako podklad legitimity zpracování. Tyto skutečnosti jsou užitečné jak pro řešení souladu s GDPR na úrovni podniku, tak i jako solidní základ pro datovou architekturu budovaného DWH.

5.2.4 Procesy související se zpracováním osobních údajů

Jednotlivé procesy, které budeme muset být schopni v rámci provozu DWH řešit vyplývají z práv *Subjektů údajů* (viz *Podkapitola 2.1.6*) a povinností *Správce osobních údajů* (viz *Podkapitola 2.1.7*).

Právo na **přístup** k osobním údajům zajistíme tím, že definujeme jednotlivá metadata, která v DWH budeme považovat za osobní údaje, jak jsem učinili v *Podkapitole 5.2.1*. Dalším krokem bude zpracování datové architektury tak, aby bylo možno tyto informace pro konkrétní záznam v databázi (*Subjekt údajů*) extrahovat do formy reportu. Následně zajistíme, aby bylo možno dodatečně identifikovat další osobní údaje, tzn. atributy zahrnout do režimu zpracování osobních údajů. A na závěr připravíme dotaz do databáze ve formě uloženého skriptu, který zajistí načtení jednotlivých osobních údajů pro daný záznam.

Právo na **opravu** či **doplnění** bude obecně realizováno v primárních aplikacích. Do DWH budeme tyto informace přenášet v rámci procesů ETL, kdy budou standardně nahrávána nová data a zároveň budou aktualizována data již v DWH uložená. Tento postup se v rámci ETL běžně využívá viz např. (Microsoft, 2019). Z našeho pohledu je podstatné si uvědomit rozdíl mezi opravou a aktualizací dat. Pokud např. evidujeme chybně jméno člena, pak je namístě jej opravit i v historických datech (tedy napříč všemi daty v DWH). Pokud však jde o aktualizaci dat, měla by se taková změna promítnout v nových datech, tzn. od okamžiku aktualizace. Nemůžeme zpětně opravovat třeba adresu člena v informacích o již zaslaných zásilkách. To by ohrozilo jak byznys procesy (např. řešení reklamací zásilek, reporting prodejů vztažený např. na úroveň geografické jednotky nebo třeba řádná evidence vyplacených provizí), ale zároveň by to vedlo k nekonzistenci zpracování dat a tedy i osobních údajů a tudíž by bylo v rozporu se zájmy a právy *Subjektu údajů* (porušení zásady integrity a důvěrnosti dat viz *Podkapitola 2.1.8*).

Právo na **výmaz** bude opět nejpravděpodobněji realizováno prostřednictvím primárních aplikací a z důvodu zachování integrity dat bude realizováno formou anonymizace dat. Osobní údaje budou přepsány hodnotou, z níž nebude možno zpětně původní údaje dešifrovat. V našem příkladu tak zůstane v evidenci členské číslo, které slouží jako primární klíč v adresáři členů a jako cizí klíč v tabulkách na adresář navázaných. Jeho zachováním zajistíme integritu dat. V reálné praxi sice nastane situace, kdy bude možno i z členského čísla identifikovat fyzickou osobu, protože si někdo prostě bude pamatovat[13], které osobě dané členské číslo náleželo. Z logiky struktury sítě a její provázanosti však může být nebezpečné toto číslo změnit (protože v systému existují byznys pravidla jako např. to, že členské číslo nadřízeného člena nikdy nemůže být vyšší – tzn. nadřízený člen musel být v evidenci dříve než daný člen). V tomto případě je oprávněný zájem *Správce osobních údajů* silnější než riziko vůči *Subjektu údajů*.

Právo na **omezení zpracování** může být v rámci architektury DWH řešeno poměrně jednouchým způsobem. Postačí anonymizovat osobní údaje (jako by došlo ke smazání).

[13] Autor této práce si např. i po sedmi letech po opuštění zaměstnavatele pamatuje, kdo byl evidován pod číslem 0, 1, 2, 530, 755, 1009, 1055, 10555 a je schopen tyto osoby identifikovat.

Zpracování je omezeno na určitou dobu a omezení musí proběhnout i v primárních systémech. Pokud tedy máme korektně nastavené ETL procesy, při dalším načtení dat z primárních systémů se buď zpět načtou původní data (protože omezení bylo mezitím na straně primárního systému ukončeno), anebo zůstanou data anonymizována (protože omezení na straně primárního systému zůstává v platnosti a záznam je tak vyřazen ze zpracování).

Jiná varianta je označení původního záznamu za ukončený a vytvoření nového (anonymizovaného) záznamu. Tento dočasný anonymizovaný záznam bude po skončení lhůty omezení smazán. Původní záznam bude zpět označen za platný a buď zůstane netknutý, anebo dojde k jeho smazání (anonymizaci) v souladu s výsledkem řešení důvodu omezení.

Právo na **přenositelnost** údajů bude procesně realizováno stejně jako právo na přístup k osobním údajům. Cílem zde je předat uživateli osobní údaje ve strukturované digitální podobě tak, aby je mohl využít pro předání jinému *Správci*. Pokud bychom měli zavedený systém řešení práv na úrovni celého podniku, bylo by řešení na úrovní DWH zbytečné, pokud by v rámci ETL nedocházelo k modifikaci údajů.

Právo vznést **námitku** ke zpracování není třeba nijak implementovat do DWH. Důsledkem námitky může být omezení zpracování, které jsme řešili výše.

Právo **nebýt předmětem automatizovaného individuálního rozhodování** bude na úrovni DWH řešeno pouze v případě, že jeho výstupem bude rozhodování o konkrétní fyzické osobě. Tento úkol však bude typicky probíhat na úrovni primárních aplikací, kdy budou např. rozhodovat o nově vkládaném záznamu. DWH může sloužit jako zdroj dat pro klasifikační úlohy, které mohou být využity pro individuální rozhodování.

V našem případě tedy uvážíme, zda nebude uvnitř DWH docházet k automatizovaným rozhodovacím procesům. Zkusme si představit situaci, kdy jsme např. díky aplikaci data miningu odhalili, že členové s určitými atributy obvykle nakupují určitý produkt. Nabízí se tedy nového člena pro účely zasílání marketingových materiálů zařadit do kategorie těch, které informujeme právě o daném produktu. To je typické profilování dle GDPR (viz *Podkapitola 2.1.6*). Je však třeba si uvědomit, že DWH se týká klasifikačních úloh data miningu. Samotné posouzení nového člena ale proběhne v primárním systému. Jeho profilování se tak DWH netýká.

Jak bylo uvedeno na začátku této podkapitoly, procesy se týkají jak realizace práv *Subjektů údajů*, tak i povinností *Správce osobních údajů*.

Zde je třeba připomenout, že zodpovědnost má vždy *Správce* osobních údajů bez ohledu na to, jestli technickým řešením pověří třetí stranu. V našem případě proto musíme všechny řešené skutečnosti zajistit nebo odkontrolovat, pokud bychom provoz DWH outsourcovali.

Povinnost *Správce* **informovat** o **narušení integrity** dat ať už dozorový orgán nebo *Subjekt údajů* nijak neovlivňuje návrh DWH. Standardní řešení DWH zahrnují logování

přístupů, *problem* a *incident management*[14]. Jde tedy jen o to na dané procesy napojit řešení informační povinnosti.

Další povinnosti vyplývají z dodržování zásad a principů GDPR, kterým je věnována následující podkapitola.

5.2.5 Dodržování zásad a principů

Zásady a principy jsme definovali v *Podkapitole 2.1.8*. Právě ony jsou klíčovým vodítkem k celému řešení, protože veškeré dříve rozebrané skutečnosti (např. co je a co není osobní údaj) mohou být vyřešeny jen v případě, že se budeme těmito zásadami a principy řídit.

Zásada zákonnosti, korektnosti a transparentnosti definuje to, že pro každý osobní údaj musíme mít konkrétní právní důvod zpracování, jak jsme si ukázali v předchozí podkapitole. **Zásada omezení účelu** pak k tomu doplní potřebu zpracovávat údaje za konkrétním a legitimním účelem.

Zásada minimalizace údajů v našem případě bude znamenat to, že budeme ve fázi konceptuální datového modelování DWH hledat způsob, jak přenášet minimum údajů. V našem konkrétním případě je na místě uvážit, jestli je vůbec nutné do DWH vkládat údaje jako jméno, rodné číslo, adresu, e-mail, telefon apod. Je jasné, že pro účely datových analýz to nemusí být nutné.

Pokud bychom např. chtěli analyzovat nákupy z hlediska lokality, bude nejspíše dostatečný přenos atributu *město* namísto celé adresy. E-mailová adresa není pro analytiku také potřeba. Rovněž v našem případě místo jména postačí členské číslo. To jako primární klíč musíme přenést tak jako tak. Analytika se pak dostane až na atomický detail v dimenzi jednotlivých členů. A pokud bychom přeci jen chtěli identifikovat konkrétního člena z členského čísla, můžeme si jej dohledat v primárním systému (či se zpětně dotázat na data mimo DWH).

Zásada omezení uložení usiluje o zpracování dat na dobu nezbytně nutnou pro daný účel. Pro každý osobní údaj může být více právních důvodů či účelů zpracování. Každá kombinace údaje, důvodu a účelu má svou konkrétní dobu zpracování. Tato doba musí být stanovena vždy na co nejkratší dobu.

Příklady uvádí *Tabulka 5*. Např. číslo účtu člena by mělo být zpracováváno jen po dobu platnosti smluvního vztahu. V praxi ovšem nastane situace, kdy je třeba v měsíci následujícím po ukončení smlouvy ještě vyplatit provizi za poslední měsíc. Je tedy zjevné, že je zde oprávněný zájem *Správce* evidovat číslo účtu ještě v té době. Na druhou stranu by patrně neobstálo tvrzení, že musíme evidovat číslo účtu dalších 10 let, protože se např. může stát, že později dojde k reklamaci vyplácených provizí. V takovém případě by totiž v rámci úspěšného řešení takové reklamace předal *Subjekt údajů Správci* aktuální číslo účtu pro

[14] Pojmy *problem management* a *incident management* označují procesy řešení technických potíží spojených s provozem (v našem případě) DWH. Rozdíl mezi nimi je ten, že *problem management* se zaměřuje na příčinu potíží, zatímco *incident management* na odstranění následků.

vyrovnání nedoplatku. Jiná situace by ale byla, kdyby např. uzavřená smlouva zahrnovala ustanovení o nároku reklamace provizí do určité doby po ukončení. Pak by bylo na místě tento údaj mazat až po skončení takové lhůty.

Zásada přesnosti vyzývá k udržování dat v aktuálním stavu. *Správce* tak musí dbát na to, aby v rozumné míře usiloval o aktualizaci dat. V našem případě bude vhodné nastavit pravidelnou kontrolu uložených údajů v primárních systémech (např. jednou za rok vyzvat uživatele po přihlášení do systému ke kontrole osobních údajů). Případně je to příležitost k zaslání e-mailové zprávy, kde členům vysvětlíme, že nám na ochraně jejich osobních údajů záleží a žádáme o jejich aktualizaci. Můžeme se tak legitimně připomenout i těm, kteří za poslední rok nenakoupili žádné výrobky.

Interakce se členy nebude probíhat na úrovni DWH. Tam budeme aplikovat proces *Quality data management* (viz *Podkapitola 1.2*) a budeme se zaměřovat na formální správnost zpracovávaných údajů. Tím zásadu naplníme v dostatečné míře a zároveň prospějeme i vlastnímu byznysu, protože evidence chybných údajů vede k chybám v analytice (z chybných dat vznikají chybné závěry).

Zásada integrity a důvěrnosti bude naopak velkým tématem na úrovni DWH. Řešíme zde otázku zabezpečení DWH jak kybernetické, tak i fyzické. Důležité je technické zachování integrity dat, jak jsme si jej popsali v *Podkapitole 1.1*, kde jsme řešili relace v rámci databáze (referenční, entitní a doménová integrita). Uvedená zásada koresponduje se zásadami implementace kvalitního DWH, takže při dodržení všech popsaných postupů bychom s ní neměli mít problém.

Princip odpovědnosti Správce nás v našem případě nabádá, abychom k plánování DWH přistupovali zodpovědně a po tvůrci DWH požadovali dodržení uváděných postupů a pravidel bez ohledu na to, zda se jedná o interní zaměstnance nebo o externí dodavatelskou společnost (model outsourcing). V případě externího dodavatele musíme mít zajištěné plnění uvedených parametrů a zpracování dat v souladu s GDPR smlouvou typu **SLA** (*Service Level Agreement*), která stanoví přesná kritéria poskytnutých služeb. Zároveň je nutné, abychom nastavili odpovídající kontrolní mechanismy, protože zodpovědnost za správnost zpracování osobních údajů nemůžeme v rámci SLA přenést na dodavatele a v případě problému se na něj vymlouvat. Jde o data našich členů a my vůči nim máme zodpovědnost.

Princip **přístupu založeném na riziku** s tím úzce souvisí. Pokud si na začátku jako součást strategie podniku stanovíme zájem o ochranu osobních údajů našich členů, nemělo by se nám při realizaci jakýchkoliv aktivit (včetně plánování DWH) stát, že bychom tyto údaje ohrozili, protože to bude v přímém rozporu s podnikovou strategií. Pokud se nám pak podaří dostat ochranu zájmu členů do podvědomí každého pracovníka, stane se takový přístup součástí firemní DNA a tomu se již jistě dá říkat *privacy by design*.

5.2.6 Navrhovaná změnová opatření

Na základě popsaného stávajícího stavu databáze a s ohledem na uvedená pravidla jsou navrženy následující změny v přístupu k osobním údajům.

Důkladně vyřešit fyzické umístění serveru. S ohledem na nedostatečné personální zajištění kybernetické bezpečnosti a fyzické zajištění **umístění severu** bude nejspíše vhodné využít outsourcingu a server si pronajmout. Ačkoliv tím předáme data do rukou třetí strany, díky kvalitně nastavené SLA a díky nastavení kontrolních mechanismů zajistíme, že budou data spravována daleko profesionálněji a budou uložena bezpečněji.

Zrušení evidence **rodného čísla**. Smyslem jeho dosavadní evidence byla jednoznačná identifikace člena jako subjektu smluvního vztahu. Pro tento účel postačuje datum narození. V primárních systémech dojde k transformaci pole *RC_adr* na *nar_adr*. Bude upraven datový typ pole a do DWH se tak rodná čísla vůbec nedostanou.

Dále **nebudou** do DWH pro potřebu analytiky **přenášena** tato **data**: jméno, příjmení, titul, ulice a číslo popisné (sídlo), ulice a číslo popisné (korespondenční), telefon, e-mail, bankovní spojení. Naopak ponecháme přenos údajů: členské číslo, členské číslo přímo nadřízeného člena, město a PSČ (sídlo), město a PSČ (korespondenční), výše provize. Důvodem je minimalizace zpracování.

Bude **zaveden systém výmazu** (anonymizace) osobních údajů po ukončení členské smlouvy a zároveň bude tento výmaz aplikován na aktuálně ukončené smlouvy. Současně dojde ke skartaci stávajících ukončených papírových smluv a dále budou aktuální platné smlouvy uchovávány uzamčené s omezeným přístupem zaměstnanců.

S ohledem na to, že řada členů negeneruje provize (nemají podřízené členy a jsou v evidenci jen pro účely vlastního nákupu produktů), **nebude bankovní spojení** automaticky **vyžadováno** od všech členů při registraci. Informaci o bankovním spojení poskytne člen teprve v okamžiku, kdy mu vznikne první nárok na výplatu provize.

Díky uvedeným opatřením bude DWH odpovídat požadavkům legislativy.

Konkrétní technické řešení pak již bude vycházet z uvedeného přístupu a bude realizováno v souladu s obecným postupem dle *Kapitol 3* a *4*. S ohledem na požadovaný rozsah této práce již nebude tento postup duplicitně rozepisován.

Závěr

Tato diplomová práce nás provedla možným přístupem k ochraně osobních údajů při implementaci a provozu datového skladu. Zpočátku byl představen teoretický úvod do problematiky dat a datových skladů jako takových (*Kapitola 1*) a dále do právních souvislostí zpracování osobních údajů (*Kapitola 2*). Následovaly kapitoly, kde autor přinesl vlastní návrh přístupu k řešení předmětné problematiky jak ve fázi implementace (*Kapitola 3*), tak i ve fázi provozu (*Kapitola 4*) datového skladu. Závěr práce (*Kapitola 5*) patřil případové studii, jejímž prostřednictvím autor demonstroval své návrhy na praktickém příkladu.

V následujících řádcích bude rozebráno, jak autor naplnil cíle této práce, jaké přinesl vlastní myšlenky a čím je práce inovativní. Rovněž bude provedeno srovnání s diplomovou prací (Tůma, 2017), na niž tato práce navazuje.

Naplnění cílů práce

Za **hlavní cíl** si autor vytyčil přinést **obecný návrh řešení problematiky ochrany osobních údajů na úrovni datového skladu** a to jak ve fázi implementace datového skladu, tak i během jeho následného provozu. Tohoto cíle autor dosáhl v *Kapitolách 3* a *4* tím, že z problematiky ochrany osobních údajů extrahoval relevantní nařízení (*Kapitola 2*) a propojil je s teoretickým poznáním o implementaci datových skladů a zpracování dat obecně (*Kapitola 1*). To na závěr autor doplnil případovou studií (*Kapitola 5*), která posloužila k demonstraci navrženého řešení na konkrétním praktickém příkladu.

Jednotlivé **vedlejší cíle**, jak byly uvedeny i v předchozím odstavci, byly naplněny takto:

Představit teoretické koncepce zpracování dat a tvorby DWH – v *Podkapitole 1.1* autor představil teoretický úvod ke zpracování dat, informací a znalostí, dále v *Podkapitole 1.2* představil teoretický úvod k datovým skladům a jejich architektuře a v *Podkapitole 1.3* pak koncept datového skladu v pojetí společnosti Microsoft, který se stal základem pro technické řešení použité v této práci.

Vytvořit základní přehled legislativy ve vazbě na DWH – *Podkapitola 2.1* nás postupně provedla stručnou právní historií ochrany osobních údajů, dotčených subjektů, definicí osobního údaje a zpracováním osobních údajů a dále právními důvody jejich zpracování, právy *Subjektu údajů* a povinnostmi *Správce osobních údajů*. Na závěr jsou představeny zásady a principy, na nichž celá úprava legislativy GDPR stojí.

Navrhnout koncept propojení legislativního rámce s konkrétním technickým řešením implementace a provozu DWH – v *Kapitolách 3* a *4* nás autor seznámil se svou vlastní představou o přístupu k implementaci a provozu DWH s využitím technologií společnosti Microsoft, kdy v rámci implementace navrhl specifické řešení datové architektury a v rámci provozu pak ukázal, jak lze díky této architektuře dostát požadavkům

legislativy. Navrhované řešení je zpracováno v obecné rovině tak, aby bylo aplikovatelné do praxe a nebylo specificky určené jen pro konkrétní situaci.

Demonstrovat navržený postup ve formě případové studie – *Kapitola 5* je tvořena případovou studií, díky níž autor demonstroval využití vlastního navrženého postupu na reálných datech a reálné situaci z praxe.

Vlastní přínosy práce

Svou prací se autor snaží přispět k osvětě v oblasti ochrany osobních údajů a demonstruje, že jakkoliv jde o téma komplikované, lze jej při troše dobré vůle řešit s lehkostí. Už jen etický přístup ke zpracovávaným datům je plně v souladu se zásadou GDPR *Privacy by design* a je tedy zjevné, že implicitní záměr osobní údaje chránit prakticky vede i k řešení souladu s GDPR.

Autor této práce přináší řešení na optimální úrovni detailu tak, aby bylo dostatečně obecné, ale zároveň reálně použitelné v praxi. Řešení, které rovněž přináší dostatek nadhledu a koncepční přístup. Řešení, které není závislé pouze na konkrétní technologii (ačkoliv je konkrétní technologií inspirované).

Tato práce vykresluje problematiku GDPR v odlišném světle než je všeobecně praxí vnímána a ukazuje ji jako příležitost k pozitivní změně myšlení a přístupu podniků ke svým zákazníkům a zaměstnancům. Celá legislativa se snaží metodickým přístupem nabídnout řešení ve formě principů a zásad, jejichž dodržování vede automaticky k naplnění požadavků této legislativy. Jejím záměrem není taxativní výčet povinností a příkazů a jim přináležejících trestů, nýbrž jakýsi přístup formou osvěty a nabídnutí koncepce. V tomto smyslu považuje autor této práce legislativu GDPR za přelomovou a domnívá se, že je právě proto legislativa přijímána s rozpaky. Měli bychom si však uvědomit, že současná doba již není dobou zákazů a příkazů, nýbrž dobou inovativního přístupu ve všech odvětvích, kdy hlavní roli hraje lidský a rozumný přístup založený na zdravém selském rozumu postaveném na odpovídající míře zodpovědnosti každého z nás.

To je vidět třeba na definici osobního údaje, kdy je jeho identifikace postavena na kontextu a přenáší zodpovědnost za definici na konkrétního *Správce osobních údajů*. Tím legislativa dokáže zachytit i ty nejjemnější nuance praxe, které by jinak řešila těžkopádnými obecnými definicemi, které podstatu problému zachytí jen okrajově, anebo by přinášela obšírné a komplikované definice ve snaze podchytit všechny praktické dopady. Přístup založený na přenosu zodpovědnosti na *Správce osobních údajů* tak kromě zachycení všech myslitelných detailů z praxe vede i k tomu, že nejsou *Správci* zatěžováni komplikovaným řešením, které by pro ně nebylo vhodné. Každý z nich provede právě taková opatření, která jsou optimální pro danou situaci, což vede v konečném důsledku k významnému zefektivnění zajištění souladu s legislativou v globálním měřítku.

Pojetí legislativy po každém, kdo osobní údaje zpracovává, vyžaduje, aby se zamyslel nad tím, jak s nimi nakládá. Neochota dělat věci korektně nebo neochota se vůbec nad problematikou zamýšlet pak vede k častému odmítání legislativy a jejímu vnímání jako

nadměrné byrokratické zátěže. Je to však většinou dáno nepochopením toho, že právě tato legislativa poskytuje *Správcům osobních údajů* prostor k vlastnímu optimálnímu řešení jak po stránce technické, tak i po stránce nákladů. V tomto bodě autor shledává největší neporozumění.

Toto nepochopení bylo navíc eskalováno zavedením velmi vysokých sankcí. A ačkoliv legislativa sankce uvádí jako výchovné a vždy přiměřené situaci (tedy analogicky k tomu, jak přiměřená opatření sama požaduje) a deklaruje záměr nestanovovat sankce jako likvidační, vnikla ve veřejnosti do jisté míry panika. Dle názoru autora této práce je smutné, že „zájem" o problematiku ochrany osobních údajů narostl až díky zavedení těchto razantních sankcí. Vždyť samotná ochrana osobních údajů v obdobném rozsahu platila již desítky let a GDPR je tak spíše reakcí na to, že dosavadní legislativa nebyla ze strany praxe dostatečně respektována. Díky rozvoji informačních technologií pak začalo docházet k ohrožování osobních údajů na globální úrovni a právě jako možnost citelně zasáhnout i světové hráče na trhu byly sankce stanoveny.

Díky pozitivnímu přístupu ke GDPR tak tato práce přináší ukázku toho, že se k zajištění souladu s legislativou dá přistoupit s lehkostí a řešení je mnohdy velmi jednoduché.

Přímé srovnání s (Tůma, 2017)

Jak bylo již zmíněno, tato práce svým zaměřením navazuje na diplomovou práci (Tůma, 2017). V ní autor nahlížel na řešení v rámci datových skladů v předstihu před samotnou platností legislativy. Díky tomu měl těžší úkol, protože v té době ještě praxe problematiku GDPR prakticky neřešila a autor tak musel vycházet zejména z vlastní interpretace legislativy. Jeho pojetí řešení tak odpovídá následné reakci praxe, která mnohdy řeší ochranu osobních údajů jen částečně či nekoncepčně.

Autor této diplomové práce ji naopak řeší v době, kdy je již legislativa jeden rok v platnosti. Odpor praxe k legislativě však vedl k obdobné situaci, protože v praxi jsou často zavedena jen omezená řešení, která jsou postavena na obavě ze sankcí. Na rozdíl od předešlé práce přichází tato práce s komplexnějším přístupem, který lépe podchycuje pojetí legislativy jako souboru zásad a principů. Řešení konkrétních situací souvisejících s jednotlivými právy *Subjektu údajů* je v této práci takové, aby bylo dlouhodobě udržitelné a mohlo pružně reagovat na další změny v legislativě.

Jako příklad může posloužit řešení práv *Subjektů*, kdy v předešlé práci jsou konkrétně zavedena řešení vybraných situací (např. výmaz osobních údajů) a již není vyřešeno to, že k výmazu může či nemusí dojít a že tato skutečnost závisí na právním důvodu zpracování, anebo na době zpracování. Navíc jsou v původní práci práva řešena taxativně formou přidání atributu do tabulky. V případě, že by došlo ke korekci legislativy, mohlo by být třeba přepracovat datovou strukturu a tím by vznikaly další dodatečné provozní náklady daného řešení.

Tato práce přistupuje k řešení situací koncepčně a tak, aby bylo řešení modifikovatelné a mohlo reagovat na změny bez toho, aby to vyžadovalo další významné zásahy do architektury.

Shrnutí

Tato diplomová práce je pro autora vyvrcholením studia oboru *Podniková informatika* na VŠE v Praze a slouží jako ukázka schopnosti autora přenést teoretické poznatky získané studiem do praxe.

Zároveň tato práce posloužila autorovi jako konfrontace vlastních zkušeností z praxe s teoretickými poznatky získanými studiem.

S ohledem na naplnění cílů této práce a propojení teoretických a praktických znalostí je autor s touto prací i se studiem jako takovým spokojen.

Použité informační zdroje

Evropská unie, 2019. *Web Evropské unie*. [Online]
Available at: https://europa.eu/european-union/index_cs
[Přístup získán 2019].

Gála, L., 2015. *Podniková informatika*. Praha: GRADA Publishing.

Harvard University, 2018. *CS50 Computer Science for Business Professionals*. [Online]
Available at: www.edx.com
[Přístup získán 2018].

Hrabě, P., 2016. *Přehled metod kvalitativního výzkumu v informatice.* [Online]
Available at: http://www.cssi.cz/cssi/prehled-metod-kvalitativniho-vyzkumu-v%C2%A0informatice
[Přístup získán červen 2019].

Humpries, M., 2002. *Data warehousing - návrh a implementace.* Brno: Computer Press.

Janoušek, J., 2017. *Aplikace principů podnikové informatiky v malém podnikání,* Praha: Bakalářská práce, VŠE v Praze.

Janoušek, J., 2018. *Reaálná úloha dobývání znalostí,* Praha: VŠE v Praze.

Janoušek, J., 2019. *Osobní zkušenosti z vlastní praxe autora.* Praha: Jan Janoušek.

KPMG, 2019. *Data festival.* Praha, VŠE v Przae.

Laberge, R., 2012. *Datové sklady - agilní metodiky a business intelligence*. Brno: Computer Press.

Lacko, L., 2003. *Datové sklady, analýza OLAP a dolování dat.* Brno: Computer Press.

Lacko, L., 2009. *Business Intelligence v SQL Serveru 2008*. Brno: Computer Press.

Microsoft, 2018. *Ethics and Law in Data and Analytics ,* místo neznámé: edX.org.

Microsoft, 2019. *Delivering a Relational Data Warehouse,* místo neznámé: edX.org.

Navrátil, J., 2018. *GDPR pro praxi.* Plzeň: Aleš Čeněk.

Nezmar, L., 2017. *GDPR: Praktický průvodce implementací.* Praha: Grada publishing.

Olecká, I., nedatováno *Web University Karlovy*. [Online]
Available at:
http://web.ftvs.cuni.cz/hendl/metodologie/pdfwww/oleckacasestudyclanek.pdf
[Přístup získán červen 2019].

Red Hat, 2019. *What are microservices?*. [Online]
Available at: https://www.redhat.com/en/topics/microservices/what-are-microservices
[Přístup získán duben 2019].

Škorničková, E., 2019. *Eva Škorničková*. [Online]
Available at: https://www.skornickova.eu/cz/
[Přístup získán 2019].

TechDifferences.com, 2017. *TechDifferences*. [Online]
Available at: https://techdifferences.com/difference-between-star-and-snowflake-schema.html
[Přístup získán červen 2019].

Tůma, D., 2017. *Ochrana osobních údajů v kontextu datového skladu,* Praha: Diplomová práce, VŠE v Praze.

University of Michigan, 2018. *Data Science Ethics,* místo neznámé: cursera.org.

Úřad pro ochranu osobních údajů, 2019. *GDPR sekce webu Úřadu pro ochranu osobních údajů*. [Online]
Available at: https://www.uoou.cz/gdpr-obecne-narizeni/ds-3938/p1=3938
[Přístup získán 2019].

Veber, J., 2014. *Management: základy, moderní manažerské přístupy, výkonnost a prosperita*. Praha: Management Press.

Voříšek, J. & kolektiv, 2015. *Principy a modely řízení podnikové informatiky*. Praha: Oeconomica.

Yin, R. K., 2009. *Case Study Research: design and methods*. místo neznámé:SAGE Publications.

Zentut.com, nedatováno *Kimball vs. Inmon Data Warehouse Architectures*. [Online]
Available at: http://www.zentut.com/data-warehouse/kimball-and-inmon-data-warehouse-architectures/
[Přístup získán květen 2019].

Žůrek, J., 2017. *Praktický průvodce GDPR*. Olomouc: ANAG.