TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Vytěžování a vizualizace nestrukturovaných a semistrukturovaných dat

Bakalářská práce

Studijní program: B6209 – Systémové inženýrství a informatika
Studijní obor: 6209R021 – Manažerská informatika

Autor práce: **Bára Smolová**
Vedoucí práce: Ing. Vladimíra Zádová, Ph.D.

Liberec 2016

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta
Akademický rok: **2015/2016**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bára Smolová**
Osobní číslo:	**E13000043**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Manažerská informatika**
Název tématu:	**Vytěžování a vizualizace nestrukturovaných a semistrukturovaných dat**
Zadávající katedra:	**Katedra informatiky**

Zásady pro vypracování:

1. Specifika zpracování nestrukturovaných a semistrukturovaných dat
2. Technologické platformy pro zpracování dat - charakteristika, srovnání
3. Zpracování dat na zvolené technologické platformě
4. Zhodnocení přínosu řešení

Rozsah grafických prací:

Rozsah pracovní zprávy: **30 normostran**

Forma zpracování bakalářské práce: ***tištěná/elektronická***

Seznam odborné literatury:

POUR, Jan, Miloš MARYŠKA a Ota NOVOTNÝ. Business Intelligence v podnikové praxi. Praha: Professional Publishing, 2012. ISBN 978-80-7431-065-2.
ZADROZNY, Peter. Raghu Kodali. Big data analytics using Splunk. Berkeley: Textstream, 2013. ISBN 143025761.
AGGARWAL, Charu C. Data Mining: The Textbook. Berlin: Springer International Publishing, 2015. ISBN 978-3-319-14142-8.
TAN, Michael STEINBACH a Vipin KUMAR. Introduction to Data Mining. Harlow: Pearson Education Limited, 2013. ISBN 1292026154.
Elektronická databáze článků ProQuest (knihovna.tul.cz).

Vedoucí bakalářské práce: **Ing. Vladimíra Zádová, Ph.D.**
Katedra informatiky

Konzultant bakalářské práce: **Ing. Štěpán Aubrecht**
Trask solution a. s.

Datum zadání bakalářské práce: **31. října 2015**

Termín odevzdání bakalářské práce: **31. května 2017**

doc. Ing. Miroslav Žižka, Ph.D.
děkan

Technická univerzita v Liberci
Ekonomická fakulta

L.S.

doc. Ing. Jan Skrbek, Dr.
vedoucí katedry

V Liberci dne 31. října 2015

Prohlášení

Byla jsem seznámena s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu TUL.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědoma povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Bakalářskou práci jsem vypracovala samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

Poděkování

Ráda bych tímto poděkovala Ing. Vladimíře Zádové, Ph.D., z katedry informatiky, za cenné rady, věcné připomínky a vstřícnost při konzultacích a vypracování bakalářské práce.

Mé poděkování patří též společnosti Trask solutions, a. s. za poskytnutí odborných konzultací, podkladů a cenných rad.

Anotace

Bakalářská práce se zabývá problematikou zpracování a vizualizace nestrukturovaných a semistrukturovaných dat. Cílem bakalářské práce je představit dostupné technologické platformy pro zpracování těchto dat, zvolené dvě porovnat a pomocí té vhodnější platformy zpracovat data. Důležitou částí práce je popis principů zpracování a následné vizualizace nestrukturovaných a semistrukturovaných dat. V závěru práce se nachází zhodnocení přínosu užití zvolené technologické platformy.

Klíčová slova

Big Data, Colix, dolování dat, ELK, ELK Stack, nestrukturovaná data, semistrukturovaná data, Splunk, strojová data, strukturalizace, systémové logy, vizualizace dat.

Annotation

The thesis address a problem with the data mining and data visualization of unstructured and semi-structured data. The goal of the thesis is to introduce available technological platforms for processing such data, compare selected two platforms and process data using with more appropriate platform. An important part of this work is to describe principles of processing and visualization of unstructured and semi-structured data. The final part of this thesis contains an assessment of benefits by using the selected platform.

Key Words

Big Data, Colix, data mining, data visualization, ELK, ELK Stack, machine data, semi-structured data, Splunk, structuralization, system logs, unstructured data.

Obsah

Seznam zkratek

API	Rozhraní pro programování aplikací (Application Programming Interface)
BI	Business Intelligence
CIM	Informační model, definující standardní sémantiku (Common Information Model)
CPU	Centrální výpočetní jednotka (Central Processing Unit)
EB	Exabajt (Exabyte)
EDI	Elektronická výměna dat (Electronic Data Interchange)
ER	Entitně vztahový model (Entity-relationship model)
GB	Gigabajt (Gigabyte)
GUI	Grafické uživatelské rozhraní (Graphic User Interface)
HTTP	Hypertextový přenosový protokol (Hyper Text Transfer Protocol)
IBM	Společnost International Business Machines
ICT	Informační a komunikační technologie (Information and Communication Technology)
IE	Extrakce informací (Information extraction)
ISO	Mezinárodní organizace pro normalizaci (International Organization for Standardization)
IT	Informační technologie (Information Technology)
JDK	Vývojový nástroj pro Javu (Java Development Kit)

JSON	Datový formát nezávislý na platformě (JavaScript Object Notation)
MB	Megabajt (Megabytte)
MDS	Vícerozměrné škálování (MultiDimensional Scaling)
MS	Microsoft
OS	Operační systém
PB	Petabajty (Petabyte)
RAM	Paměť (Random Access Memory)
SaaS	Software jako služba (Software as a Service)
SIEM	Management bezpečnostních informací a událostí (Security Information and Event Management)
SOM	Samo-organizující se mapy (Self-organizing map)
SPL	Dotazovací jazyk (Search Processing Language)
TUL	Technická univerzita v Liberci
XML	Rozšiřitelný značkovací jazyk (eXtensible Markup Language)
XPath	Počítačový jazyk sloužící k adresaci XML dokumentu (XML Path Language)
ZT	Zettabajt (Zettabyte)

Seznam tabulek

Seznam obrázků

Úvod

V současné době jsou informace a znalosti jednou z nejdůležitějších komodit. Jejich využití při rozhodování může vést k získání konkurenční výhody, zlepšení podnikaní, získání nového know-how a přinést další nepřeberná množství výhod.

Zdroji informací a znalostí jsou data. Díky neustálému rozvoji informačních a komunikačních technologií a jejich uplatňováním v praxi významně rostou objemy dat, mění se i jejich charakter, dostupnost. Významně se na nárůstu dat podílí Internet. Pro velký objem různorodých dat se vžil pojem Big Data. Tato data mohou být strukturovaná (databáze), nestrukturovaná (dokumenty, audio) a semistrukturovaná (XML dokumenty). V souvislosti s různorodostí dat a růstem objemu se mění požadavky na jejich zpracování.

Bakalářská práce je zaměřena především na data nestrukturovaná a semistrukturovaná. Zabývá se jejich vytěžováním a vizualizací. Hlavní cíl práce je zaměřen na charakteristiku technologických platforem pro vytěžování a vizualizaci Big Dat, které splňují v práci formulovaná kritéria. Dvě vybrané platformy jsou srovnány dle klíčových hledisek pro výběr Business Intelligence technologické platformy. Pomocí té vhodnější platformy jsou zpracována nestrukturovaná či semistrukturovaná data a zhodnoceny přínosy zvoleného řešení. Dílčím cílem práce je ukázka zpracování dat pomocí vybrané platformy.

První část práce je zaměřena na charakteristiku Big Dat a proces dolování dat (druhá kapitola) a vizualizaci nestrukturovaných a semistrukturovaných dat (třetí kapitola). Následující tři kapitoly obsahují cíle práce. Charakteristice technologických platforem je věnována čtvrtá kapitola. Srovnání technologických platforem je provedeno v páté kapitole práce. V šesté kapitole jsou zpracována nestrukturovaná data pomocí vybrané platformy.

Tato práce byla zpracována v rámci roční řízené praxe ve společnosti Trask solutions.

1. Zhodnocení současného stavu

Problematika vytěžování nestrukturovaných a semistrukturovaných dat je velmi aktuální téma. V nedávné minulosti bylo velmi složité zpracovat nestrukturovaná data. Zpracování těchto dat probíhalo obvykle manuálně.

Velký nárůst objemu všech druhů dat, především těch nestrukturovaných a semistrukturovaných, v posledních letech způsobený působením nových technologií (internet věcí, digitalizace tištěných dokumentů, přepisu telefonních hovorů, sociální sítě atd.) vyvolal potřebu tato data efektivně a s minimalizací potřebného času analyzovat. [1]

Problematika zpracování Big Dat je ale velmi komplikovaná a dosud ne zcela prozkoumaná. Jednotlivé druhy dat potřebují své individuální přístupy. Již existuje velké množství publikací, které popisují různé metody a přístupy ke zpracování nestrukturovaných a semistrukturovaných dat ale vzhledem k vysokým požadavkům na zpracování těchto dat v reálném čase a s co nepřesnějšími výsledky je nutné tyto metody a přístupy dále zdokonalovat. [2]

V současné době existuje nespočet různých metod pro analýzu nestrukturovaných a semistrukturovaných dat. Technologické platformy obvykle využívají metodu transformace nestrukturovaných a semistrukturovaných dat na data strukturovaná (tzv. metoda strukturalizace). Momentálně ale není možné tuto metodu použít pro všechny druhy nestrukturovaných a semistrukturovaných dat například pro vytěžování multimédií (tzv. Multimedia Mining). Z toho vyplývá, že ne všechny typy nestrukturovaných dat lze zpracovat stejnou metodou. [3]

Vzhledem k rostoucím nárokům na vytěžování dat, dochází k rychlému vývoji nových technologií a metod specializujících se na určité druhy dat, které nebylo donedávna možné zpracovat.

Aktuálně je velmi rozvíjena metoda dolování dat z webů tzv. Web Mining. Tyto postupy se zaměřují především na získávání dat z interakcí na sociálních sítích ((Facebook, Twitter, LinkedIn, Instagram atd.). Bohužel tato metoda má dosud svá úskalí, např. rozpoznání sarkasmu či ironie je dosud nevyřešená problematika. [4]

Je nutné poznamenat, že cílem nových technologií je i zprostředkování moderních vizualizací, které informace zobrazí ve srozumitelné, přehledné a interaktivní formě. Právě z vizualizace je lidská psychika schopna vytěžit nejvíce informací. Kombinace tvarů a barev je pro lidský mozek více pochopitelná než pouhá tabulka. [5] [6] [7]

Současné vizualizace ve velkém množství případů hraničí s uměním. Jsou vytvářeny vizualizace např. spotřeby elektrického proudu, interakce uživatelů sociálních sítí a další zajímavé grafiky. [8] [9]

V současné době je možné dolovat nejen strukturovaná ale i nestrukturovaná a semistruktuovaná data pomocí tzv. Text Miningu. Vizualizace těchto dat se v současné době také více zaměřuje na interaktivní a zajímavá grafická znázornění, která jsou znázorněna v kapitole 3. Vizualizace.

2. Big Data

S nástupem nových technologií se zásadně zvyšuje objem dat, která je potřeba zpracovat. Tato data jsou strukturovaná, semistrukturovaná, ale především nestrukturovaná. Analytici společnosti IBM odhadují, že 80–90 % všech dat jsou data nestrukturovaná nebo semistrukturovaná a jejich podíl se dle prognostik s příchodem nových technologii bude zvyšovat až o 40–60 procentních bodů ročně. [1]

Big Data je nově používaný termín, který ovšem mylně naznačuje, že se jedná pouze o „velké objemy dat“. Odkrývají problematiku strukturovaných a nestrukturovaných dat, jak externích, tak i interních, a to i z hlediska jejich analýzy v potřebném čase.

Big Data jsou často charakterizována pomocí tzv. V. V roce 2001 publikoval Doug Laney ze společnosti MetaGroup (IT společnost, která byla v roce 2005 připojená ke společnosti Gartner) studii, ve které se zabýval data managementem a zmínil zde existence tzv. 3V (Volume, Velocity, Variety). Vzhledem k neustále dramaticky rostoucí produkci dat se v dalších letech společnost IBM rozhodla definovat čtvrté V (Veracity). V současnosti je za nejvýstižnější charakteristiku Big Dat považována tzv. charakteristika 5V (Volume, Velocity, Variety, Veracity a Value). Páté V (Value) se rozhodl definovat uznávaný datový odborník Bernard Marr. Mezi dalšími charakteristikami bývá také zmiňováno Complexity, Validity, Volatility, Variability a Visualization. [10]

V bakalářské práci je charakterizováno 5V, které uvedl datový odborník Bernard Marr. Domnívám se, že tato charakteristika nejlépe vystihuje princip Big Dat.

Big Data charakterizuje:

Volume (objem) – každým rokem množství dat exponenciálně vzrůstá. Moderní technologie jsou již schopny velké objemy dat analyzovat. Abychom mohli opravdu hovořit o Big Data, musí se jednat o objemy minimálně v petabajtech. V České republice můžeme hovořit o Big Data i o terabajtech. [11]

Velocity (rychlost, rychlost změny) – důležitou charakteristikou je i rychlost, jelikož neustále narůstá počet dat, která je potřeba analyzovat v reálném čase a to zejména zásluhou

vzrůstající digitalizace většiny transakcí, mobilních zařízení a vzrůstajícímu počtu internetových uživatelů.

Variety (různorodost, variabilnost) – typickým znakem pro Big data je jejich různorodost, jelikož obsahují data strukturovaná, nestrukturovaná textová, nestrukturovaná netextová i semistrukturovaná, data o poloze a různé další multimediální typy dat. Data lze dělit i podle zdroje, ze kterého pochází, na externí a interní.

Veracity (věrohodnost) – charakteristická je také určitá nevěrohodnost v důsledku toho, že data jsou nekonzistentní, neúplná, nejasná, zkreslená. Například data získaná ze sociálních sítí.

Value (hodnota) – Big Data by také měla mít určitou hodnotu a cenu pro byznys. Společnosti zpracovávající svá interní i externí data jsou schopny získat jisté obchodní výhody, nové pohledy a fakta.

2.1 Různorodost dat

Big Data jsou různorodá, jsou rozdělena v závislosti na jejich struktuře.

Z hlediska strukturovanosti dat, rozlišujeme dvě krajní meze – **data strukturovaná** a **nestrukturovaná**. Vyskytne-li se kombinace těchto druhů dat, je označována za **semistrukturovaná data**. Datové kontinuum představuje obrázek 1.

Obrázek 1: Datové kontinuum
Zdroj: Vlastní

2.1.1 Strukturovaná data

Jedná se o data s pevně stanovenou strukturou. Za strukturovaná data se považují například databáze, tabulky, data o chování jedinců v rámci využívání webových stránek (Web Analytics), rejstříky, adresní údaje, katastry atd.

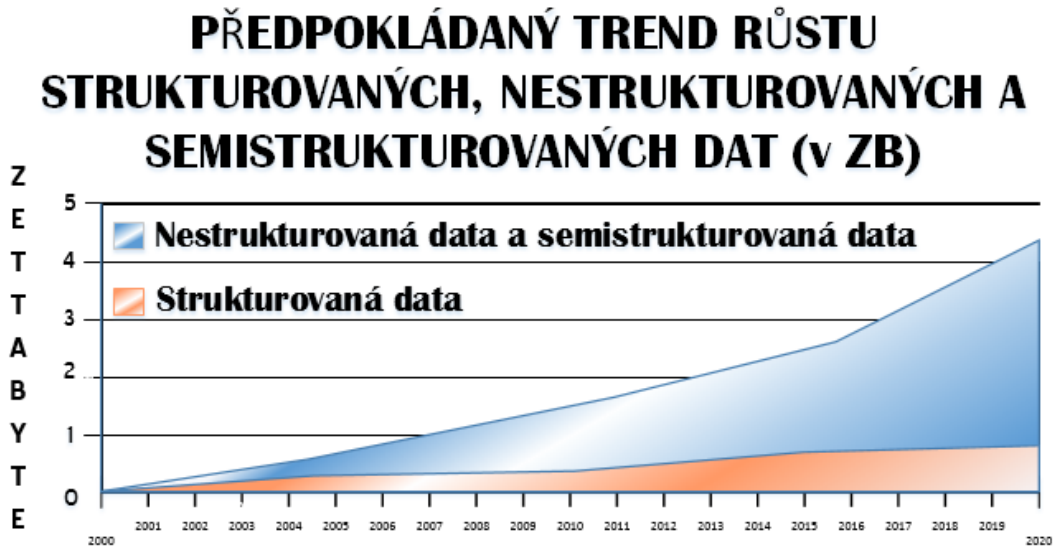
Tato data se lehce zpracovávají. Jsou uspořádána (strukturována) a je předem jasné, kde jaká data hledat. Stále se nacházejí na stejném místě, pouze mění svoji hodnotu.

2.1.2 Nestrukturovaná data

Jedná se o data volně strukturovaná nebo bez pevně stanovené struktury. Jedná se například o text, video, audio, e-maily, události, geolokační údaje a další.

V nestrukturovaných datech leží většina potenciálu pro zlepšení informační základny. Velkým zdrojem těchto dat jsou vlastní systémy, sociální sítě, webové stránky, různá multimédia a další.

Trend růstu objemu strukturovaných, nestrukturovaných a semistrukturovaných dat popisuje obrázek 2.



Obrázek 2: Trend růstu nestrukturovaných a strukturovaných dat
Zdroj: Vlastní
Zpracováno dle: [12] a http://image.slidesharecdn.com/nickyhekster-151002101805-lva1-app6891/95/big-data-expo-2015-ibm-5-predictions-7-638.jpg?cb=1443781127

Dle Molnára [13] je vhodné zmínit i rozdělení nestrukturovaných dat na data textová (tzv. pravá nestrukturovaná data) a netextová (video, audio, obrazy).

2.1.3 Semistrukturovaná data

Semistrukturovaná data (také tzv. částečně strukturovaná nebo polostrukturovaná data) jsou v problematice strukturovanosti prezentována jako data nacházející se mezi strukturovanými a nestrukturovanými daty, viz obrázek 2. Nelze je považovat za strukturovaná, jelikož nemají pevně stanovenou formální strukturu datových modelů relačních databází, ani za nestrukturovaná, protože obsahují bližší specifikace (metadata) a mají nepravidelnou strukturu. [14]

Mezi semistrukturovaná data se řadí například formáty XML, JSON, EDI a další.

Semistrukturovaná data jsou obvykle zpracována jako nestrukturovaná. Výjimku tvoří například novodobý typ XML dokumentů (od verze Word 2007), který je možné zpracovat jako strukturovaný typ dat (do relačních databází). [15]

Postupným vývojem nových technologií se objevují nové metody specializované na zpracování semistrukturovaných dat (např. Structure Mining). Tyto metody umožňují zpracovat jak obsahovou část dokumentu tak i jeho bližší specifikace a údaje vzájemně propojit.

Metadata

Podávají bližší informace o jiných datech, například u odborné publikace poskytují informace o jménu autora, nakladateli, počtu stran, vydání, edici, kontextu atd.

„Jejich prostřednictvím se lze na data dotázat, jsme schopni data doplňovat, konsolidovat je, vzájemně je synchronizovat a integrovat.“(Jak uvádí GÁLA, 2015, s. 54)

2.2 Dolování dat

Dolování dat, též známé pod výrazy vytěžování dat a Data Mining.

„Dolování dat se definuje jako analýza a zkoumání vstupních dat prostřednictvím automatických nebo částečně automatických prostředků s cílem odhalit vzájemné vztahy, pravidla, nové informace, které doposud nebyly známy nebo definovány.“ (Jak uvádí POUR, 2012, s. 205)

Definic existuje mnoho. Dolování dat je nauka o sběru, čištění, zpracování, analyzování a získávání užitečných poznatků z dat. Existuje velké množství výrazů spojených s touto naukou. Tento odborný výraz se používá jako široký „zastřešující“ termín k popisu různých aspektů zpracování dat. [2]

V současné moderní době existuje nespočet různých zařízení, která neustále generují data, obvykle nestrukturovaná, a to buď pro diagnostické či pro analytické účely. Tento fakt způsobuje prudký nárůst dat, jejich objem dosahuje k řádům Petabajtů nebo Exabajtů.

Příklady zdrojů těchto dat jsou následující:

- **Webové stránky** – množství dokumentů v rámci webových stránek je v současné době v řádech bilionů. Významnými zdroji dat jsou v současnosti i sociální sítě (Facebook, Twitter atd.). Uživatelé využíváním webových aplikací, procházením různých dokumentů a při dalších aktivitách vytvářejí logy, které jsou uloženy na serverech. Z těchto logů lze pak získat informace o chování zákazníků, poptávce atd.
- **Finanční interakce** – běžné operace s finančními prostředky, jsou součástí každodenního života, například užití bankovní karty nebo kreditní karty, využití služeb internetového bankovnictví, výběr hotovosti z bankomatu. Takovéto informace jsou uchovány a zpracovány například pro dohledání podvodů nebo zjištění neobvyklé aktivity na účtu či s kartou.
- **Interakce uživatelů** – samotná uživatelská komunikace pomocí moderních zařízení vytváří obrovské objemy dat. Například telefonní hovory vytvářejí nahrávky, které telekomunikační společnosti monitorují/ukládají i s bližšími informacemi např. o trvání hovoru a jeho přibližné poloze. Tato data slouží k analýze a k nalezení určitých vzorců chování zákazníka.
- **Senzory a internet věcí** – v současné době má téměř každý člověk chytrý telefon, tablet či jiná chytrá zařízení (např. tzv. wereables) stále sebou. Tato zařízení obsahují senzory různého druhu, poskytující informace například o poloze, o srdečním tepu, o fyzické aktivitě, kvalitě spánku, atd. Tyto senzory neustále generují data, která jsou ukládána k další analýze. [2]
- **Strojová data –** jedná se o záznamy (tzv. logy), které jsou automaticky vytvářeny velkým množstvím různorodých systémů (např. aplikacemi, podnikovými/výrobními systémy, různými servery, internetem věcí atd.). Tato data poskytují informace o chodu systému na základě událostí, které jsou v logu zaznamenány. V posledních letech dochází k prudkému nárůstu strojových dat.

Vzhledem k obrovským objemům dat produkovaných moderními technologiemi a elektronizací téměř všech aspektů současného života, je zcela přirozené zkoumat, zdali je možné z těchto dat získat poznatky.

Existuje mnoho metod dolování nestrukturovaných a semistrukturovaných dat, například:

- dolování textu;
- dolování webu;
- dolování multimédií;
- dolování toků dat;
- dolování časových řad;
- a další.

Tato bakalářská práce je zaměřena na dolování nejběžnějšího typu nestrukturovaných dat – textu. Této problematice je věnována následující kapitola.

2.2.1 Dolování textu

Dolování textu je též známé pod výrazy vytěžování textu a Text Mining.

Jedná se o proces vytěžování nestrukturovaných dat. Prostřednictvím dolování textu dochází k získávání znalostí, jeho cílem je identifikovat a analyzovat užitečné informace v textu. Tato data v textové podobě jsou jedním z nejběžnějších druhů dat, se kterými se uživatel setkává. Nachází se na webových stránkách, na sociálních sítích, v knihovnách atd.

Mezi nejdůležitější kroky vytěžování textu patří následující procesy. Ne všechny kroky musí být naplněny.

- **Kategorizace** = určování druhu textu/dokumentu.
  Text/dokument je zařazen do předem definovaných kategorií (např. kategorie informatika, politika, zdravotnictví atd.). Text/dokument může být zařazen do jedné či více kategorií, případně do žádné kategorie. Obvykle je rozdělení uskutečněno na základě četnosti slov v textu/dokumentu.
- **Shlukování**
  Slučování textů do shluků, které si jsou vzájemně co nejvíce podobné dle zadaných parametrů. Cílem je, aby si jednotlivé shluky byly co nejméně podobné. Obvykle

jeden dokument náleží jednomu shluku. Skupina dokumentů/textu ve shluku nemusí odpovídat kategoriím stanoveným během kategorizace.

- **Extrakce informací (IE)**

  Získávání strukturované nebo semistrukturované informace z nestrukturovaných dokumentů. Podúkoly extrakce informací jsou následující:

  - Rozpoznávání pojmenovaných entit – rozpoznávání objektů, které mají z pravidla stejné jméno, označení (příjmení, státy, produkty atd.).
  - Koreference – identifikace všech výrazů, které se vztahují ke stejnému subjektu v textu.

- **Sumarizace**

  Cílem sumarizace je shrnout obsah textu uživateli tak, aby mu byl v krátkém čase schopen porozumět.

- **Analýza sentimentu**

  Jedná se o detekci a extrakci subjektivně zabarvených informací z textů. Hlavním úkolem je zjistit postoj řečníka k tématu, jeho názor.

- **Určení vztahů mezi entitami**

  V případě, že lze identifikovat pojmenované entity (viz úloha extrakce informací) je možné i určit vztahy mezi nimi. [13]

Metody dolování nestrukturovaných a semistrukturovaných dat

Novodobé technologie umožňují zpracovávat nestrukturovaná data pomocí různých metod a technik založených na matematických či statistických výpočtech. Cílem práce není podrobné vysvětlení všech těchto technik, vzhledem k tomuto faktu se bakalářská práce zaměřuje pouze na stručný popis pouze některých z nich a na jejich základní principy, vlastnosti.

- **Metoda transformace nestrukturovaných a semistrukturovaných dat na data strukturovaná - Strukturalizace**

Velmi častou metodou je transformace nestrukturovaných dat na data strukturovaná například pomocí procesů normalizace, kategorizace a syntaktické analýzy. Tuto metodu využívají i všechny technologické platformy, které jsou v této práci charakterizované. Existují různé způsoby transformace na strukturovaná data, ale následující procesy jsou využívané nejčastěji.

Normalizace – data, která jsou vyjádřena různými způsoby, jsou zredukována do tzv. kanonického tvaru (do standardizovaného formátu).

Kategorizace – rozlišení dat z různých zdrojů do takzvaných kategorií pomocí tzv. značkování (tagging). Jedná se o seskupení objektů podle určitých vlastností.

Syntaktická analýza (též. parsování) – analýza posloupnosti formálních prvků, za účelem určení gramatické struktury oproti formální gramatice. Transformace textu do určité datové struktury.

- **Obsahová analýza**

Cílem této metody je identifikovat klíčová témata, termíny, které se v textu vyskytují, použitím různých statistických metod.

- **Kontextová analýza**

„Kontextová analýza umožňuje automaticky vyhodnocovat výskyty definovaných témat v plných textech zpracovávaných dokumentů.“ (Jak uvádí Molnár, 2012, s. 162)

Jedná se o kvantitativní typ analýzy dat využívaný pro analýzu nestrukturovaných dat.

- **Analýza sentimentu**

Tato metoda zjišťuje, jaký postoj autor textu zaujímá ke konkrétnímu tématu a následně ho rozdělí do jedné ze tří skupin: pozitivní, negativní a neutrální.

- **Structure Mining**

Jedná se o speciální metodu vytěžování semistrukturovaných dat, nejčastěji souborů typu XML. Používá se v případech, kdy je potřeba zpracovat jak obsah dokumentu, tak i jeho bližší specifikace (metadata atd.). K dolování těchto dat se využívá takzvaný XPath počítačový jazyk, který umožňuje vyhledávat ve struktuře XML dokumentů určitá data.

3. Vizualizace dat

Vizualizace je proces, ve kterém jsou data a informace převedeny do grafické podoby. Grafické znázornění umožňuje jednodušší pochopení zkoumaného jevu a vniknutí do problému. Mezi grafické výstupy patří například grafy (koláčový, sloupcový, spojnicový atd.), heatmapy, infografiky, diagramy a další.

Jak uvádí Černý [5] „*Vizualizaci dat je možné dobře využít také proto, že lidská psychika dokáže vytěžit mnohem více informací z vnímaných tvarů nebo barev než při čtení textu nebo při přelétnutí tabulky. A když je navíc grafika interaktivní, může uživatel lépe vystopovat nové souvislosti.*“

V současné době je datová vizualizace na prudkém vzestupu, vzhledem k následujícím faktům:

- **velké množství dat** – v roce 2012 lidská populace vyprodukovala 2,8 ZB (Zettabajtu) dat, IDC odhaduje, že v roce 2020 vyprodukujeme přibližně 40 ZB dat.
- **dostupnost dat** – tzv. open data – veřejně dostupné databáze, datasety vědeckých, státních či jiných institucí, dostupná data o chování na sociální síti a další.
- **kreativní nástroje –** nové a dostupné nástroje pro vizualizaci. [7]

3.1 Formy vizualizace dat

Vizualizovat je možné všechny běžně dostupné druhy dat. Při výběru vhodné formy vizualizace textu je nutné zohlednit formu dat, jejich množství a především také to, o jaké informace nás má daná vizualizace obohatit, jaké informace chceme získat.

Mezi základní druhy vizualizace patří grafy výsečové, sloupcové, plošné, spojnicové, různé metriky a geografické mapy.

3.2 Vizualizace nestrukturovaných dat

Nejdůležitějším krokem k získání informací z nestrukturovaných dat je jejich vizualizace.

Standardní nástroje pro vizualizaci byly nedávno vhodné pouze pro zpracování převážně strukturovaných dat, ale vzhledem ke stále narůstající potřebě společností zpracovávat nestrukturovaná data bylo nutné, aby tyto prostředky rozšířily svou vizualizační funkcionalitu. Při analýze textu je jedním z nejzákladnějších úloh vyhledávání slov, frází, souvislostí a témat textu.

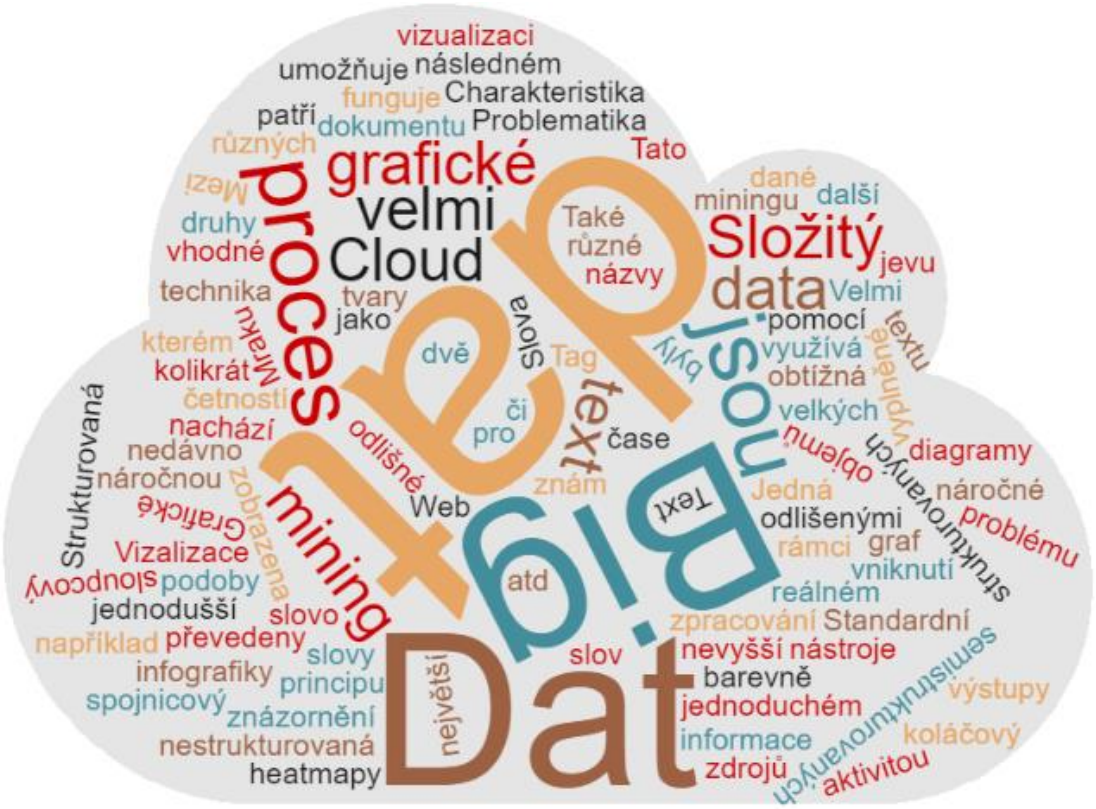
Tato kapitola je dále zaměřena na způsoby vizualizace nestrukturovaných dat – textů.

3.2.1 Vizualizace jednoho textového dokumentu

K vizualizaci nestrukturovaných dat nacházejících se v jednom textovém dokumentu, se používají následující metody:

- **Word Clouds –** také znám po názvy Tag Cloud či Text Cloud.

Jedná se o různé tvary vyplněné barevně odlišenými slovy. Tato technika funguje na jednoduchém principu – kolikrát se dané slovo nachází v rámci dokumentu. Slova s nejvyšší četností jsou i zobrazena jako největší v následném word clouds viz obrázek 3.[18]



Obrázek 3: Word Clouds (Mrak slov)
Zdroj: http://www.wordclouds.com/ Tvorba: Vlastní

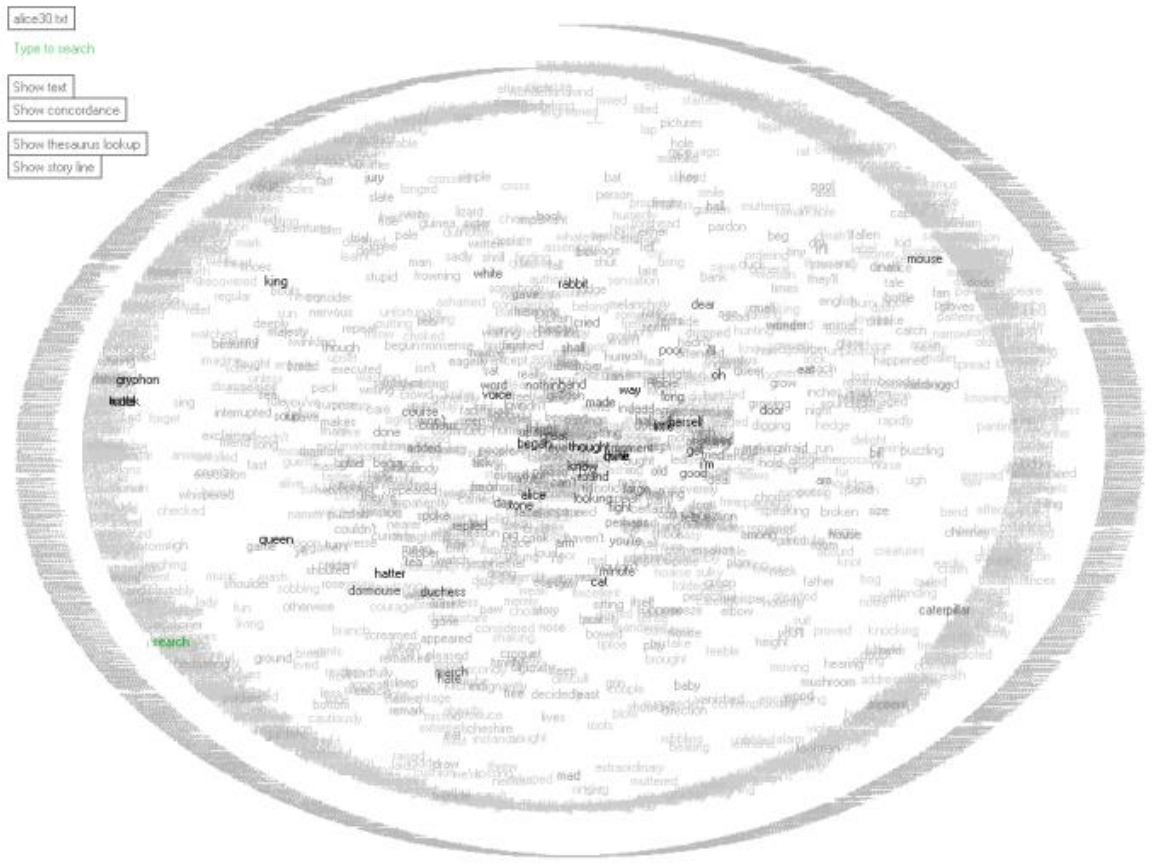
- **WordTree**

Jedná se o vizualizaci frekvence výskytu slova i jeho kontextu. Velikost slov či frází reprezentuje jeho četnost. Kořen stromu tvoří slovo či fráze, které specifikuje uživatel, jeho větve jsou kontexty, ve kterých se uživatelem specifikované slovo nebo fráze nachází. WordTree znázorňuje obrázek 4.

Obrázek 4: Word Tree (Strom slova)
Zdroj: https://www.jasondavies.com/wordtree/ Tvorba: Vlastní

- **TextArc**

Prostřednictvím této vizualizace můžeme rozšířit rozdělení slov o zobrazení jejich propojenosti (konektivity). TextArc je vizuální reprezentace toho, jak jsou jednotlivé výrazy vztažené k řadám textů, ve kterých se vyskytují. Příklad této vizualizace viz obrázek 5. [19]



Obrázek 5: TextArc
Zdroj: http://www.textarc.org/TextArcOverview.pdf

3.2.2 Vizualizace více textových dokumentů

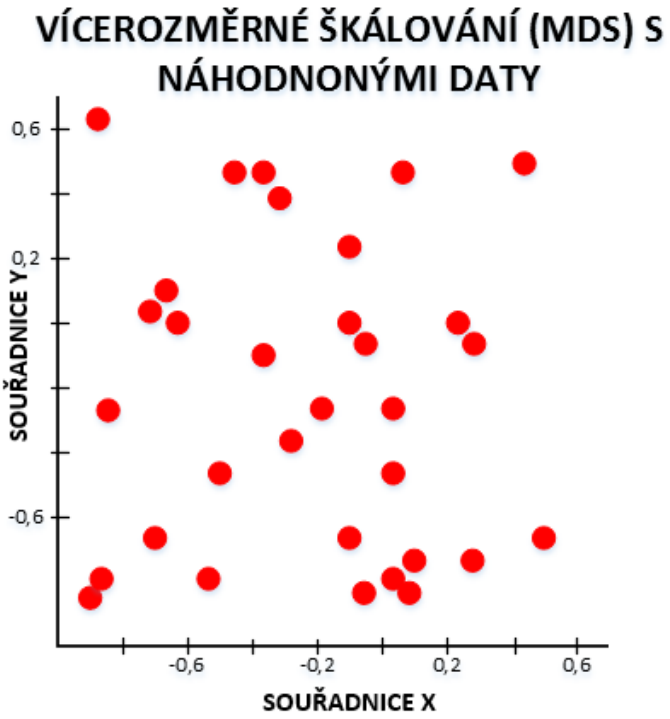
Cílem vizualizace více textových dokumentů je umisťovat pomocí charakteristik podobné dokumenty „blíže k sobě“ a ty odlišné „dále od sebe“. Podobnost je počítána mezi všemi páry dokumentů a na základě této podobnosti je poté určeno jejich rozvržení. Standardně

jsou využívány následující postupy – mnohorozměrné škálování, shluková analýza, samoorganizující se mapy a další. [19]

Vícerozměrné škálování (MDS)

„Vícerozměrné škálování objektů (MultiDimensional Scaling, MDS) je technika vytvoření subjektivní mapy relativního umístění objektů v rovině dvojrozměrného grafu, a to na základě vzdáleností či podobností mezi objekty, tzv. matice proximity (blízkosti).“ (Jak uvádí Meloun, 2011, s. 75)

Za objekty jsou považovány např. názory, kandidáti, produkty, různé hodnoty apod. Vícerozměrné škálování zobrazuje obrázek 6.

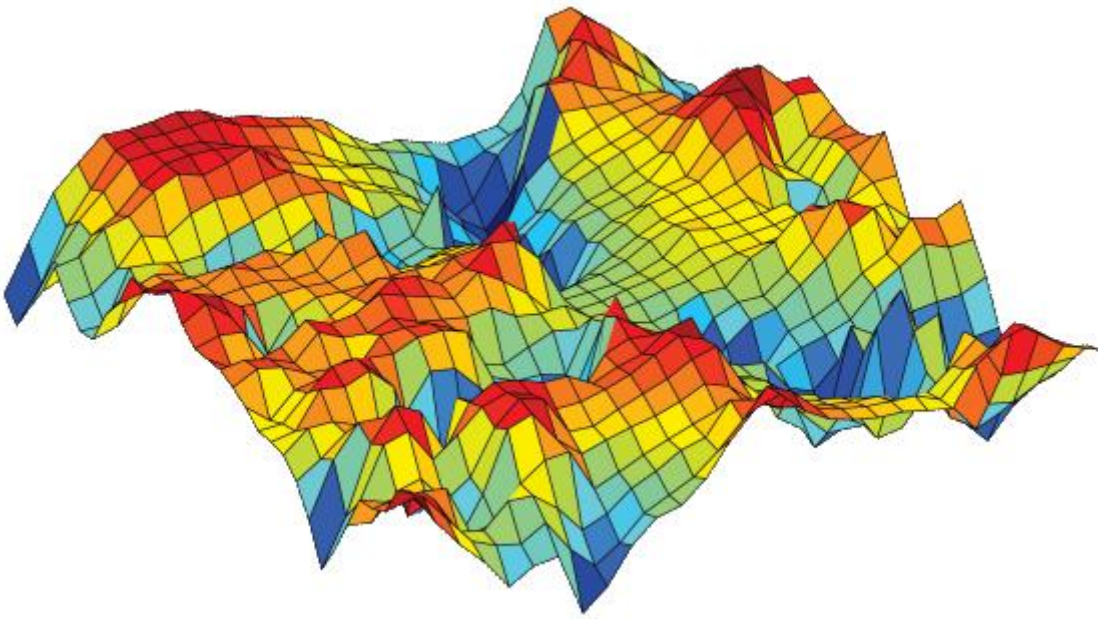


Obrázek 6: Vícerozměrné škálování
Zdroj: Vlastní

- **Samoorganizující se mapy**

Základním druhem samoorganizující se mapy (SOM) je tzv. Kohonenova mapa. Tato metoda je založena na umělých neuronových sítích a je vhodné ji využívat i pro explorační analýzu dat. Kohonenovy mapy jsou často aplikovány např. pro rozpoznávání řeči (při

přepisu řeči na text), zpracování obrazu, bezpečnostní aplikace atd. Příklad této vizualizace viz obrázek 7.



Obrázek 7: Samoorganizující se mapa
Zdroj: https://www.viscovery.net/self-organizing-maps

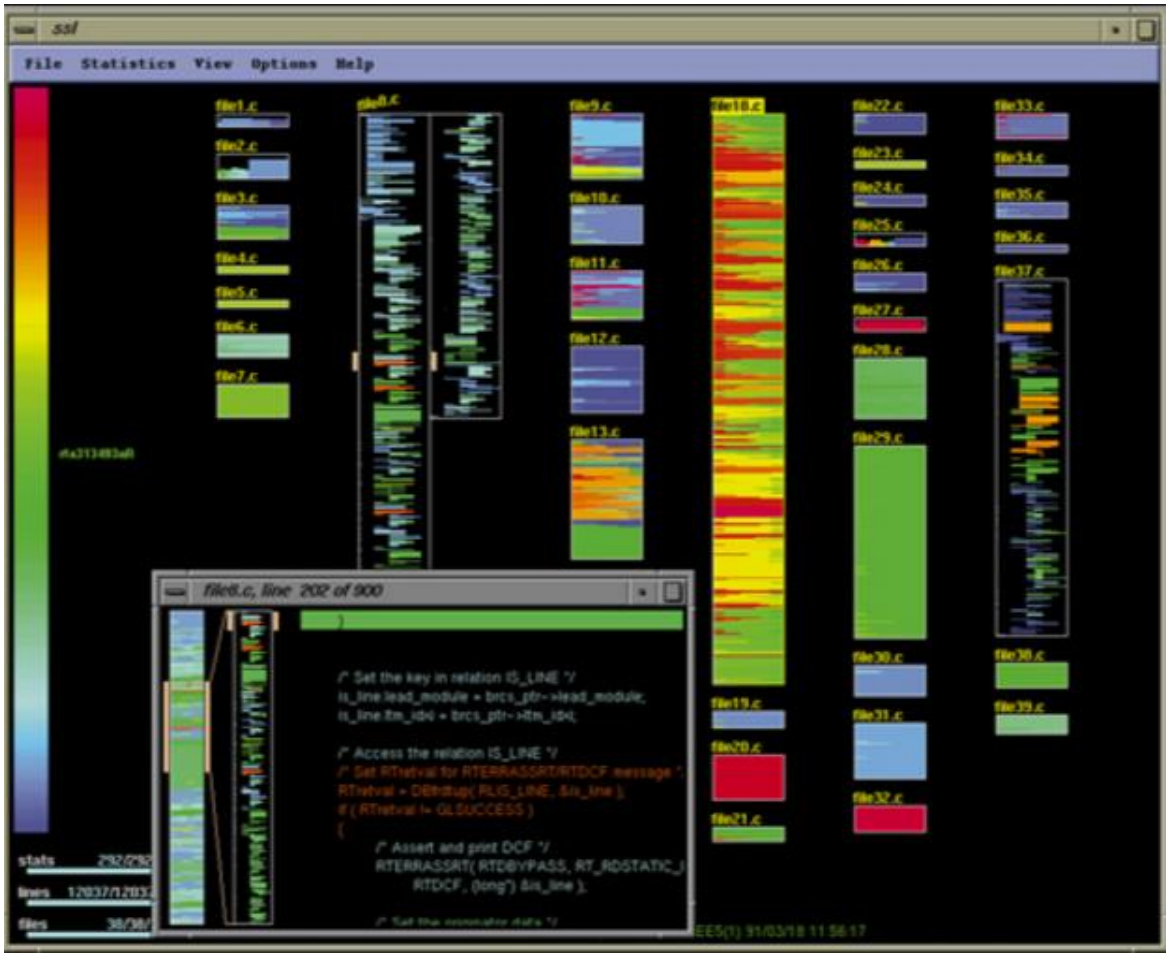
3.2.3 Pokročilá vizualizace

V následující kapitole je popsáno několik různých druhů pokročilé vizualizace.

3.2.3.1 Vizualizace softwaru

Tato metoda vizualizuje soubory tvořící zdrojový kód aplikace a znázorňuje vytíženost (četnost volání, užití) určitých jeho úseků, metod, funkcí viz obrázek 8.[19]

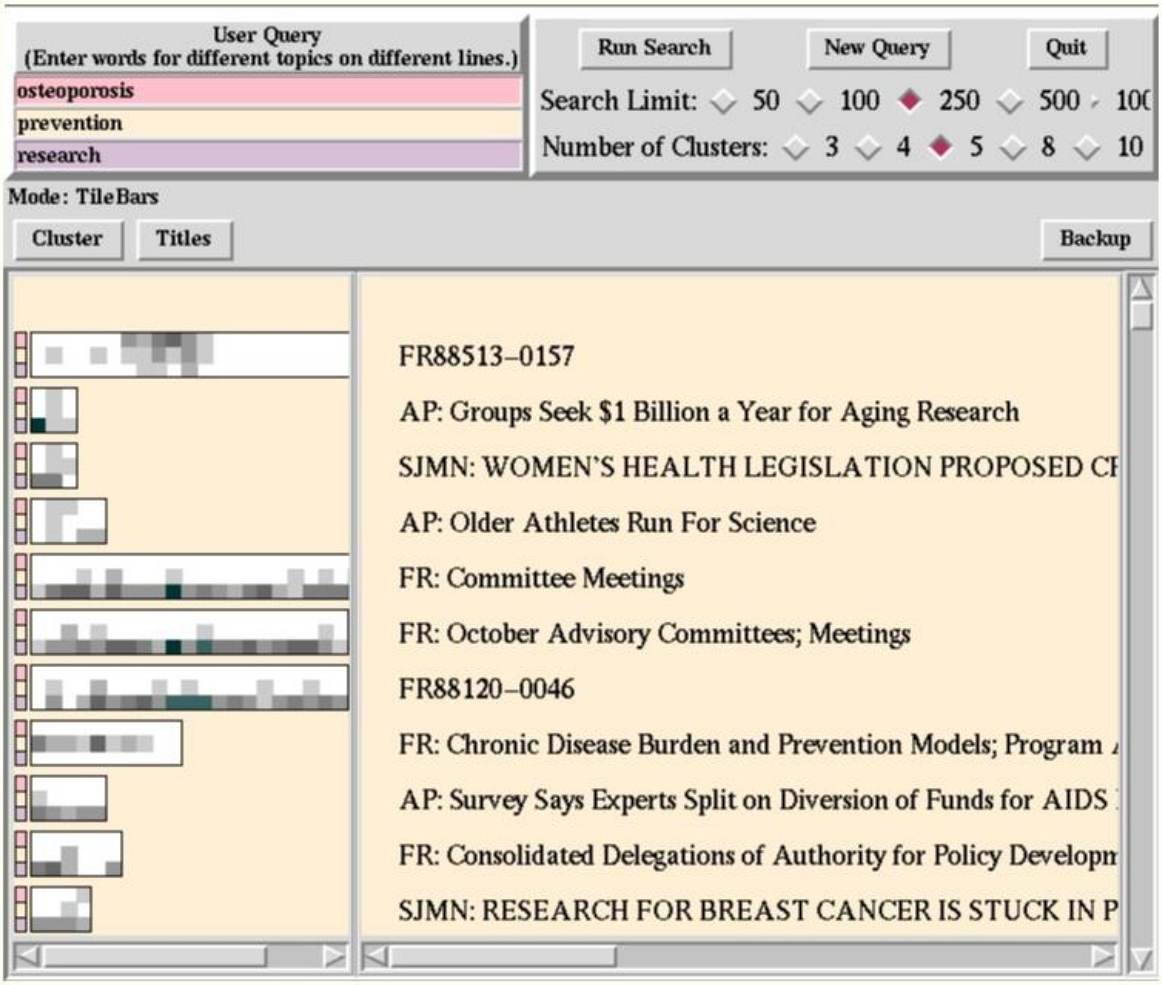
Na základě takto provedené analýzy je možné provést optimalizaci často volaných funkcí či metod.



Obrázek 8: Vizualizace softwaru - nástroj Seesoft
Zdroj:
http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Parija_Spacco/old_viewtips.html

3.2.3.2 Vizualizace výsledků hledání

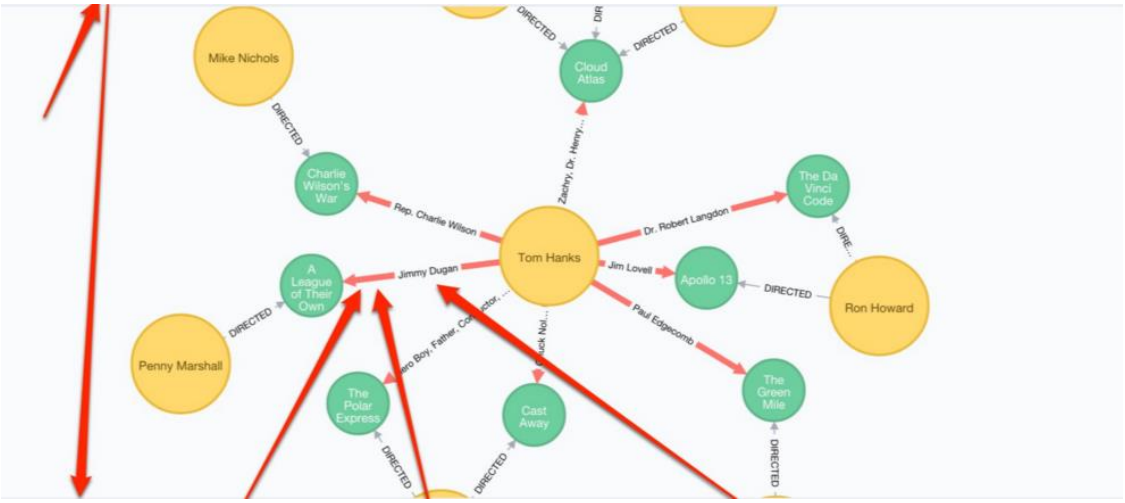
Tato metoda znázorňuje polohu a četnost vyhledávaného výrazu v rámci několika dokumentů. Délka prohledávaného dokumentu odpovídá délce obdélníku, který ho vyobrazuje (čím je delší obdélník, tím delší je prohledávaný soubor). Čtverečky v obdélníku zastupují jednotlivé kapitoly, odstín čtverečku znázorňuje četnost výrazu v dané kapitole (čím tmavší barvu čtvereček má, tím frekventovaněji se hledaná fráze či výraz v tomto místě dokumentu nachází) viz obrázek 9.



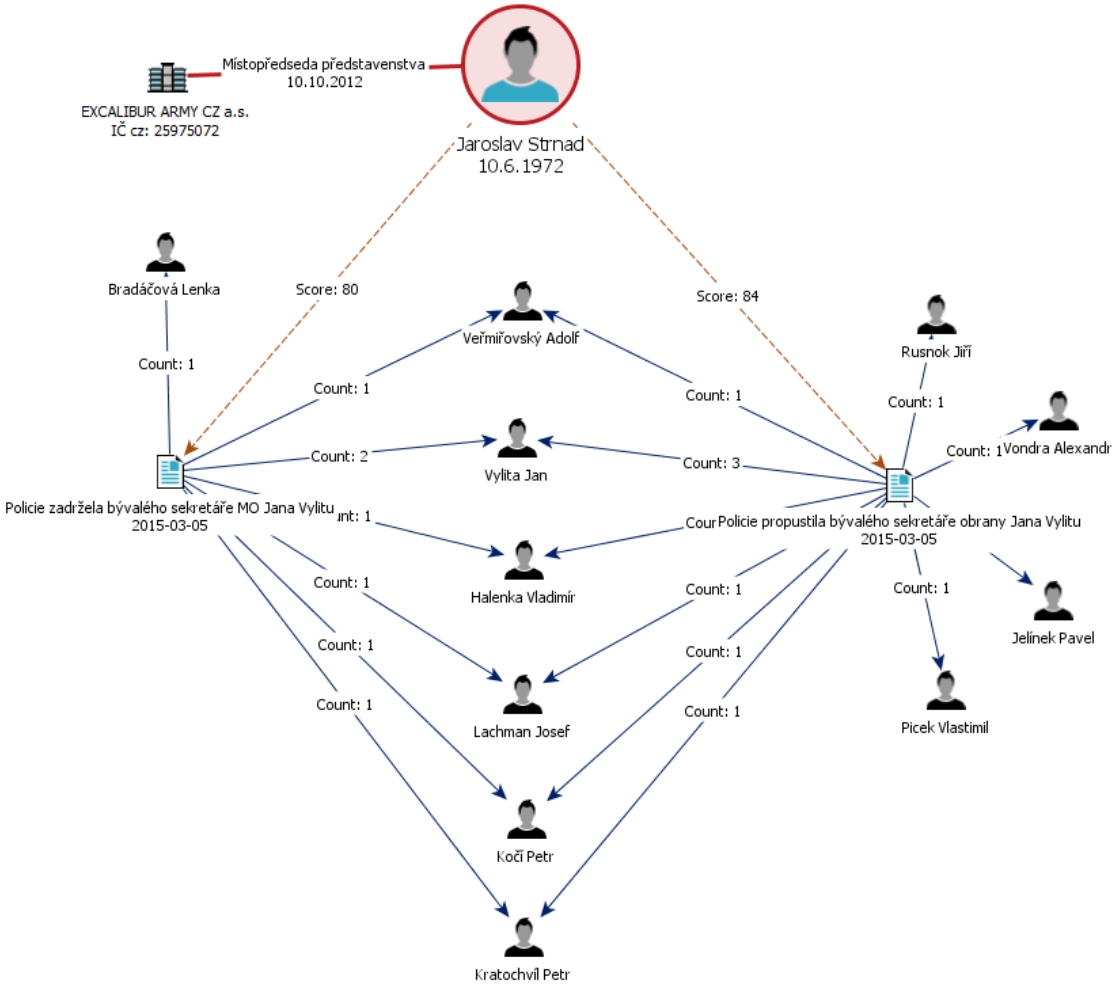
Obrázek 9: Vizualizace výsledků hledání
Zdroj: http://searchuserinterfaces.com/book/sui_ch10_visualization.html

3.2.3.3 Souvislosti a vztahy v nestrukturovaných datech

V současné době existuje hned několik řešení zaměřujících se také na zjištění souvislosti mezi různými entitami a následnou vizualizaci vztahů mezi nimi. Velmi časté je užití ER a jiných diagramů. Nalezené souvislosti jsou uspořádány do mapy, viz obrázek 10 a 11. Danou mapu je možné procházet a nacházet vztahy a vazby. [19]



Obrázek 10: Vizualizace vztahů
Zdroj: http://neo4j.com/developer/guide-neo4j-browser/

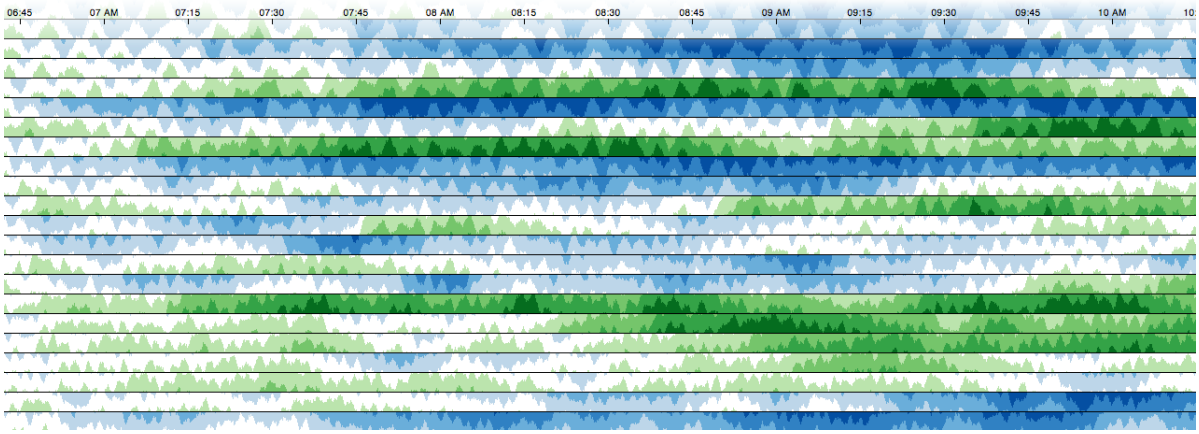


Obrázek 11: Vizualizace souvislostí v nestrukturovaných datech
Zdroj: http://www.cssi.cz/cssi/system/files/all/6-11-2015_TOVEK_0.pdf

3.2.3.4 Horizon graphs

Tato vizualizace zobrazuje vývoj různých hledisek (osa Y – např. ceny, počty produktů, hlasitosti atd.) v rámci času (osa X) z několika různých zdrojů. Kombinuje pozici a barvu pro lepší využití vertikálního prostoru viz obrázek 12. [18]

Tento graf se velmi často používá pro vizualizaci videa a audia.



Obrázek 12: Horizon Graph
Zdroj: https://www.perceptualedge.com/blog/wp-content/uploads/2009/01/horizon-graph-large.jpg

4. Technologické platformy pro zpracování dat

Zpracování nestrukturovaných a semistrukturovaných dat je velmi složitý proces. V současné době existuje hned několik technologických platforem, které jsou schopny tato data zpracovat velmi efektivně.

Business Intelligence (též BI) technologické platformy jsou schopny sběru, analýzy a vizualizace dat za účelem získání potřebných informací. Společnosti, které mají podrobný přehled o svém podnikání, jsou schopny lépe reagovat na požadavky zákazníků a také si efektivněji organizovat své podnikání.

Tato kapitola je věnována vybraným BI technologickým platformám.

4.1 Technologické platformy

Existuje nespočet různých technologických platforem zpracovávajících nestrukturovaná a semistrukturovaná data. V této bakalářské práci je popsáno pouze několik z nich. Všechna zvolená řešení splňují následující požadavky:

- zpracovávají Big Data (strukturovaná, nestrukturovaná a semistrukturovaná data);
- vstupní data platforem mohou pocházet z několika různých zdrojů;
- schopnost zpracovat historická data a data v reálném čase;
- využívá nástroje pro analýzu z rodiny „index&search“;
- nestrukturovaná a semistrukturovaná data transformují na strukturovaná.

Jednotlivé platformy budou popsány dle následujících klíčových hledisek: funkcionalita, zprostředkovatel, klienti a podporované operační systémy (dále jen OS).

4.1.1 Splunk

Obrázek 13: Logo platformy Splunk
Zdroj: http://logonoid.com/splunk-logo/

Funkcionalita: Splunk je technologická platforma, která zprostředkovává sběr a explorační analýzu dat (včetně strojových, tzv. logy), produkovaná novodobými technologiemi z různých zdrojů, např. z databází, mobilních zařízení, aplikací atd. Splunk pomáhá porozumět a reagovat na chování zákazníků, služeb, přístrojů a koncových uživatelů v reálném čase. Splunk Enterprise je perfektní platforma umožňující sběr, analýzu a vizualizaci všech strojových dat a jiných dat z fyzického, virtuálního či cloudového prostředí. [21]

Zprostředkovatel: Splunk působí na trhu od roku 2003. Největším a nejaktivnějším tuzemským partnerem Splunku je od roku 2013 společnost Trask solutions, a. s. Tato společnost je licenčním i technologickým partnerem, který zprostředkovává tuto platformu pro český a slovenský trh.

Klienti: Celkem má přes 11 000 zahraničních zákazníků. Mezi největší klienty společnosti Splunk patří Adobe, Bosch, ING, Symantec, Vodafone, GoodData a další. [21]

OS: Podporovány jsou všechny dostupné druhy operačních systémů, např. Windows, Linux, Solaris atd.

4.1.2 Sumo Logic

Obrázek 14: Logo platformy Sumologic
Zdroj: http://www.featuredcustomers.com/vendor/sumologic

Funkcionalita: Sumo Logic je technologická platforma provozovaná v cloudu. Toto řešení zpracovává a analyzuje Big Data v reálném čase a z různých zdrojů. Tato platforma svým klientům zprostředkovává interaktivní datové vizualizace a pomáhá porozumět datům. Sumo Logic využívá patentované technologie Log Processing, LogReduce a Push Analytics. [22]

Sumo Logic je považována za SaaS verzi platformy Splunk.

Zprostředkovatel: Platformu zprostředkovává společnost Sumo Logic, která byla založena v roce 2010 specialisty ze společnosti ArcSight

Klienti: Společnost má přes 1000 zákazníků. BBC, Orange, Medidata, Estee Lauder, WebJet, Concur , Ibotta a další.[22]

OS: Aplikace je hostována provozovatelem služby.

4.1.3 Colix

Obrázek 15: Logo platformy Colix
Zdroj: Vnitropodniková data společnosti Trask solutions, a. s.

Funkcionalita: Tato platforma dokáže zpracovávat Big Data v reálném čase. Umožňuje sbírat velké množství např. strojových dat hned z několika různých zdrojů. Prostřednictvím Colixu je možné získat komplexní pohled na podnikání, získat konkurenční výhodu a další výhody plynoucí z analyzování a vizualizace Big Dat. [23]

Jedná se o nový produkt společnosti Trask solutions, a.s., který se stále vyvíjí a doplňují se dodatečné funkcionality. Hlavní části systému jsou následující:

- **FluentD;**
- **Elasticsearch;**
- **Kibana.**

Tato platforma obsahuje i další části, které zajišťují například zabezpečení (šifrování, autentizace a další) či zvyšují funkcionalitu základních tří částí. Názvy dalších částí systému nebyly zveřejněny.

Zprostředkovatel: Colix je systém vytvořený společností Trask solutions, a. s., jako cenově dostupnější alternativa technologické platformy Splunk, kterou firma také zprostředkovává.

Klienti: Jedná se o nový systém, který se momentálně nabízí potencionálním zákazníkům. Novým klientem je ČSOB, kde se Colix používá pro monitoring části infrastruktury. [23]

OS: Tato platforma podporuje servery s operačním systémem Linux.

4.1.4 ELK stack

Funkcionalita: Elastic Stack je další řešení pro zpracovávání a explorační analýzu Big Dat. Toto řešení se skládá ze tří částí z Elasticsearch, LogStash a Kibana. Každá z těchto komponent vykonává určitou operaci:

- **LogStash;**
- **Elasticsearch;**
- **Kibana.**

Tato platforma, kterou společnost Elastic nabízí, může být doplněna o další produkty firmy:

- **Beats** – tato platforma rozšiřuje funkcionalitu Logstash, umožňuje sbírat a přeposílat více druhů dat.

- **Shield** – zajišťuje vysokou úroveň zabezpečení Elastic Stack pomocí šifrování, stanovení přístupových rolí a auditu. Splňuje současné rostoucí požadavky na zabezpečení obchodních dat.
- **Watcher** – tento produkt zasílá upozornění a zprávy do Elasticsearch. Upozorňuje na nestandardní anomálie v systémových lozích.
- **Marvel** – systém, jehož úkolem je zajistit nejvyšší výkon Elasticsearch, informuje o stavu jeho vytížení a tím se snaží předejít možným problémům. [24]

Zprostředkovatel: Toto řešení nabízí společnost Elastic, dříve Elasticsearch, která byla založena roku 2012 lidmi z open source projektů Elasticsearch a Apache Lucene.

Toto řešení nabízí jako SaaS službu společnost Logz.io. [25]

Klienti: Mezi klienty této platformy patří významné korporace jako například Netflix, LinkedIn, Fujitsu, Stack Overflow a další. [25]

OS: Podporuje servery s operačními systémy Linux, Ubuntu a další.

4.1.4.1 ELK

Funkcionalita: Pokročilý uživatelé si mohou nahradit systém Elastic Stack využitím pouze jeho základních součástí: **Logstash, Elasticsearch** a **Kibana**, popřípadě nahradit jinými open source systémy, se stejnou či velmi podobnou funkcionalitou. Velkou nevýhodou ELK platformy je nutnost porozumět jeho třem částem a tomu, jak pracují.

V případě znalosti těchto systémů je možné si „utvořit“ platformu, která systém Elastic Stack a jiné platformy nahradí a je bezplatná.

Zprostředkovatel: Hlavní tři open source řešení zprostředkovává společnost Elastic.

Klienti: Nelze zjistit, vzhledem k volné distribuci jeho základních komponent.

OS: Podporuje servery s operačními systémy Linux.

4.2 Open source nástroje využívané ve vybraných technologických platformách

Tato kapitola je zaměřena na popis nástrojů, které využívají výše zmíněné platformy.

4.2.1 Sběr dat

4.2.1.1 FluentD

Obrázek 16: Logo FluentD
Zdroj: http://docs.fluentd.org/articles/logo

Tento open source systém zajišťuje sběr dat.

FluentD se snaží co nejvíce strukturalizovat data jako JSON. To umožňuje sjednotit všechny aspekty zpracovávání logů: sběr, čištění, buffering (ukládání do vyrovnávací paměti) a výstupní logy v rámci několika zdrojů a cílů (Unified Logging Layer). [26]

Podporuje servery s operačním systémem Linux nebo Windows a je vlastnictvím společnosti Treasure Data. Pro FluentD je k dispozici 300+ rozšíření (pluginů) [27]

Směrování událostí (Event routing)

FluentD se spoléhá na značky (tagy) sloužící ke směrování událostí. Každá událost ve FluentD má znak (tag), který určuje, kam má být přesměrována – jedná se o deklarativní přístup.[27]

Výkonnost

Výkonnost je u těchto systémů složitě měřitelná, důležité jsou okolnosti měření. Dle Noni Peri FluentD zpracuje přes 10 000 událostí za sekundu a vyžaduje pouze 40 MB paměti. [27]

FluentD je naprogramován pomocí programovacího jazyka C, ale především v Ruby. [26]

4.2.1.2 LogStash

Obrázek 17: Logo Logstash
Zdroj: http://logz.io/blog/logstash-tutorial/

Jedná se o flexibilní, open source systém zprostředkovávající sběr dat.

Logstash pracuje na serverech s operačními systémy Windows nebo Linux a je jedním z produktů společnosti Elastic. Pro rozšíření funkcionalit systému je k dispozici mnoho dodatků (pluginů) (200+). [27]

Směrování událostí (Event routing)

Logstash směřuje všechna data do jednoho „proudu" dat, poté využívá algoritmů if-then. V případě, že je splněna podmínka tak pošle data do správného místa určení. Tato metoda je procedurální. [27]

Výkonnost

Logstash zpracuje přes 10 000 událostí za sekundu, jak uvádí ve svém článku Noni Peri. Toto zpracování vyžaduje více paměti, kolem 120 MB. [27]

Logstash je naprogramován v jazyce Java a Ruby. [24]

4.2.2 Zpracování dat

4.2.2.1 Elasticsearch

Obrázek 18: Logo Elasticsearch
Zdroj: http://www.sitepoint.com/introduction-to-elasticsearch-in-php/

Jedná se o open source nástroj, který slouží jako vyhledávací a analytický engine, navržený pro horizontální škálovatelnost, spolehlivost a jednoduchou správu. Kombinuje rychlost vyhledávání s výkonem analytiky prostřednictvím sofistikovaného, uživatelsky přívětivého dotazovacího jazyka, pokrývající strukturovaná, nestrukturovaná a data časových řad. [24]

Vzhledem k vysoké rychlosti systému Elasticsearch je možné analyzovat data prakticky v reálném čase.

Jedná se o distribuovaný systém, který je možné škálovat dle jeho zvyšující se zátěže serveru, na kterém pracuje. Při nedostatečném výkonu je nutné přidat další server, zátěž se poté rozloží mezi spolupracující servery. [24]

Vyhledávání

Elasticsearch využívá nejvýkonnější fulltextové vyhledávání dostupné v open source Apache Lucene. Podporuje několik jazyků, umožňuje vyhledávat příbuzné i podobné záznamy.[24]

Indexování

Proces indexování je jednou ze základních funkcí, které Apache Lucene zprostředkovává. Umožňuje data efektivně vyhledávat.

4.2.3 Vizualizace dat

4.2.3.1 Kibana

Obrázek 19: Logo Kibana
Zdroj: http://blog.arungupta.me/tag/kibana/

Kibana je open source platforma umožňující vizualizaci dat pomocí ohromující a vkusné grafiky, od histogramů přes geomapy. Kibana přivádí data k životu pomocí zajímavých vizualizací, které mohou být sloučeny do vlastních dashboardů, které pomohou sdílet poznatky z vybraných dat. [24]

Kibana pomáhá porozumět prostřednictvím vizualizace velkému množství dat.

5. Porovnání vybraných technologických platforem

Cílem bakalářské práce je porovnání zvolených technologických platforem z hlediska jejich funkcionalit, ceny, výkonu, podpory a dalších kritérií, která jsou klíčová při výběru vhodného řešení pro zpracování Big Dat.

Na zadání společnosti Trask solutions, a. s., v rámci které je vykonávána řízená praxe jsou v této bakalářské práci porovnány technologické platformy Splunk a Colix. Obě tyto platformy společnost zprostředkovává a umožnila i jejich testování.

Kritéria pro porovnání technologických platforem

Kritéria byla zvolena na základě nejdůležitějších otázek, které by si měli potencionální zákazníci klást při výběru vhodné BI technologické platformy pro zpracování nestrukturovaných a semistrukturovaných dat.

Mezi nejdůležitější kritéria pro volbu vhodné BI platformy patří výkon, funkcionalita, cena, uživatelská přívětivost a možný rozvoj aplikace.

5.1 Výkon

Jedno z nejdůležitějších kritérií je výkon dané technologické platformy. V tomto případě je důležité, jaké množství dat (MB) je daná platforma schopná zpracovat za jednu sekundu a kolik paměti přitom spotřebuje.

Nejdůležitější dopad na výkon má architektura a způsob zpracování dat danými platformami.

5.1.1 Architektura a způsob zpracování dat platformy Splunk

Splunk je založen na enginu MongoDB.

5.1.1.1 Popis řešení získávání a zpracování dat

Forwarder (*Splunk> heavy forwarder*)

- realizuje „data routing“, tzn. že přeposílá události splňující definovaná kritéria na jinou instanci platformy Splunk.

Master node (*Splunk> master node*)

- řídí replikaci dat „bucketů“ (metaúložišť s nastavenou životností dat a vlastnostmi dat) mezi jednotlivými indexery a také spravuje licence.

Indexer (*Splunk> indexer)*

- realizuje indexaci událostí. Indexer provádí normalizaci, kategorizaci a získává data z polí. Takto normalizovaná a kategorizovaná data se později využijí pro reportování, vytváření alertů atd.

- **Normalizace**

Cílem normalizace je převod dat na normalizovanou formu. Proces normalizace využívá vyhledávacích tabulek k nahrazení názvů a hodnot polí standardizovanými názvy a hodnotami. Příkladem nahrazení hodnoty pole je např. severity=sm“ za „severity=small nebo názvu pole např. „sev=medium“ na „severity=medium“. Jednotné formátování umožňuje bezproblémové zobrazení polí v reportech a dashboardech.

Cílem normalizace je využít stejné názvy a hodnoty mezi událostmi z různých zdrojů – převedení do jednotných (normalizovaných) názvů polí a hodnot.

Splunk pro sjednocení dat používá model CIM (Common Information Model). [23]

- **Kategorizace**

Různé aplikace a dodatky Splunku využívají tzv. Splunkový typ události a způsob značkování (tagging) typu událostí pro kategorizaci různých typů dat. Rozlišuje data z různých zdrojů např. ze serveru Exchange atd.. Pole source typ – tzv. pomocné pole. [23]

- **Parsování a získávání dat z polí**

K parsování dochází při transformačním procesu a indexaci dat. Zdrojové informace získané z událostí jsou ukládány do jednotlivých polí. Tato pole jsou normalizována a standardizována i pro další rozšíření Splunku (Exchange, Mobile Access, Splunk for Twitter atd.). [23]

Search Head (*Splunk> search head)*

- zprostředkovává „search management“. V tomto případě „search management“ je v clusterovaném prostředí schopnost rozprostřít jeden dotaz mezi několik Splunk indexerů. [23]

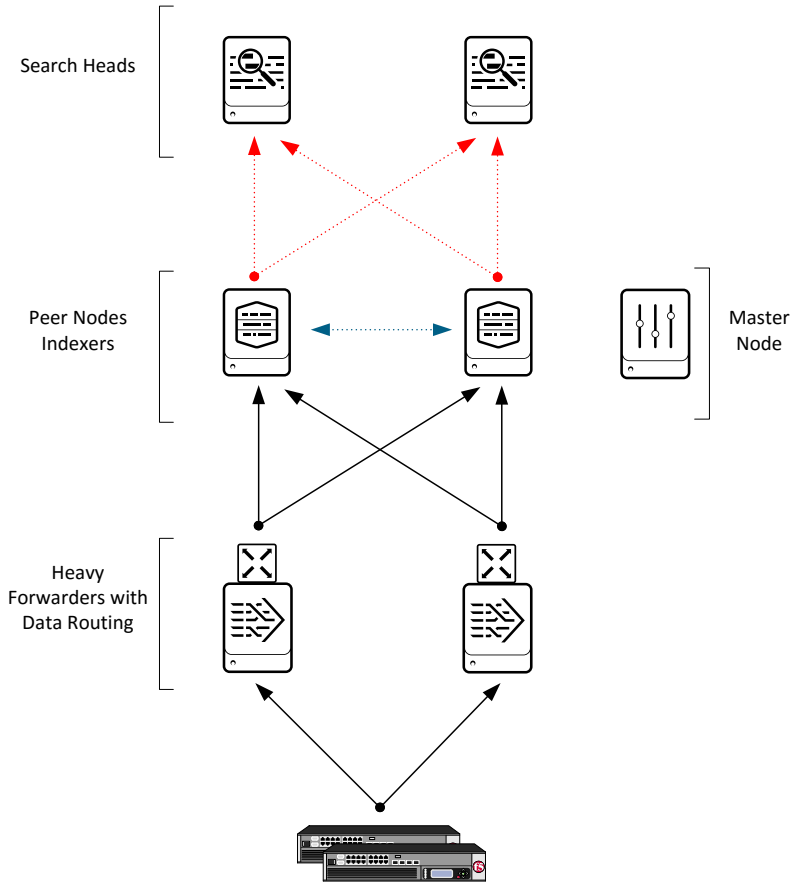
- **Princip získávání dat platformou Splunk**

Raw data (kolekce dat, která dosud nebyla analyzována či formátována) jsou odesílána na server, který slouží ke komunikaci mezi Splunk>heavy forwardery a zdroji RAW dat. Jakmile se Raw data dostanou na Splunk>heavy forwardery, dojde k jejich přesměrování (tzn. Data Routing) na Splunk> indexery. Ke směrování dat dochází na základě kritérií definovaných uživatelem. Splunk>indexery příchozí data zaindexují a provedou kategorizaci a normalizaci (*pro pozdější využití*) a přepošlou takto zpracovaná (již strukturovaná data) do konzole událostí. [23]

Vysoká dostupnost dat na úrovni Splunk>indexerů je zajištěna cluster technologií Splunk>Enterprise, která zajišťuje replikaci všech dat mezi Splunk>indexery, které jsou součástí clusteru. [23]

- **Architektura technologické platformy Splunk**

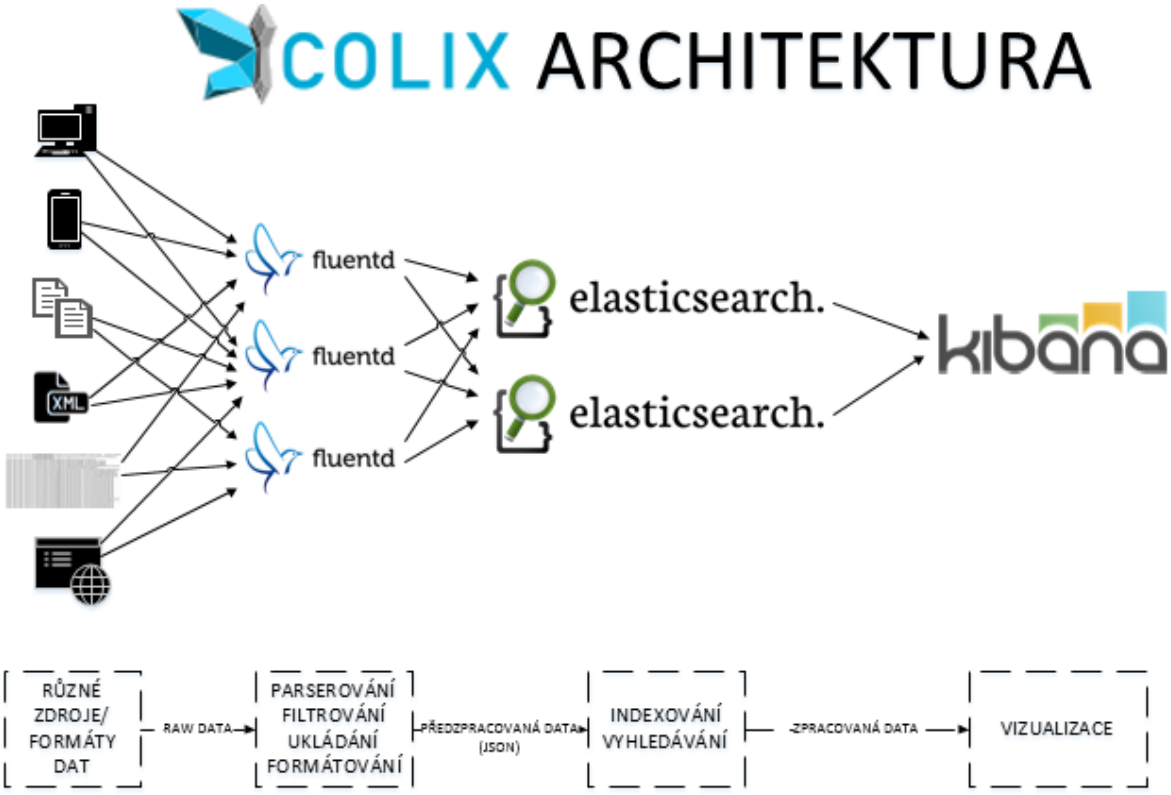
Architekturu jednotlivých komponent znázorňuje obrázek 20.



Obrázek 20: Architektura technologické platformy Splunk
Zdroj: Vnitropodniková data společnosti Trask solutions a. s.

5.1.2 Architektura a způsob zpracování dat platformy Colix

Platforma Colix se skládá z několika open source řešení, která jsou vzájemně propojená, viz obrázek 21.



Obrázek 21: Architektura technologické platformy Colix
Zdroj: Vlastní

Jak popisuje architektura výše, Colix se skládá z řešení FluentD, Elasticsearch a Kibana. Z důvodu zamezení výpadků při zpracování dat běží procesy na více instancích. Tato trojice open source nástrojů je doplněná dalšími technologiemi, aby byly zajištěny i jiné nežli základní funkce (sběr dat, zpracování dat, analýza, vizualizace). Bližší informace dalších řešení, které Colix obsahuje, firma neposkytla.

5.1.2.1 Popis získávání a zpracování dat

- **Sběr dat**

Jak již bylo zmíněno v předešlých kapitolách, FluentD sbírá data (např. datové logy). Jednotlivé komponenty tohoto nástroje jsou uvedeny na obrázku 22.

Obrázek 22: Komponenty FluentD
Zdroj: Vlastní

Vstup – jedná se o komponentu, která zajišťuje příjem logů z různých zařízení. Oficiálně podporované formáty je JSON přes HTTP, syslog a další. V případě, že je potřeba poslat data v jiném, než oficiálně podporovaném formátu, je nutné napsat plugin v programovacím jazyce Ruby.

Parser – tato komponenta zajišťuje transformaci vstupu na definovaný formát.

Filtr – jedná se o část nástroje, která obstarává filtrování a případnou modifikaci proudů událostí.

Vyrovnávací paměť (Buffer) – zajišťuje spolehlivost FluentD. V případě, že dojde k neočekávané chybě na výstupu, jsou zde události uloženy (v bufferu) a po obnově je pokus o výstup automaticky opakován. Buffer může být paměťový nebo ve formě souborového uložiště.

Výstup – jedná se o komponentu, která přijímá bloky logů, které odesílá vyrovnávací paměť a přeposílá je dál. (v případě Colixu je posílá na server, kde se nachází Elasticsearch).

Formátor – je použit v případě, že výstup neodpovídá definovaným požadavkům. Tato komponenta obsahuje kodeky, které umožňují naformátovat výstup do vhodné formy. Obsahuje například kodek, který umožňuje transformovat data do formátu JSON, který je vhodný pro Elasticsearch.

- **Indexace a vyhledávání**

Elasticsearch je nástroj založený na knihovně Apache Lucene, využívá přístup k analýze index&search.

Elasticsearch vykonává následující dva základní procesy:

- **Indexace**

Cílem indexace je reprezentace dat, v takové struktuře, ve které je možné efektivně data vyhledávat.

Indexace je proces ukládání dat do indexů. Index je základní datový kontejner, který obsahuje data, která spolu souvisejí.

Ukládání dat v Elasticsearch

V indexech jsou uloženy dokumenty ve formátu JSON (datový typ určený pro přenos a zápis dat, data jsou uložena jako dvojice klíč-hodnota).

Do indexu je možné uložit více různých typů dokumentů. V průběhu vkládání záznamu obsahujícího jednotlivá pole, dochází k jeho zpracování pomocí analýzy.

Analýza

Analýza je prováděna při vyhledávání i indexaci dat.

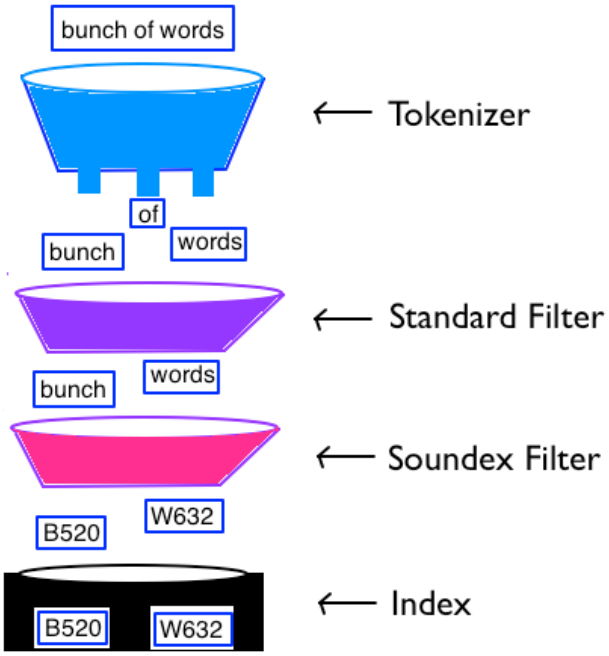
Proces analýzy se skládá z procesu:

- tokenizace
- normalizace

Procesy analýzy dat vykonávané tzv. analyzery:

- **Tokenizace –** rozdělí text na jednotlivá slova. Výstupem tokenizace je stream tokenů.
- **Normalizace –** převede termíny do normalizované formy pomocí procesů:
  - **Stematizace** – převede slova v různých tvarech na kmen slova (odstraněním předpon a přípon)
  - **Lemmatizace** – převede slova do základního tvaru.
- **Filtruje –** zpracovává streamy tokenů. Odstraňuje tzv. stop words (např. spojky). Stop words jsou v každém jazyce jiné, v českém jazyce například a, i atd.

Příklad tohoto procesu znázorňuje obrázek 23.



Obrázek 23: Colix – proces indexování
Zdroj: http://blog.jessitron.com/2012/04/configuring-soundex-in-elasticsearch.html

- **Vyhledávání**

Elasticsearch je nástroj sloužící k fulltextovému vyhledávání. Aby bylo možné dotaz zpracovat, je nutné transformovat data pomocí analýzy, která je popsána v předchozí kapitole.

Colix je kompatibilní se všemi Elasticsearch indexy, tzn. je možné v nich vyhledávat. Například lze pomocí Colixu vyhledávat v indexech připravených v platformě ELK stack a to včetně použití dotazovacích skriptů.

- **Princip získávání dat technologickou platformou Colix**

Raw data (data dosud nezpracovaná, neformátovaná) z různých zařízení jsou odesílána na server, kde se nachází FluentD, který data parseruje, filtruje, ukládá do vyrovnávací paměti (bufferu), formátuje a takto modifikovaná data posílá na server s Elasticsearch. Elasticsearch data indexuje (ukládá), normalizuje a umožňuje jejich jednoduché vyhledávání.

5.1.3 Měření výkonu

Toto kritérium je velmi složitě měřitelné, důležité jsou okolnosti měření např. typ vstupních dat, operační systém, verze jednotlivých technologických platforem či jejich komponent atd.

Měření výkonu bylo provedeno na platformě MS Windows 2012 R2, CPU 4core Intel Xeon E5-2630 ver2, 16GB RAM. Verze platforem byly následující: Splunk 6.4 a Colix 2.3 (Elasticsearch 2.1, FluentD 4.x, Kibana 4.3, JDK 1.8.45)

- **Popis měření**

Měření výkonu probíhalo prostřednictvím aplikace, která posílala data ve formátu JSON k indexaci.

Vzor dat:


```
{    "timestamp":"2015-07-
08T17:14:46.359","logBatch":{"logMessage":{"header":{"OPERATION_ID":"ID14362
820862960","logHost":"ntrosa2","TIFCOMPONENT":"TAS"},"mq": {
"messageID":"0F000000000000000000000000000000000000000000000000",
"correlationID":"0F000000000000000000000000000000000000000000000000",
"targetQueue":"TRO.REQ",  "targetQueueManager":"QM1",
"replyToQueue":"TRO.REP",  "replyToQueueManager":"QM1"  },"applicationLog":
{    "logMessage":"XMLParserPipe - Missing element 'RqUuid' in XML message",
"logTimestamp":"2015-07-07T17:14:46.359",    "logLevel":"WARN",
"logThread":1  },"performanceLog":{"agent":{    "call":1436282086359}}}}}
```


Vstupními daty pro platformu Splunk je RAW TCP Stream, jednotlivé záznamy jsou odděleny novým řádkem. [23]

Platforma Colix využívá pro filtrování a formátování logů systém FluentD. Ten byl nakonfigurován tak, aby otevřel nový TCP port a přijímal vstupní data ve stejném formátu jako Splunk. Tato data poté předával zformátované jako JSON nástroji Elasticsearch. Měření bylo provedeno třikrát, výsledné hodnoty jsou průměrem z těchto měření (tabulka 1). [23]

Poznámka: TCP stream dat byl odesílán z .NET aplikace jedním threadem, REST JSON do Elasticsearch byl odesílán z Java aplikace s 50 thready. [23]

- **Výsledek měření**

Tabulka 1: Výsledek měření výkonu porovnávaných technologických platforem

Aplikace	Čas [ms]	Počet záznamů	Num/sec (TIF)
Splunk	81 020	1 000 000	12 342,63
Colix	122 576	1 000 000	8 158,20

Zdroj: Vnitropodniková data společnosti Trask solutions a. s.

Komentář: **TIF** = typ události (podobný typ jako je JSON)

Závěr

Dle měření výkonu je za daných podmínek technologická platforma Splunk výkonnější než systém Colix.

Tento fakt je způsoben především komplexností/ucelenostís platformy Splunk. Colix je vzhledem ke své architektuře sestávající se z několika open source řešení méně výkonnou platformou – důvodem je složité propojení jednotlivých řešení.

5.2 Náklady na pořízení, rozvoj a provoz, návratnost investic

Cena technologické platformy je z pohledu zákazníků jedním z klíčových kritérií.

- **Splunk**

Náklady na pořízení: Splunk Enterprise je multioborová platforma sloužící i jako výkonný BI nástroj. Jeho cena je i vzhledem k jeho rozsáhlé funkcionalitě vyšší. Pro velké množství zákazníků je tato platforma dražší, než by považovali za vhodné. Cena licence se odvíjí od maximálního objemu nekomprimovaných dat zpracovaných v rámci jednoho dne.

Rozvoj: Pro řešení Splunk je vytvořené nepřeberné množství aplikací či dodatků, které umožňují rozšířit a uzpůsobit si funkcionalitu tohoto systému. Tato rozšíření si mohou jeho uživatelé stáhnout, některá jsou bezplatná a za některá je nutné si připlatit, např. Splunk Enterprise Security atd. Cena těchto rozšíření není vzhledem k jejich funkcionalitě vysoká.

Splunk nabízí i verzi Free, která umožňuje tuto platformu vyzkoušet. Je ale omezena na zpracování maximálně 500 MB dat denně.

Splunk vydává pravidelné updaty jak svých aplikací, tak i celého systému. K zakoupené licenci jsou tyto updaty k dispozici.

Provoz: Platí se zde za licenci i za podporu platformy.

- **Colix**

Platforma Colix je nabízena jako cenově dostupnější varianta Splunku. Především pro zákazníky, kteří nejsou ochotni investovat do vysokých pořizovacích nákladů.

Náklady na pořízení: Prvotní investice do platformy Colix je minimální. Colix je poskytován prostřednictvím poplatku za službu. Nabízeno je několik variant tohoto řešení, v rámci kterých lze platformu upravit dle požadavků zákazníka bez dodatečných nákladů.

Rozvoj: V případě zájmu o dodatečné rozšíření řešení je nutné zaslat poptávku společnosti Trask solutions, a. s., která platformu Colix uzpůsobí tak, aby splňovala kladené požadavky. Vzhledem k časové náročnosti a složitosti rozšíření funkcionalit tohoto systému patří tyto zakázky k dražším investicím.

Provoz: Platí se za službu (provoz řešení Colix) a také za updaty, inovace.

V Colixu jsou prvotní náklady minimální, dál se platí za službu. Z tohoto důvodu je toto výhodné řešení pro společnosti.

Návratnost investic

Prvotní investice do technologické platformy Splunk je mnohonásobně větší než investice do řešení Colix. Splunk poskytuje řešení pro všechna firemní odvětví, je komplexní. Lze předpokládat, že zakoupením licence Splunk ušetří podnikům velké výdaje v budoucnosti.

Zatímco prvotní investice do řešení Colix je oproti platformě Splunk velmi nízká, je zde velká pravděpodobnost dalších výdajů v případě, že bude potřeba tuto platformu rozšiřovat. Pro každé rozšíření je nutné uložit společnosti Trask solutions, a. s., novou zakázku.

Vzhledem k těmto faktům lze předpokládat, že ceny jsou v ohledu dlouhodobějšího užívání platforem, které reagují na vývoj podnikání, srovnatelné.

5.3 Funkcionalita

Mezi další důležitá kritéria patří funkcionalita technologické platformy. Mezi funkcionality je zařazena vizualizace, schopnost integrovat všechna potřebná data, zabezpečení platforem a SIEM zabezpečení.

Ačkoliv porovnávaná řešení nabízí velmi podobné funkcionality, tak jsou zde odlišnosti v rozsahu poskytovaných funkcí.

5.3.1 Vizualizace

Vizualizace má důležitou roli při výběru vhodné technologické platformy.

- **Splunk**

Technologická platforma Splunk nabízí velmi interaktivní formy vizualizací. Zprostředkovává běžné typy vizualizací – plošné grafy, spojnicové, výsečové, metriky, sloupcové grafy a geografické mapy. Toto řešení nabízí i tvorbu pokročilých vizualizací, jako např. Bubble chart.

- **Colix**

Colix pro vizualizaci výsledků analýz využívá open source nástroj Kibana. Kibana umožňuje tvořit pouze běžné typy vizualizací – grafy plošné, spojnicové, výsečové, sloupcové, různé metriky a geografické mapy.

5.3.2 Schopnost integrovat všechna potřebná data

Platformy Splunk i Colix umožňují zpracování všech běžně dostupných druhů dat.

5.3.3 Zabezpečení dat a celého systému

Řešení Splunk i Colix jsou plně v souladu se Zákonem o kybernetické bezpečnosti č.181/2014 Sb. Nad rámec zákona je řešení vhodné i pro informační systémy, které nakládají s utajovanými informacemi (§ 1, odst. 2). Předpisy z oblasti bezpečnosti informací jsou obsaženy v normách z rodiny ISO 2700x.

- **Splunk**

V tomto kritériu hodnocení má platforma Splunk velkou výhodu ve své celistvosti (jedno kompletní řešení, integrované GUI). Vzhledem k tomuto faktu obsahuje velké množství jak běžných (autentizace, řízení přístupu na základě rolí, šifrování, audit atd.), tak i nadstandardních zabezpečovacích prvků.

Jedním z nadstandardních prvků je možnost monitorovat chování uživatelů/systémů, které Splunk monitoruje. Tato platforma k nestandardním událostem v systému přiřazuje i tzv.

závažnost události. Na základě váhy, kterou Splunk stanoví, lze předcházet nežádoucím situacím.

- **Colix**

Platforma Colix je celistvá, ale nachází se v rannější fázi vývoje. Stále se vyvíjí nové a lepší funkce pro toto řešení. K zabezpečení dat využívá řešení typu Shield od Elasticu. Colix má potenciál dosáhnout v oblasti zabezpečení úrovně platformy Splunk.

5.3.4 SIEM – zabezpečení proudů dat

SIEM = složeno z výrazů SIM (Security Information Management) a SEM (Security Event Management). Jedná se o management bezpečnostních informací a událostí. Technologie SIEM sbírá logy z různých zařízení a v případě nestandardních událostí v logu upozorní na možný bezpečnostní incident. [28]

Dle zákona o kybernetické bezpečnosti má veřejný sektor povinnost tuto technologii využívat.

- **Splunk**

Splunk Enterprise s aplikací Enterprise Security je dle společnosti Gartner jedním z leaderů v oblasti SIEM – zabezpečení proudů dat. Vzhledem k tomuto faktu lze toto řešení považovat za jedno z nejlepších v oblasti zabezpečení proudů dat.

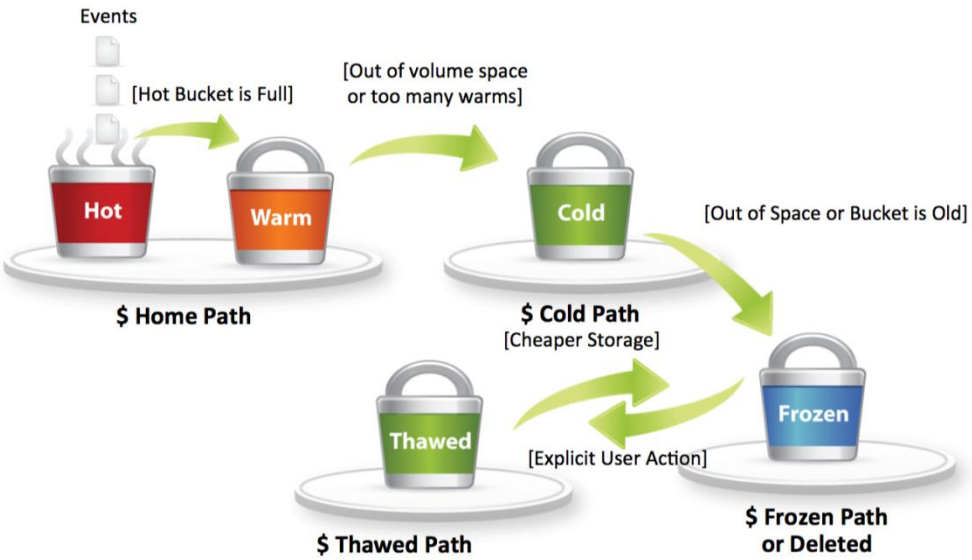
- **Colix**

Platforma Colix technologii SIEM neposkytuje.

5.3.5 Retence dat

Životnost dat (tzn. jak dlouho a kde jsou určitá data uchovávána) je řízena retenční politikou na úrovni tzv. bucketů (metaúložišť s nastavenou životností dat a vlastnostmi dat).

- **Splunk**



Obrázek 24: Splunk – Retence dat
Zdroj: Vnitropodniková data společnosti Trask solutions, a. s.

Retenci dat je možné přizpůsobit požadavkům uživatelů, lze nastavit dobu i objem dat. Retenci dat u Splunk znázorňuje obrázek 25.

Součástí clusteru indexerů je také tzv. Master node. Ten centrálně řídí celý cluster a spravuje licence.

- **Colix**

Retenci dat u řešení Colix je možné nastavit tak, aby splňovala nároky zákazníků. V této platformě je retence dat řešena na úrovni file systému pomocí skriptů.

5.4 Uživatelská přívětivost

Je velmi důležité, aby uživatelé byli schopni co nejdříve danou platformu akceptovat. Je žádoucí, aby daná řešení byla intuitivně ovladatelná, nebylo potřeba zdlouhavých školení, byla kvalitní dokumentační základna pro práci s daným nástrojem a možnost vkládat nové požadavky s co nejmenší závislostí na dodavateli nebo vlastním IT oddělením a další. [29]

5.4.1 Zákaznická podpora, dokumentace, updaty

Dalším podstatným kritériem je možnost zákaznické podpory a také dostupnost updatů či potřebné dokumentace.

- **Splunk**

Splunk má zákaznickou podporu na nejvyšší úrovni. Na webových stránkách http://www.splunk.com/en_us/support-and-services.html nabízí hned několik variant podpory. Dokumentace, návody a výukové programy jsou volně k dispozici na internetu.

Splunk svým zákazníkům poskytuje zdarma updaty na novější verze.

- **Colix**

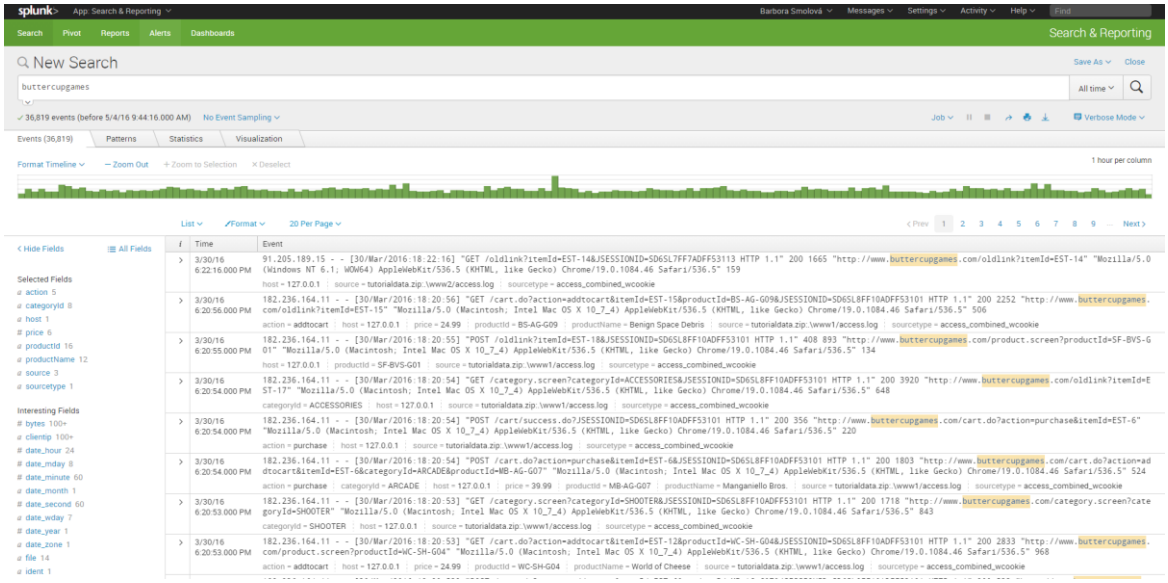
Platformě Colix zajišťuje podporu společnost Trask solutions a. s., ale vzhledem k faktu, že je tento systém poměrně nový, materiály a školení jsou zatím k dispozici pouze na vyžádání od společnosti Trask solutions, a. s..

5.4.2 Uživatelské prostředí

Ohodnocení tohoto kritéria je velmi subjektivní.

- **Splunk**

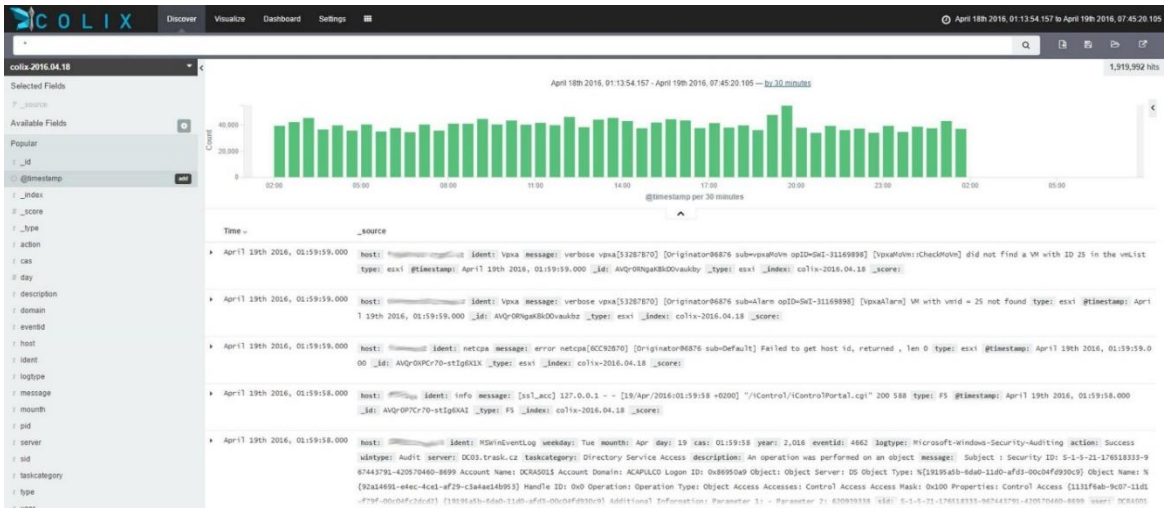
Splunk nabízí velmi přívětivé uživatelské prostředí, viz obrázek 26.



Obrázek 25: Splunk – uživatelské prostředí
Zdroj: Vlastní

- **Colix**

Uživatelé, kteří jsou zvyklí pracovat s open source Kibana, určitě upřednostní řešení Colix, které ji také využívá. Úvodní stránka platformy Colix viz obrázek 27.



Obrázek 26: Colix – uživatelské prostředí
Zdroj: Vlastní

5.4.3 Rozvoj aplikace

Dalším z klíčových kritérií je možný rozvoj daného řešení.

- **Splunk**

Jak již bylo zmíněno v kapitole o ceně, je možné Splunk Enterprise rozšířit o velké množství aplikací (např. Splunk Enterprise Security, Influx, DB Connect a dalších) a dodatků (např. PhishMe Intelligence, Cisco network atd.).

- **Colix**

V případě řešení Colix je nutné vznést specifický požadavek na rozšíření, který je možné na zakázku vytvořit. Vzhledem k tomuto faktu je doplnění systému o nové funkce složitější a časově náročnější, ve velkém množství případů i dražší.

5.5 Volba vhodné technologické platformy

Bodové ohodnocení jednotlivých platforem se výrazně liší v závislosti na sektoru podnikání. Pro účely této práce byl zvolen bankovní sektor, který vykazuje v současné době vysoký zájem o BI technologické platformy zpracovávající Big data. Bodové ohodnocení pro bankovní sektor bylo stanoveno na základě dlouhodobých zkušeností společnosti Trask solutions, a. s. (tabulka 2).

Tabulka 2: Kritéria a jejich bodové ohodnocení

Kritérium	Maximální bodové ohodnocení
Výkonnost	4
Náklady na pořízení, rozvoj, provoz a návratnost investic	48
Funkcionalita	39
Rozšiřitelnost	5
Uživatelská přívětivost	4
CELKEM	**100**

Zdroj: Vlastní

Poznámka: Maximální počet bodů je 100.

Dle porovnání, které byly provedeny v předchozích kapitolách práce, platformy Splunk a Colix obdržely následující bodové ohodnocení, viz tabulka 3.

Tabulka 3: Ohodnocení technologických platforem Splunk a Colix

Kritérium	Splunk	Colix
Výkonnost	4	3
Náklady na pořízení, rozvoj, provoz a návratnost investic	36	45
Funkcionalita	39	29
Rozšiřitelnost	5	4
Uživatelská přívětivost	4	3
CELKEM	**88**	**84**

Zdroj: Vlastní

Z ohodnocení vyplývá, že technologická platforma Splunk s 88 body je lepším řešením pro bankovní sektor než platforma Colix se 84 body.

Je nutné ale podotknout, že obě řešení získala v daném testu vysoký počet bodů, takže se jedná a vyspělé platformy, které splňují všechny kladené požadavky.

Pomocí řešení Splunk budou v následujících kapitolách zpracována nestrukturovaná či semistrukturovaná data a podrobně popsán postup práce s tímto systémem.

6. Zpracování dat technologickou platformou Splunk

Následující část bakalářské práce je věnována samotnému postupu zpracování a vizualizace dat prostřednictvím technologické platformy Splunk, která byla zvolena jako nejvhodnější dle kritérií uvedených v předchozí kapitole.

6.1 Vstupní data

K dispozici je nepřeberné množství různých zdrojů, ale pro potřeby bakalářské práce je nutné využít data volně dostupná.

Jako vstupní data pro platformu Splunk byl vybrán volně dostupný dataset z virtuálního internetového obchodu Buttercup Games (www.buttercupgames.com) umožňující mapovat chování zákazníků, oblíbenost produktů a další.

Tento dataset je volně k dispozici na docs.splunk.com. Dataset obsahuje logy, které pochází z webového serveru společnosti Buttercup Games a obsahuje informace o aktivitě, přístupu uživatelů, o prodejích atd..

Ukázka vstupních dat v RAW podobě:

112.111.162.4 - - [23/Mar/2016:18:26:37] "GET /category.screen?categoryId=NULL&JSESSIONID=SD7SL8FF5ADFF4964 HTTP 1.1" 505 2445 "http://www.buttercupgames.com/category.screen?categoryId=NULL" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.5 (KHTML, like Gecko) Chrome/19.0.1084.52 Safari/536.5" 393

112.111.162.4 - - [23/Mar/2016:18:26:38] "GET /oldlink?itemId=EST-7&JSESSIONID=SD7SL8FF5ADFF4964 HTTP 1.1" 503 1207 "http://www.buttercupgames.com/category.screen?categoryId=NULL" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.5 (KHTML, like Gecko) Chrome/19.0.1084.52 Safari/536.5" 704

74.125.19.106 - - [23/Mar/2016:18:32:15] "GET /cart.do?action=addtocart&itemId=EST-16&productId=DC-SG-G02&JSESSIONID=SD4SL7FF10ADFF4998 HTTP 1.1" 200 1425 "http://www.buttercupgames.com" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-GB; rv:1.8.1.6) Gecko/20070725 Firefox/2.0.0.6" 375

74.125.19.106 - - [23/Mar/2016:18:32:15] "GET /category.screen?categoryId=NULL&JSESSIONID=SD4SL7FF10ADFF4998 HTTP 1.1" 503 2039 "http://www.buttercupgames.com/oldlink?itemId=EST-13" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-GB; rv:1.8.1.6) Gecko/20070725 Firefox/2.0.0.6" 375

Ukázka obsahuje čtyři události, které zaznamenaly přístup a aktivitu na virtuálním internetovém obchodě Buttercup Games.

6.2 Postup zpracování dat technologickou platformou Splunk

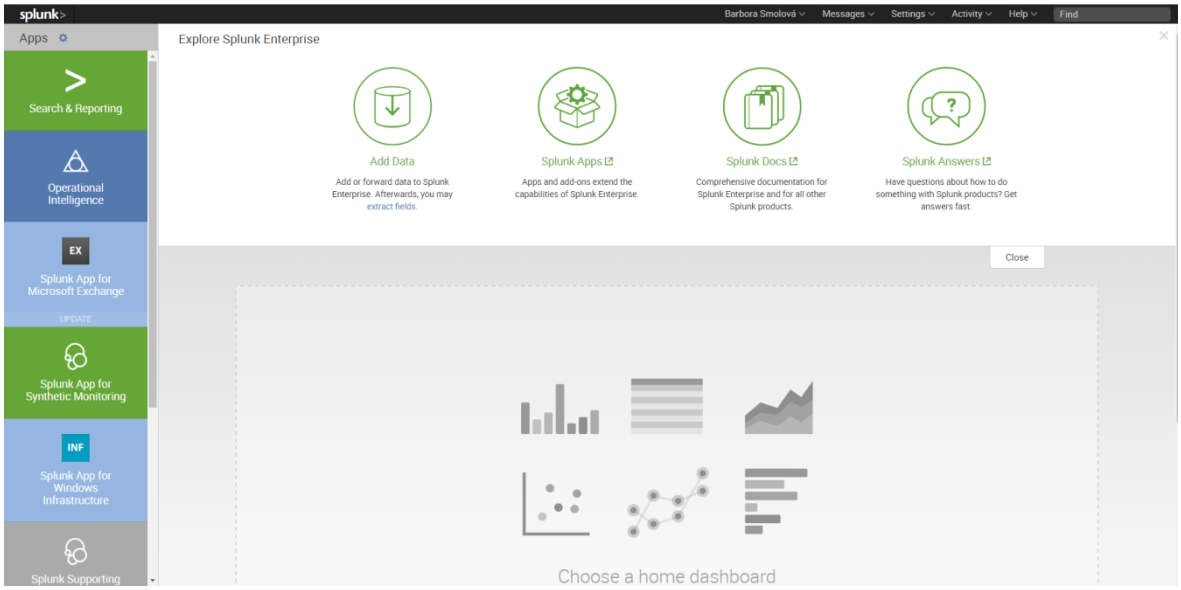
Postup zpracování dat zvolenou technologickou platformou Splunk je popsán v následujících kapitolách.

6.2.1 Proces sběru dat, nahrání dat

Nastavení procesu sběru dat či nahrání dat je prvním krokem pro zpracování dat. Tuto činnost zprostředkovává tzv. ***Splunk> heavy forwarder.*** Princip sbírání dat je popsán v teoretické části práci bakalářské práce.

V případě zpracování dat z internetového obchodu Buttercup Games je potřeba nahrát již existující logy do platformy Splunk.

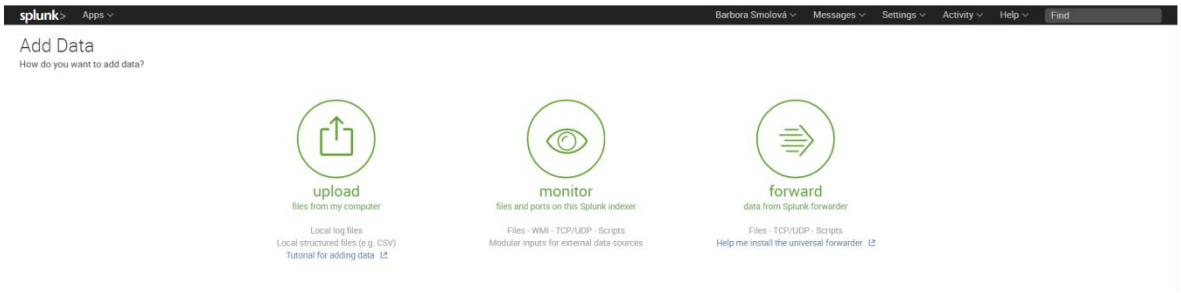
Po autentizaci do aplikace Splunk se zobrazí úvodní stránka aplikace, viz obrázek 28.



Obrázek 27: Splunk – Úvodní stránka platformy
Zdroj: Vlastní

Úvodní interaktivní stránka obsahuje navigační lištu a v levé části je umístěno menu aplikací, které obsahují definované vizualizace.

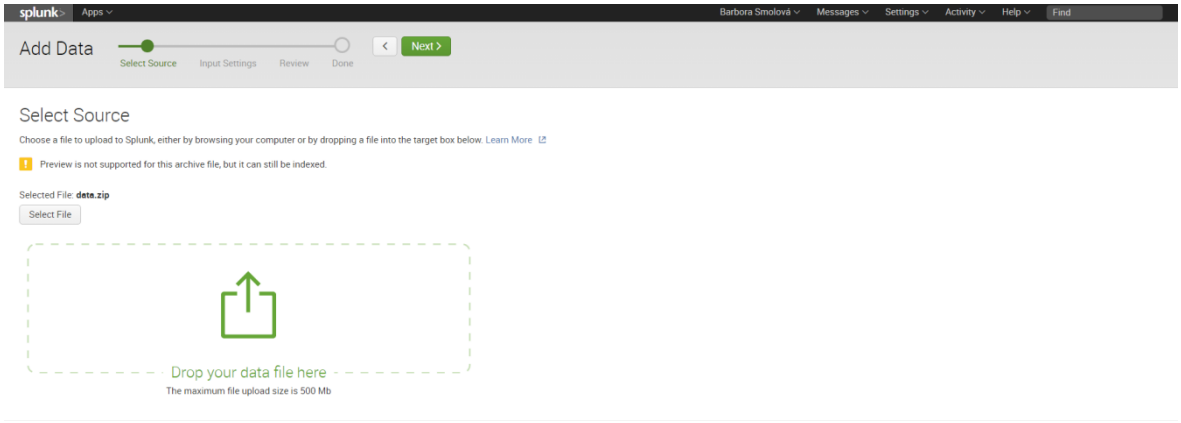
V případě potřeby nahrání dat je v navigačním panelu k dispozici volba Add Data viz obrázek 28. Následně se zobrazí možnosti upload, monitor a forward (obrázek 29).



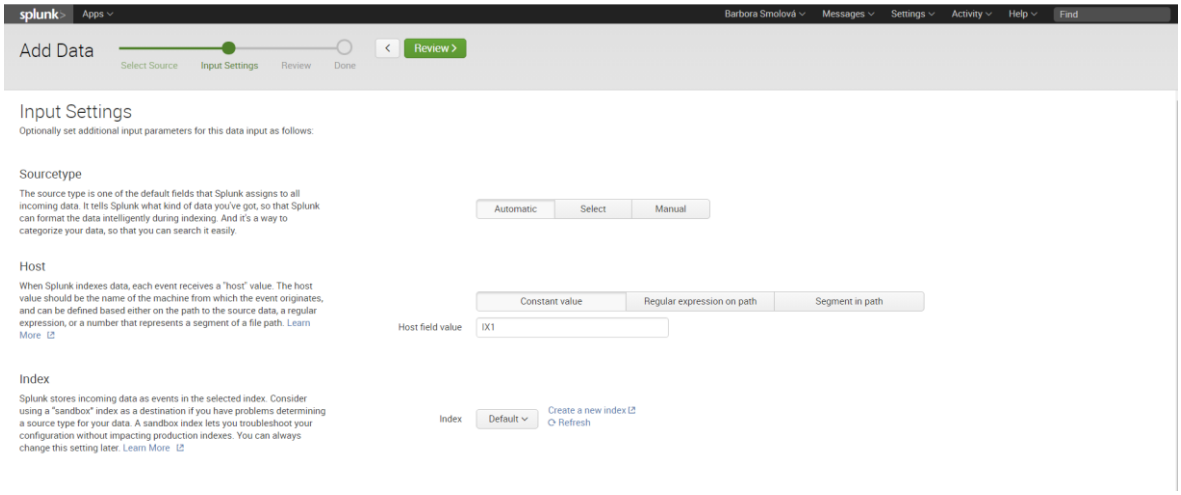
Obrázek 28: Splunk – Přidání dat – navigační panel
Zdroj: Vlastní

Pro nahrání již existujícího datasetu je využívána volba Upload.

Následujícím krokem je zvolení souboru, který bude nahrán a jeho bližší specifikace. Viz obrázek 30, 31 a 32.



Obrázek 29: Splunk – Volba dat ke zpracování
Zdroj: Vlastní



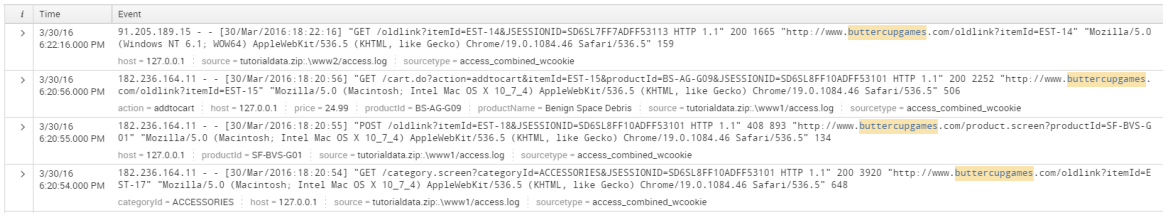
Obrázek 30: Splunk – Bližší specifikace nahrávaných dat
Zdroj: Vlastní

Obrázek 31: Splunk – Základní informace o souboru k nahrání
Zdroj: Vlastní

Po absolvování výše zmíněných kroků jsou logy nahrány do platformy Splunk.

U tohoto typu vstupních dat (webové a databázové logy) není nutné dodatečné parsování dat. Splunk automaticky rozpozná typ a rozparseruje tento zdroj (obrázek 33).

Vstupní data v technologické platformě Splunk:

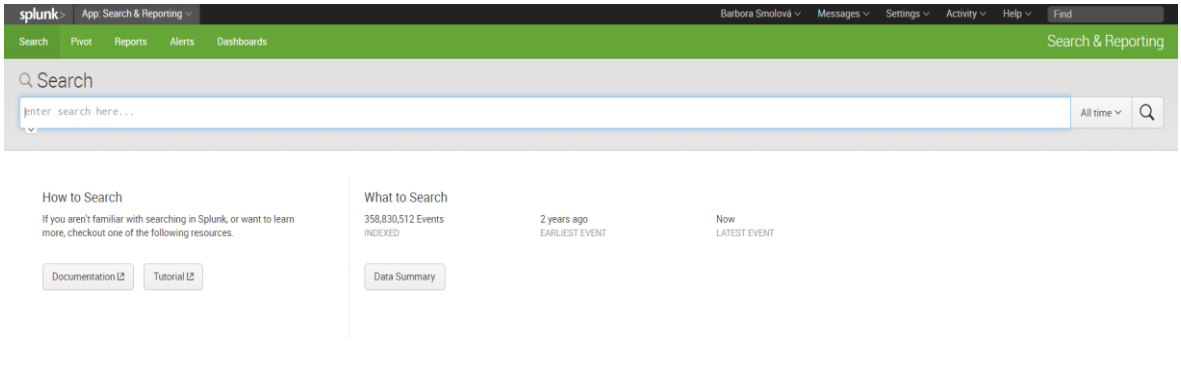


Obrázek 32: Splunk – Vstupní logy
Zdroj: Vlastní

6.2.2 Vyhledávání, filtrování, manipulace, vkládání a mazání dat

Řešení Splunk využívá jazyk Search Processing Language (SPL), pomocí něhož vyhledává, filtruje, upravuje, manipuluje, vkládá a maže data. Jeho syntaxe je založena na jazyku SQL a tzv. unixových rourách.

Pro tvorbu reportů slouží aplikace Search & Reporting, která je již v základu platformy nainstalována viz obrázek 34.



Obrázek 33: Splunk – Aplikace Search & Reporting
Zdroj: Vlastní

6.2.3 Výsledky vyhledávání – Statistiky

Platforma Splunk nabízí i možnost zobrazení výsledků transformace vyhledávání v záložce Statistiky viz obrázek 35. Dle syntaxe v poli Search je možné vytvořit různé statistiky, např. zobrazit si kolik produktů se v jaké kategorii nachází, z jakých prohlížečů se zákazníci nejvíce připojují, nejoblíbenější produkt a nespočet dalších.

Například:

- Počet prodaných produktů v jednotlivých kategoriích a jejich procentuální podíl (obrázek 35)

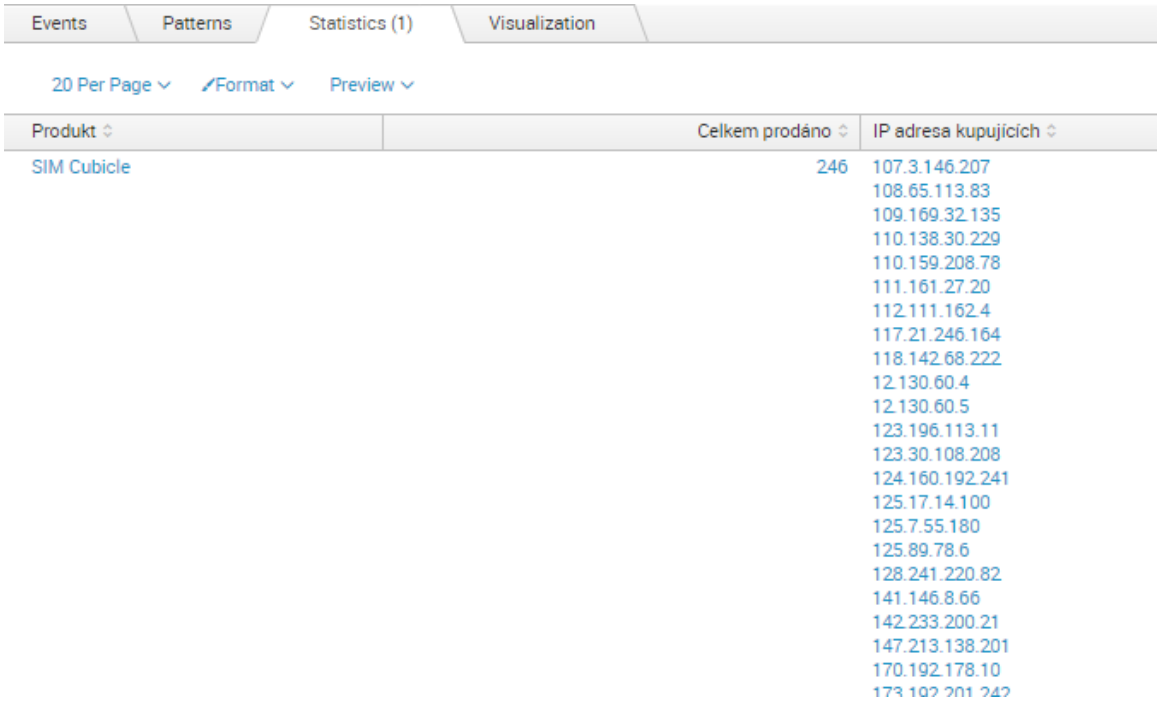
Events | Patterns | Statistics (7) | Visualization

20 Per Page | Format | Preview

Název kategorie	Prodaných produktů (v ks)	% podíl prodaných produktů
STRATEGY	806	30.495649
ARCADE	493	18.653046
TEE	367	13.885736
ACCESSORIES	348	13.166856
SIMULATION	246	9.307605
SHOOTER	245	9.269769
SPORTS	138	5.221339

Obrázek 34: Splunk – Statistika – kategorie produktů
Zdroj: Vlastní

- Název nejprodávanějšího produktu e-shopu, kolik kusů se prodalo a IP všech kupujících (obrázek 36)



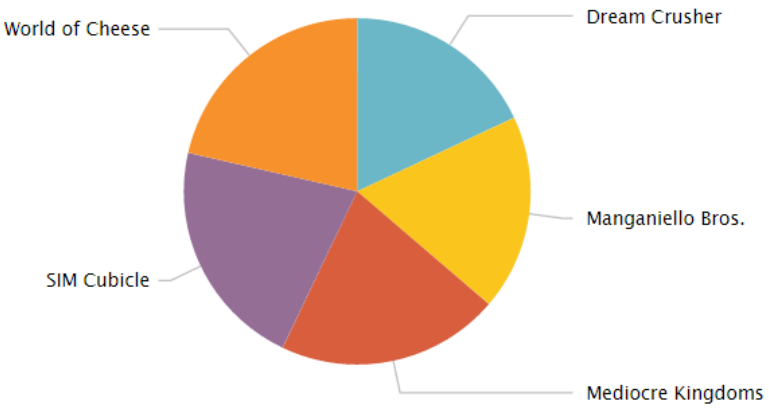
Obrázek 35: Splunk – Statistika – nejprodávanější produkt
Zdroj: Vlastní

6.2.4 Vizualizace

Platforma Splunk umožňuje vytvářet jak velmi jednoduché vizualizace, tak i ty pokročilé např. Bubble chart, Scatter plot a další.

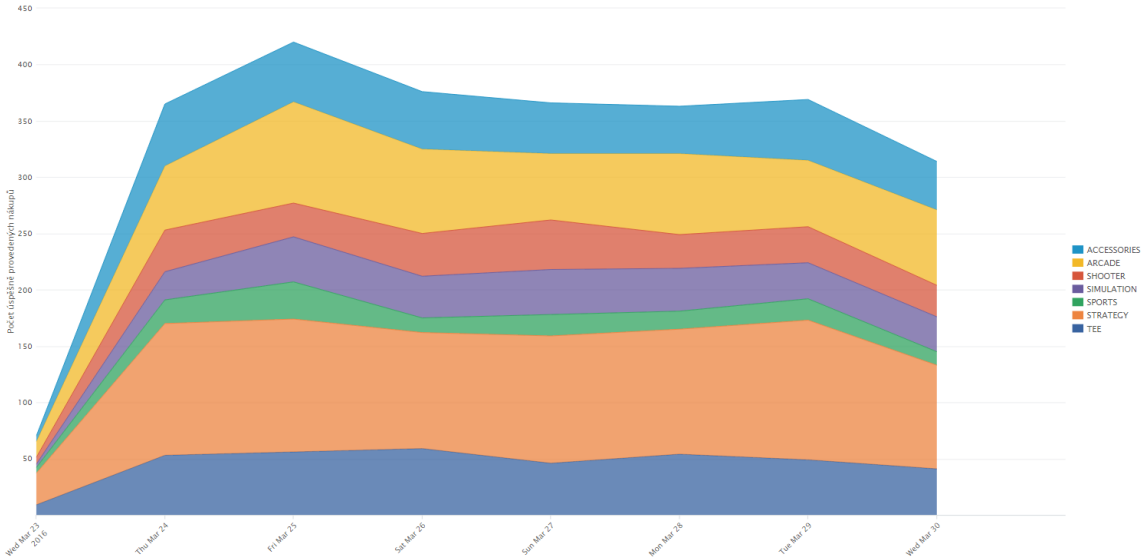
Na základě vzoru dat, která byla do platformy Splunk nahrána, je možné vytvořit velké množství různých reportů. Například:

- Výsečový graf - pět nejvíce prodávaných produktů internetového e-shopu Buttercup Games (obrázek 37).



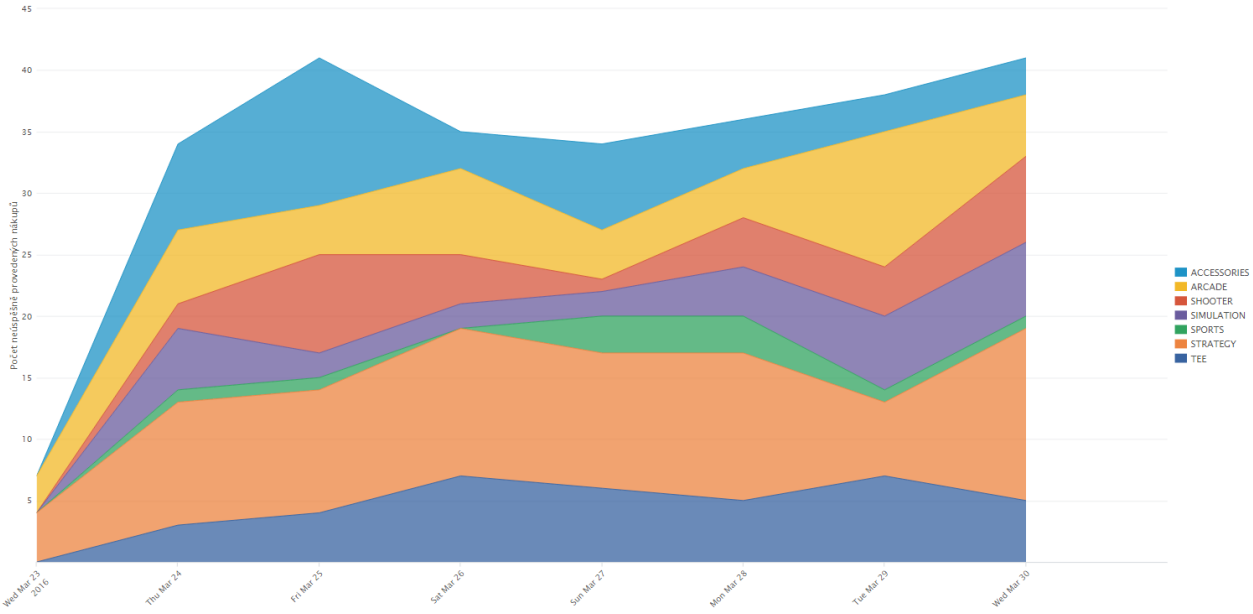
Obrázek 36: Splunk – Výsečový graf – pět nejvíce prodávaných produktů
Zdroj: Vlastní

- Vývoj počtu úspěšně provedených nákupů, rozděleného dle kategorií za období 23. března – 30. března 2016 (obrázek 38).



Obrázek 37: Splunk – Plošný graf – počet úspěšně provedených nákupů
Zdroj: Vlastní

- Vývoj počtu neúspěšných pokusů o nákup, rozdělených dle kategorií za období 23. března – 30. března 2016. Plošný graf je zobrazen na obrázku 39.



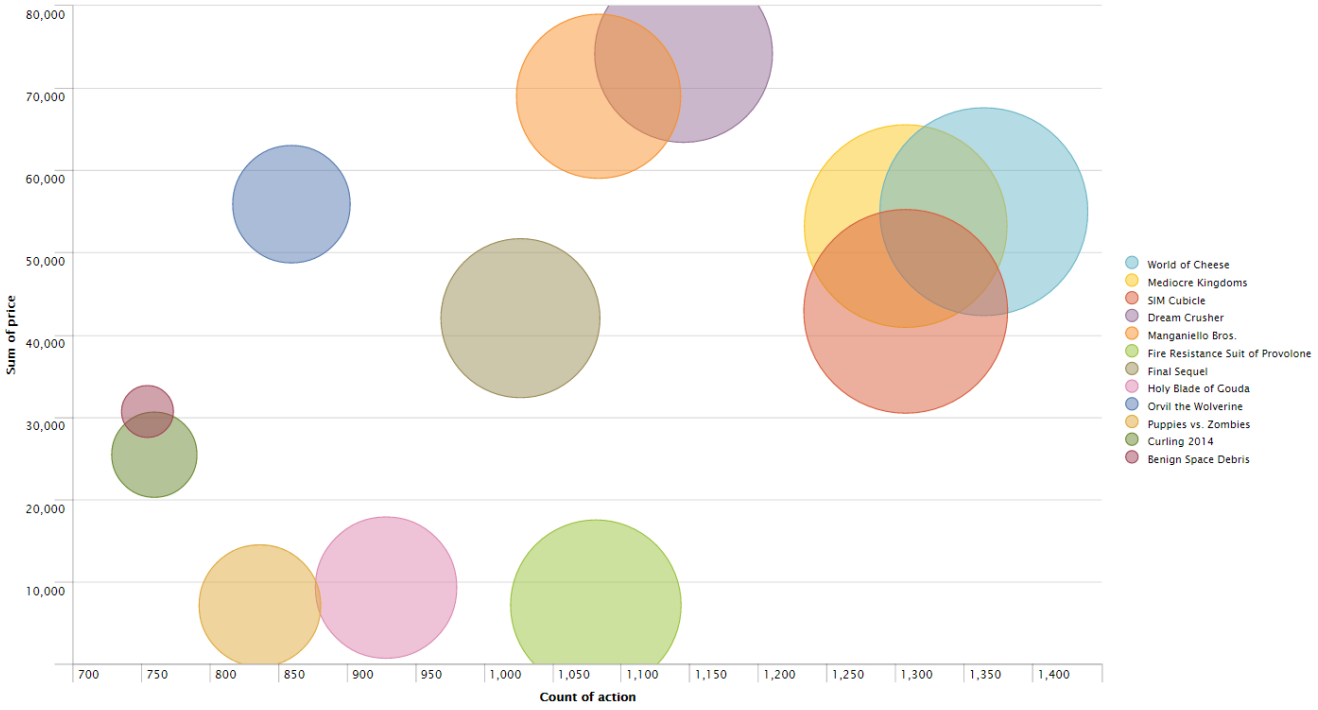
Obrázek 38: Splunk – Plošný graf – počet neúspěšných pokusů o nákup
Zdroj: Vlastní

- Mapa znázorňuje, jaké hry byly nakoupeny a z jakých zemí Evropské unie (obrázek 40).

Obrázek 39: Splunk – Geografická mapa
Zdroj: Vlastní

6.2.4.1 Pokročilé vizualizace

Jak již bylo zmíněno v předešlé kapitole, technologická platforma Splunk umožňuje tvorbu pokročilých datových vizualizací. Oproti technologické platformě Colix umožňuje Splunk tvorbu více jak dvou dimenzionálních grafů, např. Bubble chart viz obrázek 41 níže.



Obrázek 40: Splunk – Bubble chart
Zdroj: Vlastní

Tento Bubble chart znázorňuje jaká je závislost mezi počtem interakcí, které uživatel provede na celkové ceně, kterou na e-shopu utratí.

7. Zhodnocení přínosu

Zpracování dat technologickou platformou Splunk, které je popsáno v předchozí kapitole, názorně ukazuje jak jednoduchá a intuitivní je práce s touto platformou.

I uživatel laik se velmi rychle zorientuje v prostředí této platformy. V případě jakýchkoliv potíží Splunk poskytuje velmi kvalitní a podrobnou dokumentační základnu, včetně tutoriálů.

Toto řešení je i velmi uživatelsky přívětivé. Není nutné mít speciální znalosti či zkušenosti např. zkušenosti se skriptovacími jazyky.

Velmi zjednodušuje například nahrávání dat. Tato platforma má již předpřipravené skripty pro většinu běžně dostupných typů dat a má tudíž předem definováno, jak daný typ dat bude parsovat. Uživateli tím odpadne povinnost této volby a tím se mu výrazně zjednoduší práce s touto platformou.

Hlavním přínosem řešení Splunk je velmi jednoduché a rychlé zpracování dat v různých formátech a z různých zdrojů (včetně dat strojových, textových atd.). Pro potřeby vizualizace platforma nabízí širokou paletu interaktivních grafů, metrik a map (např. multidimenzionální graf zobrazený na obrázku 41 či geografická mapa na obrázku 40), na základě kterých zprostředkovává vhled do informací, které jsou v nestrukturovaných a semistrukturovaných datech obsažené a které by nebylo možné bez využití platformy takto efektivně vytěžit.

Závěr

Hlavní cíl práce je zaměřen na charakteristiku technologických platforem pro vytěžování a vizualizaci Big Dat, které splňují v práci formulovaná kritéria. Vybraná řešení jsou charakterizována na základě funkcionality, zprostředkovatele, klientů a podporovaných operačních systémů. Všechna BI technologická řešení popsaná v této práci využívají metodu transformace nestrukturovaných a semistrukturovaných dat na data strukturovaná.

Z charakterizovaných platforem jsou vybrána dvě řešení, která jsou podrobněji porovnána.

Do podrobnější komparace byly zařazeny platformy Splunk a Colix. Daná řešení byla porovnána na základě výkonu, nákladů, funkcionality a uživatelské přívětivosti. Tato stanovená kritéria jsou dle autorky rozhodující pro výběr vhodné BI technologické platformy. Dle provedeného porovnání těchto platforem, je pro zvolený bankovní sektor vhodnější varianta pro zpracování nestrukturovaných a semistrukturovaných dat platforma Splunk. Především kvůli svým rozsáhlým funkcionalitám např. SIEM řešení, které Colix neposkytuje.

Dílčím cílem práce je ukázka práce s daty pomocí vybrané platformy. Zpracování dat proběhlo na platformě Splunk s ukázkou možných grafických výstupů.

Současný stav těchto technologií je na vysoké úrovni. Technologické platformy, které zde byly popsány, jsou schopné velmi snadno zpracovat a vizualizovat všechny běžně dostupné druhy dat ve velkých objemech a v reálném čase.

Neustálým vývojem nových zařízení a technologií se budou objemy dat dále zvyšovat a bude přibývat i množství formátů dat, které je potřeba zpracovat. Lze předpokládat, že úměrně k tomuto faktu se vývoj těchto technologií pro zpracování Big Dat bude v následujících letech dále zdokonalovat.

Je velmi pravděpodobné, že s rostoucím objemem dat se bude zvyšovat i možnost jejich zneužití. Z tohoto důvodu se bude dále zvyšovat úroveň zabezpečení.

Seznam použité literatury

[1] TAKMI: Bringing Order to Unstructured Data. In: *IBM 100* [online]. New York, United States: IBM Corporation, 2015 [cit. 2016-05-02]. Dostupné z: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/takmi/

[2] AGGARWAL, Charu C. *Data Mining: The Textbook*. 1. Cham: Springer International Publishing AG, 2015. ISBN 978-3-319-14142-8.

[3] MERRET, Rebecca. 5 tools and techniques for text analytics. In: *CIO Australia: Information Technology strategy insight for senior IT management* [online]. Boston: IDG Communications, 2016 [cit. 2016-05-02]. Dostupné z: http://www.cio.com.au/article/575209/5-tools-techniques-text-analytics/

[4] RUSSELL, Matthew A. *Mining the social web*. 2nd ed. Beijing: O'Reilly, 2014. ISBN 9781449367619.

[5] ČERNÝ, Michal. Vizualizace dat: Jak odhalit utajené souvislosti. In: *VTM.cz: věda, technika, technologie, budoucnost* [online]. Praha: Serafico investment, 2016 [cit. 2016-05-02]. Dostupné z: http://vtm.e15.cz/vizualizace-dat-jak-odhalit-utajene-souvislosti

[6] Big Data s pomocí vizualizace? *SystemOnLine: ekonomické a informační systémy v praxi* [online]. Brno: Webservis, ©2001-2016 [cit. 2016-05-02]. Dostupné z: http://www.systemonline.cz/business-intelligence/big-data-s-pomoci-vizualizace.htm

[7] MAREK, Tomáš. *Od datové vizualizace k datovému umění*. Brno, 2012. Bakalářská diplomová práce. Masarykova univerzita, Filozofická fakulta. Vedoucí práce Mgr. Jan Boček.

[8] *VisualComplexity.com: A visual exploration on mapping complex network* [online]. Agoura Hills: GreenGeeks.com, ©2008-2016 [cit. 2016-05-02]. Dostupné z: http://www.visualcomplexity.com/

[9] THORP, Jer. Goodmorning!: Information Visualization - Twitter. In: *Blprnt.blg: Jer Thorp* [online]. San Francisco: Automattic, ©2005-2016 [cit. 2016-05-02]. Dostupné z: http://blog.blprnt.com/blog/blprnt/goodmorning

[10] MARR, Bernard. Why only one of the 5 Vs of big data really matters. In: *IBM Big Data and Analytics Hub: Understanding big data for the enterprise* [online]. New York, United States: IBM Corporation, 2015 [cit. 2016-05-02]. Dostupné z: http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters

[11] SUŠICKÝ, Marek a Petr MIKEŠKA. *IT Systems: časopis pro manažery, kteří vědí, že informační technologie jsou konkurenční výhodou*. Brno: CCB, 2015, **17**(11), 18-19. ISSN 1802-002X.

[12] NEEF, Dale. *Digital exhaust: what everyone should know about big data, digitization and digitally driven innovation*. Indianapolis, IN: Pearson, 2014. ISBN 9780133837964.

[13] MOLNÁR, Zdeněk. *Competitive intelligence: aneb jak získat konkurenční výhodu*. 1. Praha: Oeconomica, 2012. ISBN 978-80-245-1908-1.

[14] ROUSE, Margaret a Ivy WIGMORE. Semi-structured data. In: *Computer Glossary, Computer Terms: The Tech Dictionary and IT Encyklopedia* [online]. Newton: TechTarget, c1999-2016 [cit. 2016-05-02]. Dostupné z: http://whatis.techtarget.com/definition/semi-structured-data

[15] CRIPE, Billy. Structured and Unstructured Data: What Are They (Fusion ECM). In: *Oracle: Integrated Cloud Applications and Platforms Services* [online]. Redwood City: Oracle, ©1995-2016 [cit. 2016-05-02]. Dostupné z: https://blogs.oracle.com/fusionecm/entry/structured_and_unstructured_da

[16] GÁLA, Libor, Jan POUR a Zuzana ŠEDIVÁ. *Podniková informatika: počítačové aplikace v podnikové a mezipodnikové praxi*. 3., aktualizované vydání. Praha: Grada Publishing, 2015. ISBN 9788024754574.

[17] POUR, Jan, Miloš MARYŠKA a Ota NOVOTNÝ. *Business intelligence v podnikové praxi*. Praha: Professional Publishing, 2012, 276 s. ISBN 9788074310652.

[18] SAAGAR, KSHIRA. 5 Methods for Visualizing Unstructured Data. In: *Information Management: How Your Business Work* [online]. New York, United States: SourceMedia, 2016 [cit. 2016-05-02]. Dostupné z: http://www.information-management.com/news/news/5-methods-for-visualizing-unstructured-data-10025220-1.html

[19] WARD, Matthew O., Georges G GRINSTEIN a Daniel KEIM. *Interactive data visualization: Foundations, Techniques, and Applications*. Second edition. Boca Raaton: CRC Press, Taylor & Francis Group, 2015. ISBN 9781482257373.

[20] MELOUN, Milan. *Počítačová analýza vícerozměrných dat v oborech přírodních, technických a společenských věd* [online]. Brno, 2011 [cit. 2016-05-02]. Dostupné z: http://www.crr.vutbr.cz/system/files/brozura_05_1106.pdf. Učební texty ke kurzu. Vysoké Učení Technické v Brně.

[21] *Operational Intelligence, Log Management, Application Management, Enterprise Security and Compliance: Splunk* [online]. California: Splunk Inc., ©2005-2016 [cit. 2016-05-02]. Dostupné z: www.splunk.com/

[22] *Continuous Intelligence, Log Management & Analysis: Sumo Logic* [online]. Redwood City: Sumo Logic, 2015 [cit. 2016-05-02]. Dostupné z: https://www.sumologic.com/

[23] Vnitropodniková data společnosti Trask solutions, a. s.

[24] The Elastic Stack: Make Sense of Your Data. *Elastic: Revealing Insights from Data* [online]. Mountain View, California: Elasticsearch, 2016 [cit. 2016-05-02]. Dostupné z: https://www.elastic.co/products

[25] *Log Analysis Software With Enterprise ELK: Logz.io* [online]. New York: Logz.io, 2016 [cit. 2016-05-02]. Dostupné z: http://logz.io/

[26] *FluentD - Open source Data Collector: Unified Logging Layer* [online]. Mountain View, United States: FluentD Project, c2010-2016 [cit. 2016-05-02]. Dostupné z: http://www.fluentd.org/

[27] PERI, Noni. Fluentd vs. Logstash: A Comparison of Log Collectors. In: *Log Analysis Software With Enterprise ELK: Logz.io* [online]. New York, United States: Logz.io, 2016 [cit. 2016-05-02]. Dostupné z: http://logz.io/blog/fluentd-logstash/

[28] BUDÍN, Emil. K čemu je SIEM? In: *SystemOnline.cz: ekonomické a informační systémy v praxi* [online]. Brno: Webservis, ©2001-2016 [cit. 2016-05-02]. Dostupné z: http://www.systemonline.cz/it-security/k-cemu-je-siem.htm

[29] TŘMÍNEK, Tomáš. Přínosy a náklady Business Intelligence. In: *SystemOnline.cz: ekonomické a informační systémy v praxi* [online]. Brno: Webservis, ©2001-2016 [cit. 2016-05-02]. Dostupné z: http://www.systemonline.cz/business-intelligence/prinosy-a-naklady-business-intelligence.htm

Bibliografie

ILIINSKY, Noah P a Julie STEELE. *Designing data visualizations*. Sebastopol, CA: O'Reilly, c2011. ISBN 9781449317430.

TAN, Pang-Ning, Michael STEINBACH a Vipin KUMAR. *Introduction to data mining*. 2. Harlow, United Kingdom: Pearson Education Limited, 2013. Pearson new international edition. ISBN 9781292026152.

SIKORA, Radek. *Vyhledávání v českých dokumentech pomocí Apache Solr*. Brno, 2012. Diplomová práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce RNDr. Radek Ošlejšek, Ph.D.

WILLIAMSON, Jason. *Getting a big data job for dummies*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2015. ISBN 978-1-118-90340-7.

ZADROZNY, Peter a Raghu R. KODALI. *Big data analytics using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real–Time Streaming Sources*. 2013 ed. Berkley, United States: aPress, 2013. ISBN 978-1-4302-5761-5.