Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Analýza a vizualizace dat z ubytovacích portálů

Plzeň 2017 Miroslav Havlíček

Místo této strany bude zadání práce.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 4. dubna 2017

Miroslav Havlíček

Abstract

The text of the abstract (in English). It contains the English translation of the thesis title and a short description of the thesis.

Abstrakt

Text abstraktu (česky). Obsahuje krátkou anotaci (cca 10 řádek) v češtině. Budete ji potřebovat i při vyplňování údajů o bakalářské práci ve STAGu. Český i anglický abstrakt by měly být na stejné stránce a měly by si obsahem co možná nejvíce odpovídat (samozřejmě není možný doslovný překlad!).

Obsah

1 Úvod

Hlavním úkolem této bakalářské práce je zanalyzovat a vizualizovat data z ubytovacích portálů. Tato data jsou uložena v databázi Elasticsearch, která byla zvolena, protože se jedná o nestrukturovaná data. Data představují recenze z hotelů v okolí Dubaje, která byla získána z ubytovacích portálů. Tyto recenze byly zpracovány skupinou NLP, která působí na Fakultě aplikovaných věd Západočeské univerzity. Úkolem této práce tudíž není připravit data a uložit je do databáze, ale využít tato data k zodpovězení otázek, které zaslala Zayedova univerzita, která si získání a zpracování dat z ubytovacích portálů objednala.

Databáze Elasticsearch byla zvolena díky možnosti ukládání takzvaných Big dat, pro které je typická jejich rozsáhlost a běžně dostupné databázové nástroje nejsou schopny zpracovat v reálném čase, především kvůli velkému množství vazeb mezi daty. První částí této bakalářské práce je prozkoumání různých nástrojů sloužících k vizualizaci dat, které jsou dostupné na trhu. Tyto nástroje jsou nejčastěji součásti aplikací Bussiness Intelligence, což jsou aplikace, které slouží k analýze, reportování a vizualizaci velkého množství dat, které podniky sbírají a následně je využívají jako podporu pro strategické rozhodování. Důležitým kritériem pro výběr nástrojů k prozkoumání je podpora ukládání a zpracování nestrukturovaných dat. Nestrukturovaná data jsou běžně reprezentována fotografiemi, zvukovými nahrávkami, daty ze senzorů, ale také například souvislým textem, což získané recenze jsou. S výběrem databáze Elasticsearch souvisí i výběr vizualizačního nástroje, kterým je nástroj Kibana, který je stejně jako Elasticsearch od společnosti Elastic. Tento nástroj slouží k vizualizaci dat z nástroje Elasticsearch skrze webové uživatelské rozhraní.

Dalším úkolem bakalářské práce je tvorba webových stránek, které budou zobrazovat vybrané vizualizace, které v rámci bakalářské práce vzniknou. Jedná se především o zprostředkování nástroje Kibana běžným uživatelům bez možnosti tvorby vizualizací. Hlavní funkcionalitou by mělo být snazší porovnání vybraných hotelů podle zadaných kritérií, než je tomu u nástroje Kibana, který není pro běžné uživatele snadno ovladatelný.

První část práce je zaměřena na porovnání dostupných nástrojů a jejich detailnější představení. Druhá část bakalářské práce se věnuje funkcím nástroje Kibana a jejich využití k analýze a vizualizaci dostupných dat. Další část bakalářské práce je tvořena popisem webových stránek, které jsou v rámci bakalářské práce vypracovány. Poslední částí je zhodnocení práce, za

kterou následuje slovníček pojmů, ve kterém je možné dohledat vysvětlení odborných termínů a zkratek.

2 Nástroje k analýze a vizualizaci dat

První kapitola práce je zaměřena na nástroje, pomocí kterých je možné analyzovat a vizualizovat semi-strukturovaná popřípadě nestrukturovaná data. Na trhu je mnoho nástrojů určených pro zpracování firemních dat a většinou jsou součástí komplexnějšího řešení ve formě BI nástroje. BI je zkratka pro Business Intelligence, což je rámocvý termín pro škálu aplikací, které slouží k analýze nestrukturovaných dat organizace.[3]

2.1 InetSoft Style Intelligence

Jedná se o produkt firmy InetSoft Technology Corporation, která se zaměřuje na vývoj BI aplikací. Společnost nabízí aplikace určené k tvorbě podnikových reportů a k vizualizaci vlastních firemních dat. Tyto aplikace jsou založené na webových technologiích a například aplikace Vizualize Free a Style Scope Free Edition jsou zdarma. Při vývoji se především používá XML, což je obecný značkovací jazyk pro uchování a přenos dat, SOAP, což je protokol pro přenos XML zpráv přes síť, Java a JavaScript[1]. Hlavně díky použití programovacího jazyka Java, který je dnes celosvětovou jedničkou mezi programovacími jazyky [27] jsou aplikace snadno integrovatelné s jinými aplikacemi, které jsou dostupné jako open-source. Open-source aplikace jsou distribuovány s licencí, která dovoluje upravovat původní zdrojový kód aplikace a následně již upravenou verzi aplikace dále distribuovat jako open-source.

2.1.1 Přehled

InetSoft Style Intelligence je velmi silný nástroj, který umožňuje zpracovat a kombinovat data z různých zdrojů. Systém je nabízen jako cloud aplikace nebo jako řešení na míru. Cloudové aplikace jsou v dnešní době oblíbeným řešením, protože nekladou nároky na hardware uživatele a navíc jsou všechny operace prováděny na serverech poskytovatele těchto aplikací. Výhodou cloudu je, že uživatelé platí jen za výpočetní kapacitu, kterou skutečně využijí a jedná se tak o finančně výhodnější řešení, nežli hosting, kdy se platí za celý server, i když není zcela využitý. Řešením na míru je například produkt InetSoft Style, který je nasazen přímo na server subjektu, a společnost

InetSoft garantuje minimální zatížení serveru.

2.1.2 Technologie

Kolekce dat

Základem je technologie Data Block^{TM}, která zprostředkovává shromažďování dat z různých zdrojů. Data z databáze získává tento nástroj přímo z originálních úložišť pomocí předem připravených materializovaných pohledů. Tyto pohledy představují databázové objekty, ve kterých je uložen výsledek nějakého dotazu, takže je přístup k němě rychlejší než provádění nového dotazu.

Zpracování dat

Po kolekci a přípravě dat je možné z dat tvořit různé vizualizace, které slouží k lepšímu pochopení dostupných dat. InetSoft Style poskytuje uživatelům dashboardy, které se v pravidelně aktualizují podle dat, které jsou k dispozici, a ze kterých je dashboard tvořen. Dashboardy mohou být tvořeny jednoduchými grafy, složitějšími multidimenzionálními grafy, které jsou vhodné pro porovnání hodnot, a nebo například geografickými mapami. InetSoft Style zároveň obsahuje sadu předpřipravených dashboardů například pro oblast zdravotnictví, vzdělávání nebo podpory prodeje a nákupu.

2.1.3 Kompatabilita

Firma tvrdí, že její produkt je schopný převzít data z téměř jakéhokoliv zdroje, což zvládá díky jejich data mashup engine. Zákazník tedy může použít jak strukturovaná tak semi-strukturovaná nebo dokonce nestrukturovaná data. InetSoft Style Intelligence podporuje export do běžných formátů, díky čemuž jsou data a grafy využitelné například v aplikacích balíčku MS Office. Jelikož se jedná o službu založenou na webové technologii, tak je dostupná na všech hlavních prohlížečích na běžných systémech jako jsou Windows, Unix, Linux, Mac OS X, HP-Unix, Solaris a další.[11]

2.1.4 Hodnocení a ceník

Hlavní výhodou tohoto produktu je uživatelsky přívětivé prostředí, kde uživatelé tvoří vizualizace svých dat. To je zapříčiněno především přehledností webové stránky, ze které se nástroj ovládá. Jednotlivá tlačítka zlehčují ovladatelnost systému a také systém drag and drop, který umožňuje používat

nástroj bez nutnosti znalosti dotazovacího jazyka. Velkým pozitivem je i možnost ovládat nástroje přes mobilní telefon, ale podle mého názoru je to vhodné jen k zobrazování již vytvořených analýz a dashboardů. Dalším pozitivem je i možnost nahlédnout přímo do původních dat pouhým kliknutím na libovolný prvek vizualizace. Výhodou je taktéž to, že produkt není určen primárně pro technicky vzdělané uživatele, ale je určen pro běžné uživatele bez nutnosti podpory IT oddělení. Balíček služeb je dostupný již od 2 800 dolarů.[17] Pozitivní je, že produkty Vizualize Free a Style Scope Edition jsou zdarma na neurčitou dobu. Oba produkty jsou založeny na cloud technologii a akceptují pouze strukturovaná data.

2.2 Splunk Enterprice

Software byl vytvořen firmou Splunk Inc., která se zaměřuje na vývoj podpůrného softwaru pro vyhledávání, analyzování a monitoring strojových dat skrze webové rozhraní.[8] Splunk Enterprise patří mezi základní produkty této firmy, ale ta dále nabízí cloudovou verzi Enterprise pod názvem Splunk Cloud^{TM}a Splunk Light, což je nástroj na monitoring logů pro menší IT subjekty. Základní verze těchto produktů je možno rozšířit o další nadstavby a vytvořit tak komplexní řešení na míru každého zákazníka. Mezi nadstavby patří produkty Splunk Enterprise Security, který se zaměřuje na kolekci a analýzu dat získaných ze zabezpečovacích technologií, Splunk IT Service Intelligence, což je nástroj na sledování funkčnosti IT systémů (sleduje podezřelé aktivity systému, výkon systému a předem definované kritické částí systému), a Splunk User Behaviour Analytics, který používá strojové učení k detekci potencionálních hrozeb a kyberútoků. Právě díky nástrojům vhodným ke zlepšení kyberbezpečnosti spolupracuje firma s americkou vládou skrze zprostředkovatelskou firmu.[28]

2.2.1 Přehled

Produkt Splunk Enterprise je vhodný pro sběr, analýzu a úpravu strojových dat ve velkém množství (tzv. Big Data). Tyto data mohou být generována různými interními systémy uživatele ve formě serverových logů, aplikačních logů, dat o výrobě, logů ze sociálních sítí a podobně. Software je určen primárně pro zpracování semi-strukturovaných a nestrukturovaných dat z non-SQL databází.

2.2.2 Technologie

Kolekce dat

Velkou výhodou tohoto řešení je nezávislost na formě vstupních dat, protože ty jsou zpracována a zaindexována do formy, kterou produkty od firmy Splunk vyžadují, zcela automaticky. Nedochází avšak k normalizaci dat, ale data jsou uchovávána v raw podobě na které odkazují metadata v souborech s indexy.

Vyhledávání

Vyhledávání je zprostředkováno vlastním query jazykem, který se jmenuje Search Processing Language neboli SPL™. Jazyk je velmi rozsáhlý a obsahuje více než 140 příkazů. Výsledky z vyhledávání jsou interpretovány do vhodných interaktivních grafů, které jsou zvoleny přímo aplikací na základě množství a formy dat. Vyhledává se buďto pomocí SPL™nebo přes velké množství filtrů a případně kombinovaně. [14]

Zpracování dat

Jelikož je vytvořen soubor s indexy, tak Splunk Enterprise zvládá korelaci a analýzu z různých zdrojů najednou, což značně urychluje práci analytikům. K tvorbě modelů a k predikci anomálií v chování systému používá produkt strojové učení. Modely lze vytvářet přímo skrze webový prohlížeč pomocí speciálních příkazů jazyka SPL™. Jako základ strojového učení jsou použity knihovny programovacího jazyka Python. [19]
Data se vizualizují pomocí různých uživatelsky přizpůsobitelných grafů, které se následně skládají do interaktivních dashboardů, které lze posléze exportovat ve formě HTML. Výhodou je, že lze nastavit uživatelská práva jednotlivým dashboardům, přiřadit jim ovládací prvky a následně je sdílet s ostatními kolegy skrze společný pracovní prostor.

2.2.3 Kompatibilita

Splunk je zaměřen především na zpracování big dat, ale je možné ho připojit i na relační databáze, nebo ho propojit s tabulkovým procesorem Microsoft Excel či produktem Tableau, který je podrobně popsán v sekci 1.3. Splunk lze rozšiřovat vlastními aplikacemi, nebo využít databázi Splunkbase, kde jsou již vytvořené aplikace a rozšíření, které umožňují lepší integraci a vizualizaci. K aplikaci je také možné se připojit přes mobilní zařízení a kontrolovat tak chod sledovaného systému. Splunk je možné provozovat pod operačními

systémy Linux, Windows 7 a novější a pod operačním systémem Mac OS X. [8]

2.2.4 Hodnocení a ceník

Ke každému produktu jsou k dispozici zkušební verze, které jsou omezeny jak časově, tak množstvím přenesených dat. Od množství přenesených dat za den se odvíjí i cena produktu, ale je k dispozici produkt Splunk Free, který je určen pro jednotlivce a má omezené funkce monitoringu a strojového učení. Velkou výhodu spatřuji v existenci vlastní komunity, která zprostředkovává možnost přímo se zeptat uživatelů na přínosy, popřípadě je to místo, kde hledat pomoc při problémech se softwarem od společnosti Splunk Inc.

2.3 Tableau Desktop

Jedná se o produkt americké softwarové společnosti Tableau Software, jenž se zaměřuje na vývoj softwaru vhodného pro vizualizaci dat a na nástroje business intelligence. Firma vznikla na základě výzkumu v oblasti vizualizace dat. Konkrétně se jednalo o projekt Polaris, který byl veden katedrou počítačových věd Standfordské univerzity.[9] Mezi hlavní produkty firmy patří kromě Tableau Desktop taktéž Tableau Server, který je určený především pro spolupráci napříč organizací, Tableau Online, což je cloudová verze produktu Tableau Server, Tableau Public, který je určen pro jednotlivce na nekomerční užití a proto je zdarma, a Tableau Reader, který je taktéž volně dostupný a slouží k prohlížení a manipulaci vizualizací vytvořených některým produktem Tableau. Produkty od firmy Tableau jsou dnes především využívané datovými žurnalisty, kteří oceňují jeho snadnou ovladatelnost.

2.3.1 Přehled

Tableau Desktop je nástroj k analýze a vizualizaci vlastních dat a také o nástroj business intelligence. Produkt je nabízen ve verzi Personal, která je určena osobnímu použití a vstupní data musejí být strukturovaná, a ve verzi Professional, která zvládá i nestrukturovaná Big Data. Výhodné je spojení s produktem Tableau Server, aby bylo možné vytvořené dashboardy a ostatní vizualizace sdílet s kolegy. Předností produktu Tableau Desktop je uživatelská přívětivost, protože k jeho běžnému používání nejsou nutné žádné pokročilé technologické znalosti, ale stačí používat systém tvorby vizualizací drag and drop.

2.3.2 Technologie

Kolekce a příprava dat

Aby Tableau Desktop umožňovala používání dat z více zdrojů, je nutné jejich předzpracování, na což využívá vlastní nástroj a uživatele tak příliš nezatěžuje, protože sám hledá vztahy mezi jednotlivými zdroji dat. Zdroje dat rozlišuje do dvou hlavních skupin, a to na soubory (například z MS Excel, MS Access, textových souborů, logů a podobně) a na databázové servery, na které je možné produkt Tableau Desktop přímo napojit a data se mohou přenášet buďto přímým spojením, nebo technologií in-memory, kdy se přenáší jen virtuální obraz dat a Tableau tak pracuje rychleji bez nutnosti odesílání velkého množství dotazů serveru.
Z projektu Polaris vznikl dotazovací jazyk VizQL™. Jedná se o vizuální query jazyk, který převádí jednotlivé drag and drop příkazy na dotazy.[15] Výhodou je lepší ovladatelnost pro méně technicky zdatné uživatele, protože se nemusí orientovat v dotazovacích jazycích a přímo mohou vidět výsledky vizualizací pomocí systému drag and drop.

Zpracování dat

Data jsou rozdělena na dvě kategorie, na dimenzionální (jména, regiony) a kvantitativní data (množství, prodeje, zisk ...). Výhodou jsou filtry dat, které jsou aplikovatelné na více různých zdrojů dat najednou. U Tableau je ceněno seskupování dat, což je funkce, která automaticky seskupí data, které mají společné vlastnosti (například geografickou polohu, symptomy nemocí) a následně lze tyto data v grafu zvýraznit, či je přesunout do nově vzniklého grafu.[4]
Tableau nabízí mnoho analytických nástrojů a umožňuje provádět výpočty nad daty při tvorbě dashboardů. U dashboardu je po zveřejnění možnost ho hodnotit, popřípadě okomentovat, čehož následně využívají již zmíněné analytické nástroje, které tvoří žebříček oblíbených, případně trendy dashboardů. Obrovská výhoda Tableau je správa dashboardů. Kromě nastavování uživatelských práv lze i sledovat vývoj dashboardu pomocí interního verzovacího systému.[21]

2.3.3 Kompatibilita

Tableau Desktop nabízí možnost využít základní aplikaci pro modifikovatelný přístup k datům a vizualizacím a jelikož je napsána v JavaScriptu je možné ji dále rozšiřovat dle vlastních potřeb. Pomocí této aplikace lze vizualizace exportovat do jiných programů především z rodiny MS Office.

Aplikace je založena na webových technologiích, takže je kompatibilní s běžnými distribucemi Windows 7 a vyšší a s Mac OS X 10.10 a novější. Verzi personal je možné propojit se zdroji dat typu MS Access popřípadě přímo s textovými soubory ve formátu CSV popřípadě JSON. CSV soubory jsou určeny pro výměnu tabulkových dat, které obsahují data, která jsou od sebe vzájemně oddělena oddělovačem (například čárkou, středníkem, tabulátorem atd.). JSON „neboli JavaScript Object Notation je formát souborů určený k výměně dat.“[10] Je snadno čitelný i zapisovatelný člověkem a zároveň lze soubory ve formátu JSON snadno generovat i zpracovávat strojově. Tableau Desktop verze professional je již propojitelná s většinou databázových systémů, které se dnes používají (např. Oracle Databases, PostgresSQL, Cloudera Hadoop Hive and Impala, Cisco Information Server).[26]

2.3.4 Hodnocení a ceník

Cena se pohybuje od 999$ za osobní verzi, až po 1999$ za profesionální edici Tableau Desktop.[18] Zajímavostí je, že firma nezapomněla na své kořeny ze Stanfordské univerzity a je tak pro všechny studenty a vyučující k dispozici zdarma bez omezení. Verze Tableau Public, která je zdarma pro širokou veřejnost má podobné vizualizační nástroje jako placené verze, ale podporuje zpracování pouze již strukturovaných dat. Stejně jako u produktu Splunk Enterprise má Tableau vlastní komunitu, která je podporována přímo firmou Tableau a kam přispívají jak jednotliví zaměstnanci, tak i zákazníci. K dispozici je taktéž mobilní aplikace, která je primárně určená pro prohlížení vytvořených dashboardů.

2.4 Sisense

Tento software je produktem stejnojmenné firmy, která se zaměřuje na nástroje buisiness intelligence. Jejich řešení jsou komplexní a obsahují jak nástroje na sběr dat, tak nástroje na jejich vizualizaci. Společnost je uváděna jako lídr oblasti poskytovatelů buisiness intelligence nástrojů a její produkty se v očích odborné veřejnosti tak i v očích uživatelů jeví jako velmi spolehlivé.[12]

2.4.1 Přehled

Sisense je koncový BI nástroj, který byl vyvinut pro uživatele, kteří nemají téměř žádné zkušenosti s BI nástroji a nemusejí být technicky zdatní. Aplikace nabízí nástroje pro správu, text mining a pro interaktivní analýzu dat,

která jsou uložena v databázi ElastiCube, což je podpůrný produkt Sisense. Výhodou Sisense je velmi snadná ovladatelnost pomocí systému drag and drop, bez nutnosti znalosti query jazyka.

2.4.2 Technologie

Kolekce a příprava dat

Sisense nevyžaduje časově náročnou fázi předzpracování dat, tak jako ostatní BI nástroje na trhu. Data z různých zdrojů jsou sjednocena a importována do jednotného úložiště, což usnadňuje práci s daty a nevyžaduje nákup speciálních programů na přípravu dat, popřípadě spouštět skripty nad daty. Databáze ElastiCube je sloupcově orientovaná a je tvořena mnoha poli, kde každá hodnota v polích má odpovídající logickou hodnotu v jiném poli a tím je databáze propojena. Proto je vhodné používat tento typ uložení při velkém množství dat, nebo pokud jsou data z různých zdrojů.
Výhodou je také zpracování jednotlivých dotazů, které nejsou na rozdíl od běžných technologií zpracovány jako celek, ale jsou rozděleny do bloků, které se následně vyhodnocují. To je výhodné časově, protože při jednotném zpracování dotazů musí, CPU při každé změně v dotazu opětovně vyhodnotit celý dotaz, ale při blokovém musí zpracovat jen tu část dotazu, která byla změněna.[6]
Rychlost, za kterou je produkt velmi ceněn je dána způsobem zpracování dat, který je na rozdíl od ostatních produktů na trhu zajištěn technologií in-chip. Tato technologie je v databázových systémech jedinečná a jejím základem je maximalizace využití paměti, kterou poskytují jednotlivé CPU, protože přístup do této paměti je výrazně rychlejší, než přístup do paměti RAM, která je využívána technologií in-memory.

Zpracování dat

Vizualizace je umožněna širokou paletou grafů a geografických map, které lze skládat do interaktivních real-time dashboardů. Tyto dashboardy se následně mohou sdílet napříč organizací a sledovat jeho vývoj a případně upravit uživatelská práva. Každý dashboard obsahuje ovládací panely s filtry a lze zde přímo upravit znění dotazů a přizpůsobit tak vizualizace dle aktuálních potřeb. Aplikace již obsahuje některé dashboardy předpřipravené, což je určeno především pro nezkušené uživatele.[29]

2.4.3 Kompatibilita

Nabízí předpřipravené nástroje na import dat z Excelu, Google Adwords, Salseforce, CRM reports, Splunk bez nutnosti složitého importu dat, což je opět určeno především pro netechnické uživatele, kteří dokáží pomocí systému drag and drop jak importovat a zpracovat data, tak je následně vizualizovat a popřípadě exportovat například do formátů CSV, PDF, Excel a podobně. „PDF je formát používaný k prezentaci a spolehlivé výměně dokumentů, který je nezávislý na softwaru, hardwaru i operačním systému.“[16] Aplikace Sisense je založena na webových technologiích a je tak kompatibilní s běžnými operačními systémy jako jsou Windows, Android, Mac OS X. Splunk nabízí možnost obohatit produkt o řadu předpřipravených rozšíření, nebo je možné přímo vyvíjet rozšíření v JavaScriptu, kdy je ovšem doporučeno toto rozšíření konzultovat s komunitou.[25]

2.4.4 Hodnocení a ceník

Cena softwaru se odvíjí od velikosti společnosti a od velikosti zpracovávaných dat, ale je k dispozici až na konkrétní dotaz. K dispozici je i demo verze, která již obsahuje data a je tak možné si vyzkoušet funkcionalitu systému před jeho zakoupením. Výhodou je napojení systému na mobilní telefony. Co se týče zpracování strojových dat, sám výrobce doporučuje propojení Sisense s produkty od firmy Splunk. Stejně jako Tableau a Splunk, podporuje Sisense vlastní komunitu.[13]

2.5 Kibana

Kibana je open source nástroj na vizualizaci dat od firmy Elastic a byla vytvořena jako plugin do fulltextového vyhledávače Elasticsearch, který vychází z Apache Lucene. Elasticsearch je NoSQL bezschémová databáze, což znamená, že není nutné definovat předem přesnou strukturu databáze, ale databáze sama nastaví schéma podle dat, která obsahuje. NoSQL jsou databáze nové generace vhodné pro zpracování velkého množství dat.[7] Přesto se u Elasticsearch doporučuje zvolení alespoň základního schématu, což může usnadnit následující analýzy nad daty.[2] Oba nástroje jsou součástí produktu Elastic Stack, který je jediným produktem firmy Elastic a je dostupný jako open source, ale společnost dále nabízí rozšířené funkce, které je ovšem již nutné zakoupit. Elastic Stack se tak skládá z Elasticsearch, Kibany, Logstash a Beats.

2.5.1 Přehled

Kibana byla původně vytvořena jen jako plugin k Elasticsearch, ale nyní je z ní plnohodnotná součást celku Elastic Stack a je hojně využívána především pro monitorování logů ze serverů. Kibana umožňuje spojit data z databáze do komplexních grafických prvků, ze kterých je snazší datům přiřadit význam. Díky rychlosti databáze Elasticsearch a jejím možnostem fulltextového vyhledávání můžeme aplikaci Kibana nazývat real-time nástrojem, pokud jsou tedy správně zaindexována data, které obsahuje databáze. Problémem může být pro některé uživatele ovladatelnost, která není tak jednoduchá jako například u Tableau i přes přítomnost systému drag and drop. Na druhou stranu je oceňované rychlé fulltextové vyhledávání buďto skrze všechna data, nebo skrze data, která si pomocí query dotazu vybereme.

2.5.2 Technologie

Kolekce a příprava dat

Jelikož je Kibana součástí produktu Elastic Stack, který obsahuje i Elasticsearch, nemusí řešit ukládání dat, ale stačí definovat indexování, které odpovídá indexům v Elasticsearch. Indexy lze přidat i za běhu aplikace v záložce „Settings“. Kibana, kterou požívám při vypracování bakalářské práce je ve verzi 5.1.1 a obsahuje pouze jeden index a to konkrétně „hospitality“.

Prohlížení dat

Data lze přímo prohlížet skrze Kibana na záložce „Discover“, kde jsou zobrazena všechna data přidaná za časový úsek, který lze měnit v pravém horním rohu stránky. Data jsou zobrazena v základu v tabulce, kde jsou rozepsány jednotlivé skupiny dat, nebo je možné zobrazit soubor formátu JSON, který je přímo uložený v databázi. Jednotlivé záznamy lze řadit podle data přidání a je možné změnit rozložení tabulky vybráním specifických skupin dat ze seznamu, který je po levé straně webové stránky.

Zpracování dat

Záložka „Vizualize“ slouží ke grafickému vyjádření dat v Elasticsearch například ve formě plošných grafů, spojnicových grafů, kruhových grafů nebo například jako geografickou oblast. Jednotlivé vizualizace lze ukládat do komplexnějších dashboardů, které lze upravovat na záložce „Dashboards“. Dashboardy lze tvořit z vizualizací, které vznikly na základě různých indexů, což nebylo v předchozích verzích možné a hlavně jedna vizualizace

může být využita ve více dashboardech, protože existuje jako samostatný objekt.
Filtry lze použít přímo v jednotlivých objektech, nebo nad celými daty. Filtrování lze provádět pomocí speciálních query dotazů zadávaných do vyhledávacího pole nad vizualizacemi, daty či dashboardy. Dále je možné filtry využít již při vytváření vizualizace a poté je jen pomocí ovládacích objektů v horní části vizualizace ovládat, nebo využít filtrů, které nabízí Kibana na základě struktury dat, která byla použita při vizualizaci.[23]

2.5.3 Kompatibilita

Kibana je napsána v JavaScriptu a využívá architektury klient-server. Jelikož patří do Elastic Stack, který je distribuovaným systémem, tak je snadno integrovatelná s jinými nástroji. Například od verze 5 lze přímo skrze Kibanu vkládat do Elasticsearch data ve formátu CSV, což umožňuje integraci s nástrojem Microsoft Excel. Problémová je kompatibilita se staršími verzemi Kibany, kdy je naprosto nutné upgradovat nejprve Elasticsearch, protože každá verze má jinou strukturu clusterů, což je kolekce serverů, která shromažďuje data a poskytuje indexování a vyhledávání přes jednotlivé servery.

2.5.4 Verze 5

V říjnu roku 2016 vyšla verze 5.0, kdy Kibana doznala především grafických změn, které na její funkčnost nemají velký význam. Ke změnám funkcionalit lze řadit nezobrazování histogramu záznamů u dat, která neobsahují časovou značku, což na rozdíl od předchozí verze šetří čas při zobrazování logů. Velmi podstatnou změnou je odstranění linku nad jednotlivými záznamy v databázi, které mohli způsobovat bezpečnostní riziko, jelikož odkaz vracel reálný JSON soubor uložený v Elasticsearch skrze prohlížeč a volání GET. Toto se změnilo a nyní je k dispozici pouze náhled, který neodkazuje na JSON a je tak bezpečnější ho sdílet. Novinkou je také generování krátkých URL odkazujících na vizualizace či dashboardy, což je více uživatelsky přívětivé, než předchozí dlouhé URL adresy. Novinkami jsou i nové nabídky na hlavní liště, která byla přesunuta na levou stranu a je skryta, když není používána, což zvětšuje prostor k zobrazení vizualizací. Navíc přibyly nástroje Timelion a Console, což dříve byli jen pluginy. Timelion je nástroj, který umožňuje sledovat změny v čase v jednotlivých datech. Console je nástroj, který usnadňuje psaní query dotazů. Dotazy lze skládat do jednoho souboru, dělit je podle indexů a nástroj má dokonce funkci našeptávače, který ještě více usnadňuje psaní dotazů. Co se týče možností rozšiřování kódu, doznala

verze 5 také vylepšení, protože jednotlivé položky hlavního menu byly rozděleny do samostatných pluginů, které se dají lépe upravovat. Další novinkou je možnost upravovat názvy os u jednotlivých vizualizací, jelikož se v nabídce objevil nový parametr s názvem „Custom Label“. U filtrů, které lze připnout nad vizualizace taktéž přibyla možnost přímé úpravy query dotazu a lze tak ovlivnit chování daného filtru. [22]

3 Nevím název

V druhé části mé práce je mým úkolem prostudovat otázky, které se vztahují k dostupným datům z ubytovacích portálů a nalézt vhodnou formu jejich vizualizace. Tato kapitola popíše syntaxi a použití filtrů a query dotazů, které jsou pro vlastní vizualizaci potřebné, dále pak popisuje rozšíření aplikace Kibana, nástroje na vizualizaci, které Kibana poskytuje, a také samozřejmě použití těchto nástrojů při vizualizaci.

3.1 Query DSL

Jak již bylo zmíněno, Elasticsearch využívá syntaxi dotazovacího jazyka Lucene, nebo případně dotazovacího jazyka založeného na JSON k definování jednotlivých dotazů nad daty. Existují dva typy dotazů a to **Leaf query clauses** a **Compound query clauses**. První jmenovaný typ dotazů slouží k vyhledávání přesné hodnoty v předem určené oblasti dat. Do této skupiny dotazů řadíme dotazy obsahující výrazy *match*, *term* a *range*. Druhý typ dotazů slouží ke kombinování více dotazů v logickou posloupnost nebo k ovlivňování výsledků filtrů či dotazů. Tato třída obsahuje dotazy s výrazy, jako jsou *bool*, *dis_max* nebo *constant_score*. Chování obou klauzulí záleží také na tom, jestli jsou použity v kontextu dotazu nebo filtru. Při použití v kontextu dotazu rozhoduje klauzule, jestli záznam odpovídá dotazu a navíc vypočítává skóre, které vyjadřuje, jak moc odpovídá v porovnání s ostatními záznamy. Pokud je klauzule použita v kontextu filtru, rozhoduje klauzule jen o tom, jestli záznam odpovídá dotazu a nebo ne. [20]

3.1.1 Full textové dotazy

Tato třída dotazů se využívá hlavně na záznamy obsahující souvislý text. Zkoumají, jak byl záznam analyzován a následně podle toho aplikuje stejnou metodu analyzování na výraz, který dotaz obsahuje. Jelikož data z ubytovacích portálů obsahují převážně textové řetězce, je vhodné se s touto třídou seznámit. Tato třída dotazů se dělí na:

Match Query

Dotazy typu boolean, které akceptují parametry typu text, číslo a datum. To znamená, že je text analyzován a následně vytvořen booleovský dotaz

obsahující v základu logický operátor OR.

Match Phrase Query

Dotazy založené na match query, které slouží k vyhledávání přesných frází. Zadaná fráze se nejprve analyzuje, následně se sestaví jednotlivé dotazy formou více vnořených match query dotazů a nakonec se provede vlastní dotaz nad daty.

Match Phrase Prefix Query

Obdobně jako předchozí typ se zaměřuje na vyhledávání frází, kdy ovšem hledá záznamy obsahující text, jehož předponou je právě dotazovaná fráze. Pomocí parametru *max_expansions* lze nastavit maximální počet znaků, které následují za prefixem a omezit tak možné výsledky.

Multi Match Query

Jedná se o rozšířené dotazy typu match query, kdy lze vyhledávat ve více oblastech najednou. Pro názvy oblastí lze použít zástupné znaky, které mohou určovat prioritu oblasti, nebo název (např. „*“ je používána jako zástupný znak pro libovolný počet znaků, nebo „?“ je používán pro zvýšení priority). Výsledek dotazu ovlivňuje parametr typ, kdy jednotlivé typy využívají jinou interpretaci hodnoty skóre, kterou vrátí dotaz match query.

Common Terms Query

Tyto dotazy rozdělí zadaný dotaz na 2 skupiny a to na skupinu s vysokou důležitostí, kam se řadí výrazy z dotazu, které se v záznamech neobjevují příliš často, a pak na skupiny s nižší důležitostí, kde jsou výrazy, které jsou v záznamech velmi časté. Nejprve dojde ke zpracování skupiny s vyšší důležitostí, což způsobí vyfiltrování jen relevantních dokumentů a po zpracování druhé skupiny se počítá celkové skóre jen těchto dokumentů, což značně ovlivňuje výkon vyhledávání.

Query String Query

Dotazy, které plně odpovídají syntaxi Lucene Query Parser Syntax, specifikují logické operátory, jako jsou AND, OR a NOT a zároveň umožňují provádět vyhledávání přes více polí jedním dotazem. Zadaný dotaz se rozdělí na jednotlivé výrazy a operátory, přičemž výrazy mohou být jednoslovné anebo mohou obsahovat fráze. Pokud chceme, aby výraz obsahoval fráze, je

nutné v původním dotazu tuto frázi psát uvnitř uvozovek. Defaultně nastavený logický operátor je OR, ale lze to změnit použitím parametru „default_operator“ do struktury dotazu.
Tento typ dotazů má také definovanou skupinu znaků, které nelze použít samostatně ve vlastním podobě dotazu. Tato skupina dotazů se nazývá rezervované znaky a je tvořena následujícími znaky: „+, -, =, &&, ||, >, <, !, ( ), { } ∧, [ ],∼ ,* ,? ,: ,/, \“. Pokud je nutné některý z těchto znaků v dotazu použít, musí se před něj vložit tzv. escape znak, který indikuje výskyt znaku z množiny rezervovaných znaků. Jedná se tedy o surjekci znaků, které se v dotazu mohou vyskytovat do množiny povolených znaků a escape sekvencí. Například při hledání hotelů, jejichž celkové skóre je vetší nebo rovno 8 a zároveň menší nebo rovno 9 a jejichž skóre neplaceného WiFi připojení je vetší nebo rovna 6, by struktura dotazu ve formátu JSON vypadala následovně:

```
{
"query":{
 "query_string":{
  "query":"SCORE_TOTAL:[8 to 9] AND
   SCORE_FREE_WIFI:>=6"
  }
 }
}
```

Simple Query String Query

Obdobné dotazy jako query string s tím rozdílem, že tyto dotazy nikdy nevrátí výjimku způsobenou nesprávným zápisem dotazu, protože tyto části ignoruje. Zároveň nahrazuje logické operátory zástupnými znaky a zjednodušuje tak uživatelům psaní dotazů. Stejně jako u query_string je nastavený defaultní operátor na hodnotu OR.

3.2 Vyhledávání v Kibana

Aplikace Kibana má v základu k dispozici pole, kam lze vkládat prosté řetězce, nebo jednoduché dotazy, které respektují syntaxi Query String Query. Do tohoto pole lze taktéž vkládat dotazy ve formátu JSON, který je popsán v části Query DSL. Toto pole je v aplikaci k dispozici na záložkách „Discover“, „Visualize“ a „Dashboards“. Před samotným vyhledáváním je důležité porozumět funkci analyzátoru, který analyzuje data při přidání mezi záznamy databáze.

3.2.1 Funkce analyzátoru

Chování analyzátoru závisí na mapovaní indexu, do kterého jsou data nahrávána. Pokud není mapování nastaveno, má Elasticsearch schopnost detekovat jakého typu jsou hodnoty daného pole. Většinou volí typ „text“, který následně analyzuje skrze zvolený analyzátor. Pro Query String Query a fulltextové vyhledávání rozeberu funkci analyzátoru na polích, které jsou typu „text“, nebo ve starších verzích typu „string“. S analyzátorem je spjato několik tokenizátorů, které získají hodnotu z pole a následně ji rozdělí na jednotlivé výrazy podle mezer nebo interpunkce. Jednotlivé tokeny, které jsou výsledkem tokenizace mohou být upraveny (například převedeny na malá písmena) popřípadě využity jako hodnoty pro filtry, které jsou taktéž s analyzátorem spjaté. Výsledkem práce analyzátoru jsou tedy jednotlivé tokeny, které jsou ukládány do takzvaného invertovaného indexu, který navíc ještě obsahuje odkaz na záznam, ze kterého byly tokeny získány. Výhodou je, že se při vyhledávání jednoho výrazu nemusí procházet všechny záznamy v databázi, ale projdou se jen tokeny v invertovaném indexu a výsledkem jsou záznamy, které byly spjaty s daným tokenem. Pokud nechceme, aby bylo pole analyzováno, stačí nastavit parametr mapovaní „index“ na hodnotu „not_analyzed“. Poté neproběhne rozdělení textu na tokeny a vyhledání je citlivé na velikost písmen. U neanalyzovaného pole také není možné vyhledávat podle slov, ale jen podle přesných frází, které pole obsahuje jako hodnoty.

3.2.2 Fulltextové vyhledávání

Pokud do tohoto pole vložíme řetězec, který není v souladu s používanou syntaxí, dojde k fulltextovému vyhledávání nad defaultně nastaveným polem s hodnotami. V základu je toto pole v aplikaci Kibana nastaveno na hodnotu „_all“, což znamená, že vyhledávání proběhne ve speciálním invertovaném indexu, který obsahuje tokeny ze všech záznamů, které kdy byly přidány. Elasticsearch totiž při přidání záznamu rozdělí hodnoty podle mapování, ale zároveň si uloží celý vstup jako jeden dlouhý řetězec, který následně analyzuje pomocí standardního analyzátoru a vytvoří invertovaný index s tokeny, který známe jako pole „_all“. Pokud chceme vyhledávat fráze napříč všemi záznamy v databázi, je nutné uzavřít dotaz do uvozovek, aby nedošlo k rozdělení celkového dotazu na jednotlivé výrazy, protože pak by výsledky vyhledávání byly nerelevantní.

3.2.3 Dotazy ve formátu Query String Query

Hlavní výhodou těchto dotazů je možnost omezení výsledků na vybraná pole. Realizace této restrikce je velmi jednoduchá, protože stačí znát přesný název vybraného pole a za něj zadat frázi, popřípadě výraz. Jedná se o například dotaz HOTEL_NAME: „Villa Rotana“, který zobrazí záznamy obsahující v poli HOTEL_NAME hodnotu Villa Rotana. Při psaní dotazů lze taktéž využít následující prvky:

Zástupné znaky

Při psaní dotazů není nutné psát přesný název jednotlivých polí nebo celé výrazy, ale lze využít zástupných znaků, které jsou k dispozici. Jedná se o znaky „?“ a „*“, kdy první znak nahrazuje právě jeden znak na zadaném místě v řetězci a druhý znak nahrazuje celou skupinu znaků, která může být i prázdná. Výjimkou, kdy nelze využít zástupných znaků, jsou fráze, protože analyzátor tyto znaky nenahradí a hledá záznamy, které obsahují řetězec, který přesně odpovídá zadané frázi.
Pokud tedy využiji dotaz, který vyhledává záznamy obsahující zmínku o hotelu „Villa Rotana“ a zároveň použiji zástupné znaky, bude dotaz vypadat například následovně *HOTEL_NAME: „Villa Rot*“* a výsledkem budou všechny záznamy, které obsahují sekvenci „Villa Rot“, která může být následována libovolnou posloupností znaků.

Logické operátory

Jako každý dotazovací jazyk, i jazyk Query DSL, konkrétněji Query String Query, používá pro spojení částí dotazů do větších dotazů logické operátory. Hodnota defaultního operátoru je nastavena na hodnotu OR, což znamená, že se dotaz *HOTEL_NAME: „Villa Rotana“ HOTEL_NAME: „Royal“* přeloží na dotaz *HOTEL_NAME: „Villa Rotana“ OR HOTEL_NAME: „Royal“* , takže výsledkem budou všechny záznamy z databáze, které obsahují v poli HOTEL_NAME hodnoty Villa Rotana, nebo Royal.
Důležité je, že logické operátory se musejí psát velkými písmeny, jinak jsou brány jako součást vyhledávané fráze a ne jako spojovací výraz.
Jednotlivé části dotazu lze spojovat nejen operátory AND a OR, ale také je lze sdružovat do skupin pomocí závorek, nebo nahradit operátor AND znaky && a OR znaky ||.
Další operátory, pomocí kterých je také možné ovlivnit chování dotazu, jsou znaky „+“ a „-“, které se vkládají před vybranou část dotazu. Operátor plus zapříčiní, že vybraný výraz se musí vyskytovat v záznamech a další výrazy v

dotazu, které nejsou označeny operátorem plus jsou pouze volitelným doplňkem prvního výrazu, takže se v dokumentu nemusejí vyskytovat. Exkluze, neboli vyloučení, je možné vyjádřit hned několika operátory a to „-“, „!“ nebo výrazem NOT. Stejně jako u operátoru plus je nutné tyto operátory psát před vybraný výraz.
Ekvivalentem tohoto zápisu jsou dotazy typu match query. Například dotaz *quick OR brown AND fox AND NOT news* lze přepsat následovně:

```
{
"bool":{
 "must":        {"match":"fox"},
 "should":      {"match":"quick brown"},
 "must_not":    {"match":"news"}
 }
}
```

Dotazy s omezeným rozsahem

Pro vyhledávání v polích s numerickými hodnotami se využívají znaky „{}, [], <, >, =“ a operátor TO. Operátor TO se pojí s použitím obou typů závorek. Například dotaz *TOTAL_SCORE:[7 TO 8}* vrátí záznamy, jejichž pole TOTAL_SCORE obsahuje hodnoty 7 až 8, včetně hraniční hodnoty 7. Z tohoto příkladu je možné vypozorovat, že hranaté závorky zahrnují hraniční hodnoty a složené závorky naopak tyto hodnoty nezahrnují.
Tento typ dotazů je ovšem možné též použít na pole, která obsahují hodnoty typu text popřípadě „string“. V tomto případě je nutné si uvědomit, že hodnoty jsou řazeny podle jejich ASCII hodnoty, takže nejmenší hodnotu má znak „A“ a naopak největší hodnotu znak „a“. K vyhledávání už poté lze využít jen operátory „<“ (menší) nebo „>“ (větší).
Problém nastává při použití těchto dotazů na neanalyzované pole typu „text“, protože Elasticsearch defaultně převede zadaný dotaz na malá písmena. Tento problém lze vyřešit použitím dotazů ve formátu JSON a nastavením parametru *lowercase_expanded_terms* na hodnotu *false*.

3.3 Graph

Jedním z dostupných rozšíření aplikace Kibana je Graph, které bylo představeno v rámci ElastiCON 2016 jakožto součást připravovaného balíku rozšíření X-Pack. Graph lze rozdělit na dvě základní části a to na rozšíření možností nástroje Elasticsearch, které umožňuje uživatelům vyhledat spo-

jitosti mezi jednotlivými zaindexovanými položkami, a také jako rozšíření Kibany, kdy Graph poskytuje uživatelům vizualizaci, ze které jsou snadno rozpoznatelné váhy jednotlivých spojení.

3.3.1 Ovládání

Základním požadavkem pro vytvoření vizualizace pomocí rozšíření Graph je zvolení správného indexu, který obsahuje pole s hodnotami, které chceme prozkoumat a vizualizovat. V případě této bakalářské práce se jedná o index „hospitality“. Bez zvolení indexu není možné v tvorbě grafu pokračovat. Po zvolení indexu se musí zvolit zdroj dat pro jednotlivé vektory. Je nutné, aby tento zdroj obsahoval pouze textové řetězce nebo celá čísla a zároveň musí mít tento zdroj nastaven atribut „aggregatable“, protože komunikace nástroje Graph s aplikací Elasticsearch probíhá skrze automaticky tvořené dotazy, které obsahují atribut „aggs“. Po zvolení zdrojového pole je možné si zvolit barvu, kterou budou mít ve výsledném grafu vektory, ikonu výsledných vektorů a také počet vektorů, které se zobrazí. Pokud z výsledku chceme vynechat vybrané pole, je možné ho buďto odstranit skrze tlačítko „Remove“, nebo ho lze vynechat z dotazu tak, že podržíme klávesu „Shift“ a následně na něj klikneme. Posledním povinným polem je pole vyhledávací kam je možné vložit buďto text, kdy se provede fulltextové vyhledávání napříč všemi poli, které vybraný index obsahuje, nebo lze použít dotaz ve formátu Lucene Query Syntax a prohledat tak jen vybrané pole. Text, který je vložen do tohoto pole bude klíčový při následné tvorbě grafu, protože určuje spojení, které očekáváme mezi jednotlivými záznamy v databázi. Jak aplikace Graph interně funguje lze zjistit z následující části.

Interní komunikace

Výstupem tohoto rozšíření je pouze síť položek daného indexu, které mají stejné definované vlastnosti. Zobrazené výsledné položky se v Elasticsearch nazývají vektory a vztahy mezi nimi jsou znázorněny spojeními. Toto názvosloví ovšem nekoreluje s teorií grafů, kdy by měly být vektory správně nazývány vrcholy a spojení by měly být jednotlivé hrany mezi vrcholy grafu. Toto názvosloví bylo zvoleno z důvodu, že v Elasticsearch se již termín „vrcholy“ používá jako název pro součást topologie, která reprezentuje instanci Elasticsearch.
Jelikož je rozšíření Graph v aplikaci Kibana jen front-end aplikací, jsou jednotlivé operace uživatelů automaticky přetransformovány do patřičné podoby query dotazu a následně je odeslán požadavek aplikaci Elasticsearch,

která požadavek zpracuje a jako odpověď vrátí pole vektorů obsahující pole s názvem vektoru, což jsou vlastně hodnoty ze zadaného pole, které vyhovují vstupnímu dotazu. Následně také vrací pole obsahující informace o spojeních mezi jednotlivými vektory. Síla těchto spojení je reprezentována váhou spojení, která vyjadřuje podíl mezi počtem záznamů databáze, které obsahují oba vektory a mezi počtem záznamů, které obsahují alespoň jeden z vektorů. Váha tudíž může nabývat hodnot 0 až 1, kdy 1 znamená, že oba vektory jsou obsaženy ve všech odpovídajících záznamech.

Přidání spojení

Pokud při tvoření sítě potřebujeme zadat do vyhledávacího pole více výrazů, je nutné tyto kroky od sebe oddělit, což ale samozřejmě také znamená, že přijdeme o automatické vytvoření spojení mezi jednotlivými vektory. Ve výsledku tak získáme několik oddělených skupin vektorů, přičemž každá skupina je vnitřně propojena podle výrazu, který byl zadán při jejím vzniku. Pokud ovšem chceme vidět i spojení mezi jednotlivými skupinami, je již nutné využít možnost přidání spojení mezi existující skupiny, což nástroj Graph v aplikaci Kibana nabízí pod ikonou dvou spojených řetězů. Po stisku tohoto symbolu se výsledný graf sám aktualizuje a podle nových ohodnocení spojení vytvoří komplexnější vizualizaci.

Práce s vektory

Základním úkonem při práci s vektory je jejich selekce, která je umožněna buďto tlačítky „all“, „none“, „invert“ a „linked“, nebo lze výběr provádět ručně za pomoci klávesnice „Shift“ a myši.
Rozšíření Graph umožňuje uživatelům provádět takzvaný „spidering“, což je operace, při které uživatelé rozšiřují jen vybranou část původního vygenerovaného grafu. K této operaci jsou potřeba dva mezikroky. Tím prvním je označit si skupinu vektorů, pro které si přejeme provést „spidering“ a následně vyloučit z hledání vektory, které jsou již v původním grafu zobrazené. Toto lze provést přidáním klauzule „expand“ do dotazu pro Elasticsearch.[24] Tyto kroky lze také vykonat v rozšíření pro aplikaci Kibana. V tomto případě je pouze nutné zvolit vektory původního grafu, které si přejeme rozšířit, což je ekvivalent ke klauzuli „expand“. Dále se musí upravit vstupní zdrojová pole a následně jen stisknout ikonu s plusem, kdy se vytvoří nový dotaz a výsledkem je obnovený a rozšířený graf. Dále je možné jednotlivé vektory sdružovat do skupin. Nejprve je nutné si vybrané vektory označit a následně je pomocí tlačítka „group“ sloučit. Inverzní operaci lze provést kliknutím na tlačítko „ungroup“. Sjednocení vektorů je vhodné, pokud je použito neana-

lyzované zdrojové pole, které je citlivé na velikost písmen. Vektory lze také vyloučit z budoucího hledání jejich přidáním na černou listinu. K tomu slouží tlačítko „Blacklist selection from return to workspace“. Seznam blokovaných vektorů lze kdykoliv najít v nastavení aplikace Graph.

3.3.2 Shrnutí

Tento nástroj je vhodný především pro analýzu dat, kde lze předem očekávat spojitosti mezi daty. Především ve spojení s aplikací Logstah se jedná o velmi účinný nástroj například k analýze chování uživatelů, popřípadě na hledání vzorů chování útočníků. Další možnost použití vidím v analyzování dat ze sociálních sítí, kdy lze zjistit například jak moc se jednotlivé příspěvky šíří a jak jsou populární. Obecně je tento nástroj užitečný v oblasti bezpečnosti jako detekce hrozeb a také v oblasti komerce jako personalizované návrhy produktů. Nevýhodou je, že je dostupný jen jako součást balíku Elastic Stack, který je ovšem placený.

3.4 Vizualizace pomocí nástroje Kibana

Součástí zadání mé bakalářské práce je i seznam otázek, který zaslala Zayedova univerzita v Dubaji. Jak název této práce napovídá, jedná se o jedno z hlavních témat. V předchozích částech této práce byl popsán nástroj Kibana a jeho funkcionality, které využívám při získávání odpovědí na zadané otázky a k jejich případné vizualizaci.
Problém, na který jsem během práce narazil, jsou především sebraná data z ubytovacích portálů, která nejsou úplná, a tudíž nelze na některé zadané otázky odpovědět, tak aby byla odpověď relevantní pro všechna data.

3.4.1 Využití rozšíření Graph

Toto rozšíření je popsáno výše a je velmi vhodné především pro analýzu chování jednotlivých uživatelů portálu. Výhodou vizualizací z nástroje Graph je především hledání spojení mezi zadanými výrazy ve všech dokumentech. Následně probíhá interní ohodnocení výsledků a podle zadaných kritérií se zobrazí jen omezený počet výsledků.

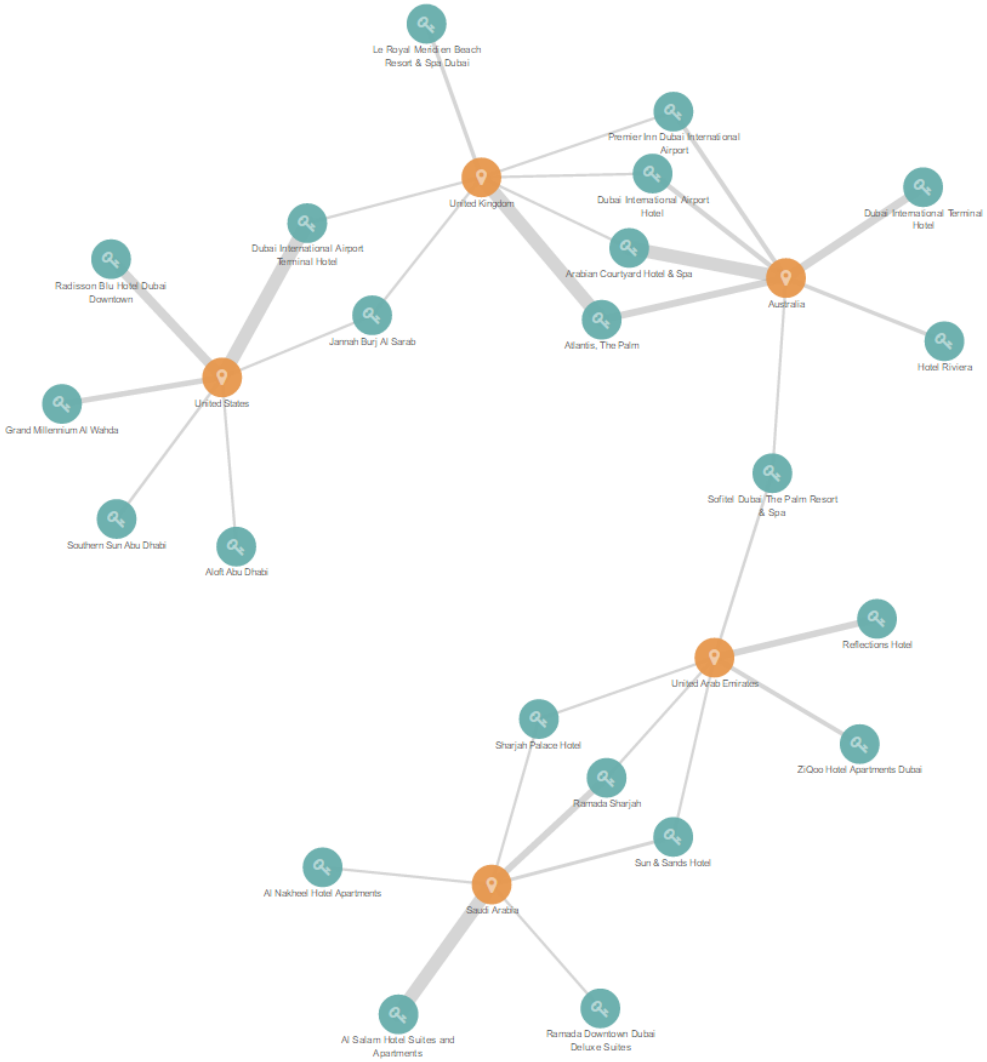
Jaké národnosti nejčastěji navštěvují Dubaj, a jaké jsou jejich oblíbené hotely?

Tato otázka je pro vhodná pro zpracování pomocí rozšíření Graph v aplikaci Kibana. Jedná se samozřejmě o hledání vztahů mezi národnostmi, které nejčastěji navštěvují Dubaj a mezi hotely, které navštívili. To že nejčastěji navštěvují Dubaj se dá chápat jako, že zanechali nejvíce recenzí na ubytovacích portálech. Tento krok ovšem samostatné rozšíření nedokáže realizovat, takže je nutné otázku rozdělit na dvě samostatné otázky. Prvním úkolem je získat odpověď na otázku, které národnosti se vyskytují v datech nejčastěji. K tomu výborně poslouží vizualizace, která se nazývá „Data table“, která je dostupná pod složkou „Visualize“. Tato vizualizace je vlastně přehledná tabulka, která poskytuje detailní přehled numerických výsledků, což přesně potřebujeme. Po otevření vizualizace je nutné nastavit parametry „metrics“ a „buckets“. Do prvního jmenovaného parametru jsem zadal agregaci typu počet, protože potřebuji zjistit počet recenzní, ve kterých se jednotlivé národnosti, respektive země objevují. Jako druhý parametr jsem zvolil možnost „Split Rows“, protože potřebuji mít data umístěna v řádcích pro lepší přehlednost. Konkrétně tedy bude v jednom řádku název země, ze které recenze pochází a vedle něj bude počet recenzí, které jsou dostupné v datech. Do pole „aggregation“ jsem vyplnil agregaci podle výrazů, která je ukryta pod názvem „Terms“ a jako zdrojové pole jsem zvolil pole USER_LOCATION_COUNTRY.keyword, které obsahuje názvy zemí, ze kterých uživatel pochází a to ve formátu string. Řazení jsem zvolil sestupné, protože hledám pět nejčastějších zemí a řadí se to samozřejmě podle počtu recenzí. Výsledkem této vizualizace je tedy tabulka čítající pět nejčastěji se vyskytujících zemí a konkrétní počet recenzí. Tento výsledek je k dispozici na následujícím obrázku.

Země	Počet recenzí
United Kingdom	50,491
United Arab Emirates	41,912
Saudi Arabia	29,147
United States	23,938
Australia	15,052

Díky rozdělení otázky na dvě části můžu využít výsledky z první části v rozšíření Graph jako výraz, pomocí kterého hledá Graph spojitosti mezi zadanými oblastmi. Nejprve jsem samozřejmě zvolil index, kterým je index „hospitality“. Následně jelikož hledám hotely, do kterých jezdí uživatelé z

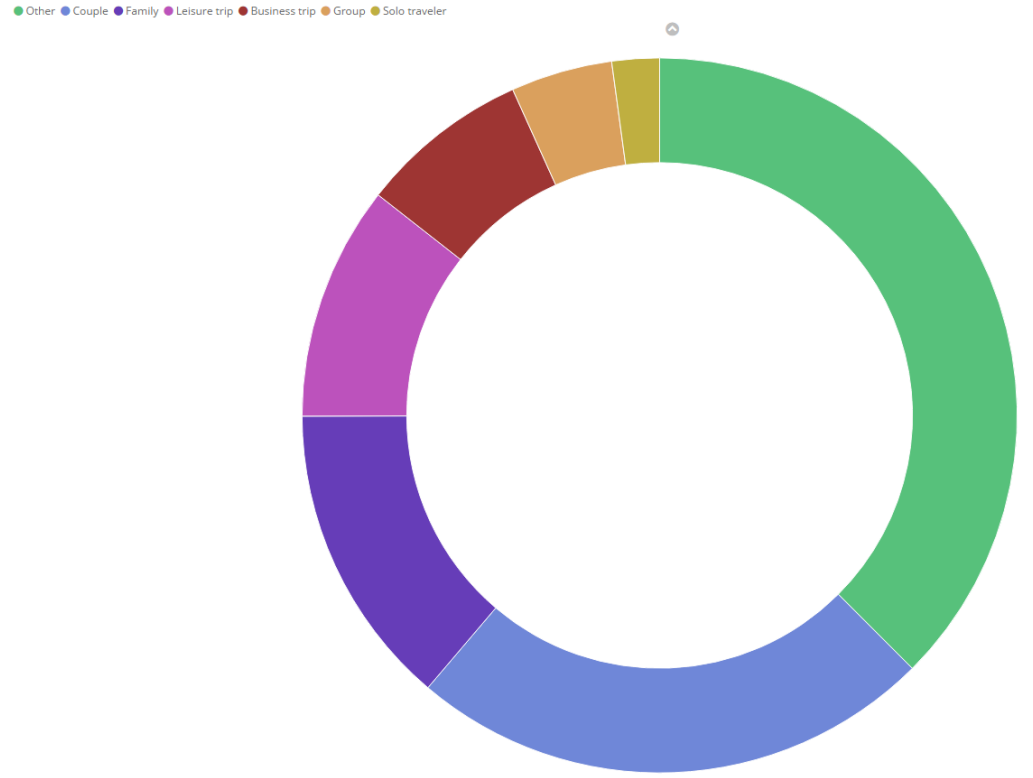
dané země nejčastěji, musím zvolit, že chci ve výsledcích zobrazit vždy jeden výsledek pro pole USER_LOCATION_COUNTRY.keyword a 5 výsledků pro pole HOTEL_NAME.keyword. Důležitý je především správný výběr polí, protože pole USER_LOCATION_COUNTRY a HOTEL_NAME nemají nastavený atribut „aggregateble“, takže je není možné využít v dotazech typu „aggs“, kterým ovšem rozšíření Graph komunikuje s databází Elasticsearch. Po zvolení zdrojových polí stačí jen zadat do vyhledávacího pole řetězec „USER_LOCATION_COUNTRY:„nazev““ a spustit vyhledávání spojení. Místo řetězce „nazev“ jsem postupně doplnil hodnoty, které jsem získal z výsledku první části otázky. Výsledkem tedy nakonec bylo pět částí, které se skládaly z jedné země a pěti hotelů, ale mezi těmito částmi nebylo žádné spojení. Na závěr tvorby vizualizace je tedy nutné použít takzvaný „spidering“, který doplní vazby mezi získanými částmi grafu. Odpovědí na otázku je tedy následující graf, který obsahuje pět nejčastějších zemí, ze kterých jezdí návštěvníci a jejich nejoblíbenější hotely.



Z grafu lze vyčíst, že anglofonní země jakou jsou Velká Británie, Spojené státy americké nebo Austrálie preferují podobné hotely, což lze vidět především u hotelů, které mají spojení mezi Austrálií a Velkou Británií. Naopak země Arabského poloostrova, které jsou v grafu reprezentovány Saudskou Arábií a Spojenými arabskými emiráty preferují jiné hotely než anglofonní země. Zároveň však uživatelé z těchto zemí mají v oblibě podobné hotely, což značí 3 hotely, které mají spojení jak se Saudskou Arábií, tak se Spojenými arabskými emiráty. Dle mého názoru se dal takovýto výsledek na otázku očekávat a to především kvůli velikosti jednotlivých zemí, které jsou v grafu zastoupeny. Překvapivým je pro mne pouze zastoupení Austrálie, ale může to být především způsobeno geografickou polohou Dubaje, přes který míří velké množství letů z Evropy do Austrálie a opačně. To může způsobit, že většina cestujících z Austrálie stráví část dovolené právě v Dubaji, kdy čeká na navazující let do cílové destinace. To potvrzují i následující tabulka

a graf, který reprezentuje zastoupení jednotlivých typů cest na celkovém počtu recenzí. Nejvýraznější zastoupení má typ cesty „other“. Obě vizualizace vznikly nástrojem Kibana a jsou uložené v indexu „hospitality“ pod názvy Australia travel type respektive pod názvem Australia travel type table.

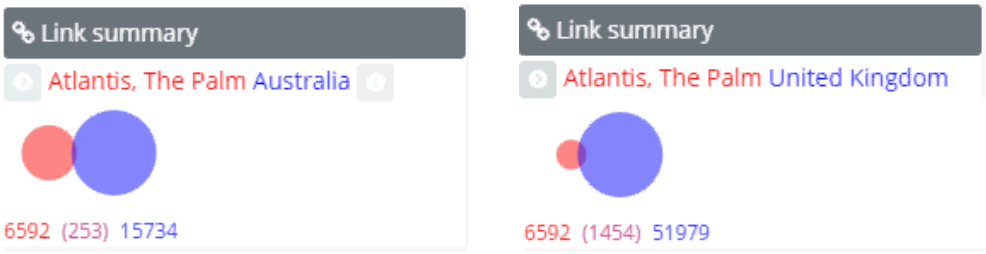
Typ cesty	Počet recenzí
Other	3,340
Couple	2,111
Family	1,224
Leisure trip	945
Business trip	684
Group	409
Solo traveler	191



Silné zastoupení zemí z Arabského poloostrova, které se umístili na druhém a třetím místě, co se do celkového počtu recenzí týče je zapříčiněno geografickou polohou, jelikož se jedná o velmi blízké sousedy Dubaje, v případě Spojeným arabských emirátů se jedná dokonce o hlavní město stejnojmenného emirátu. Velmi zajímavé je na výsledném grafu sledovat sílu jednotlivých spojení, která je symbolizována tloušťkou čáry mezi jednotlivými vrcholy grafu. Jelikož je výsledný graf fakticky složený z více výsledků hledání spojení je irelevantní porovnávat sílu spojení mezi hotely a různými zeměmi. To je možné zpozorovat pří porovnání šířky spojení mezi hotelem Atlantis, The

Palm a zeměmi Velká Británie a Austrálie. Detail tohoto spojení lze vidět na následujícím obrázku.

Na první pohled by se mohlo zdát, že do téhož hotelu jezdí zhruba dvojnásobné množství hostů z Velké Británie než hostů z Austrálie, ale ve skutečnosti je poměr větší. Hlavní příčinou je „spidering“, který sice přidá spojení mezi existujícími vrcholy, ale nebere v potaz již existující spojení, které do těchto vrcholů vedou. Další příčinou je samostatný princip rozšíření Graph, který určuje sílu spojení mezi vrcholy ne podle poměru počtu dokumentů, kde se vyskytují oba výrazy ku celkovému počtu dokumentů, ale závisí jen na poměru počtu dokumentů, kde se vyskytují oba výrazy ku počtu dokumentů, kde se vyskytuje alespoň jeden z výrazů. Jedná se tedy vlastně o průnik množiny dokumentů, které obsahují výraz A a množiny dokumentů, které obsahují výraz B. Abychom tedy zjistili reálnou sílu spojení, je nutné na kliknout na jednotlivá spojení a získat detailnější informace. Tyto informace jsou zjistitelné z následujících obrázků.



Červená čísla na obrázcích reprezentují počet dokumentů, které obsahují v poli HOTEL_NAME.keyword hodnotu Atlantis, The Palm, modrá

čísla reprezentují na horním obrázku počet recenzí, které obsahují v poli USER_LOCATION_COUNTRY.keyword hodnotu Australia a na spodním obrázku recenze obsahující v témže poli hodnotu United Kingdom. Růžová čísla reprezentují průnik množin, a jak je možné vidět, v tentýž hotel je přibližně sedm krát častěji recenzován uživateli z Velké Británie než uživateli z Austrálie, což je oproti původnímu odhadu velký rozdíl.

4 Slovník pojmů

BI: „Business intelligence, nebo také BI, je rámcový termín, označující paletu softwarových aplikací využívaných k analýze raw dat organizace.“ [3]

Big Data: „je termín aplikovatelný na soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými nástroji v rozumném čase.“ [5]

CRM neboli Customer Relationship Management je zkratka pro podpůrné informační systémy, které jsou určené pro řízení vztahů se zákazníky.

CSV neboli Comma-separated values je souborový formát určený pro výměnu tabulkových dat. Soubor se skládá z řádků, na kterých jsou uložené položky, které jsou od sebe odděleny oddělovači (například čárka, středník, tabulátor atd.).

JSON „neboli JavaScript Object Notation je formát souborů určený k výměně dat.“ Je snadno čitelný i zapisovatelný člověkem a zároveň lze soubory ve formátu JSON snadno generovat i zpracovávat strojově.[10]

NoSQL„Jedná se o novou generaci databázových systému pro správu velkého množství dat, které jsou převážně ne-relační, distribuované, škálovatelné a podporují replikaci. Často jsou to databáze bez datového schématu, s jednoduchým rozhraním pro práci s daty a open source přístupem“.[7]

Materializovaný pohled: jsou to databázové objekty obsahující výsledek dotazu. Přístup k výsledku je rychlejší než u normálních query dotazů, ale nesmí se měnit vstupní data pro materializovaný pohled, protože se neaktualizuje automaticky při změně databáze.

PDF „PDF neboli Portable Document Format je formát používaný k prezentaci a spolehlivé výměně dokumentů, který je nezávislý na softwaru, hardwaru i operačnímu systému.“ 16

SOAP: Simple Object Access Protocol je protokol zajišťující přenos zpráv založených na XML přes síť, především pak pomocí protokolu HTTP.

XML: neboli Extensible Markup Language je obecný značkovací jazyk určen pro uchování a přenos dat. Je čitelný jak pro lidi tak i pro stroje.

Literatura

[1] *Company Background* [online]. InetSoft Technology Corp, 2002. [cit. 2016/12/10]. Dostupné z: http://web.archive.org/web/20020210121546/www.inetsoft.com/cgi-bin/v6/index.pl?aboutus/backgrounder.html.

[2] Brasetvik, A. *Elasticsearch as a NoSQL Database* [online]. Elasticsearch, 2013. [cit. 2016/12/11]. Dostupné z: https://www.elastic.co/blog/found-elasticsearch-as-nosql#schema-flexible.

[3] *Co je to business intelligence* [online]. CFO world, 1999. [cit. 2016/12/10]. Dostupné z: http://businessworld.cz/ostatni/co-je-to-business-intelligence-7157.

[4] *Tableau Desktop* [online]. ZiffDavis, LLC. PCMag Digital Group., 2016. [cit. 2016/12/11]. Dostupné z: http://www.pcmag.com/business/directory/business-intelligence/182-tableau-desktop.

[5] Dolák, O. *Big data* [online]. SystemOnLine.cz, 2011. [cit. 2016/12/10]. Dostupné z: https://www.systemonline.cz/clanky/big-data.htm.

[6] *Introduction to ElastiCubes* [online]. Sisense Inc., 2014. [cit. 2016/12/11]. Dostupné z: https://www.sisense.com/documentation/v5/elasticube-manager/introduction-to-elasticube-manager/.

[7] Holubová, I. et al. *Big Data a NoSQL databáze*. Grada, 2015. ISBN 978-80-247-5466-6.

[8] *Download Splunk Enterprise* [online]. Splunk Inc., 2015-2016. [cit. 2016/12/10]. Dostupné z: https://www.splunk.com/en_us/download/sem.html?ac=ga0508_s_splunk&_kk=download%2520splunk&_kt=c02e72c3-e19a-49ad-b070-f6acd33faee3&gclid=CNK84bG06dACFUE_Gwodux8AlQ.

[9] *How To Get a 20 Million Dollar Pre-Money Valuation for Series A: Tableau Software CEO Christian Chabot* [online]. Sramana Mitra's Biography, 2010. [cit. 2016/12/11]. Dostupné z: http://www.sramanamitra.com/2010/03/04/how-to-get-a-20-million-pre-money-valuation-for-series-a-tableau-software-ceo-ch

[10] *JSON* [online]. Refsnes Data, 2017. [cit. 2017/03/13]. Dostupné z: https://www.w3schools.com/js/js_json_intro.asp.

[11] *Style Intelligence - Business Intelligence Software* [online]. InetSoft Technology Corp, 2016. [cit. 2016/12/10]. Dostupné z: https://www.inetsoft.com/products/StyleIntelligence/.

[12] *Sisense REVIEW* [online]. FinancesOnline.com, 2016. [cit. 2016/12/11]. Dostupné z: https://reviews.financesonline.com/p/sisense/.

[13] Levy, E. *SISENSE 5.7 IS OUT: SAY HELLO TO MACHINE DATA, MONGODB* [online]. Sisense Inc., 2015. [cit. 2016/12/11]. Dostupné z: https://www.sisense.com/blog/sisense-5-7-is-out-say-hello-to-machine-data-mongodb/.

[14] *Splunk* [online]. ZiffDavis, LLC. PCMag Digital Group., 2016. [cit. 2016/12/10]. Dostupné z: http://www.pcmag.com/business/directory/network-monitoring/1830-splunk.

[15] *VizQL™* [online]. TABLEAU SOFTWARE, 2016. [cit. 2016/12/11]. Dostupné z: http://www.tableau.com/about/mission#vizql.

[16] *PDF* [online]. Adobe Systems Incorporated., 2017. [cit. 2017/03/13]. Dostupné z: https://acrobat.adobe.com/cz/cs/why-adobe/about-adobe-pdf.html.

[17] *Style Intelligence - Business Intelligence Software* [online]. InetSoft Technology Corp, 2016. [cit. 2016/12/10]. Dostupné z: https://www.inetsoft.com/company/bi_dashboard_pricing/.

[18] *Pricing* [online]. TABLEAU SOFTWARE, 2016. [cit. 2016/12/11]. Dostupné z: http://www.tableau.com/products/desktop#pricing-specs.

[19] *Machine Learning Toolkit* [online]. Splunk Inc., 2016. [cit. 2016/12/10]. Dostupné z: https://splunkbase.splunk.com/app/2890/#/overview.

[20] *Elasticsearch Reference [5.2]* [online]. Elasticsearch, 2013. [cit. 2017/02/13]. Dostupné z: https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html.

[21] *Revision history* [online]. TABLEAU SOFTWARE, 2016. [cit. 2016/12/11]. Dostupné z: http://www.tableau.com/new-features/10.0#.

[22] Roes, T. *Kibana 5 Introduction* [online]. https://www.timroes.de, 2016. [cit. 2016/12/13]. Dostupné z: https://www.timroes.de/2016/10/23/kibana5-introduction/.

[23] Roes, T. *Elasticsearch/Kibana Queries - In Depth Tutorial* [online]. https://www.timroes.de, 2016. [cit. 2016/12/13]. Dostupné z: https://www.timroes.de/2016/05/29/elasticsearch-kibana-queries-in-depth-tutorial/.

[24] Roes, T. *Elasticsearch Graph [2.4]* [online]. Elasticsearch, 2017. [cit. 2017/03/13]. Dostupné z: https://www.elastic.co/guide/en/graph/current/graph-api-rest.html#spider-search.

[25] *SISENSE ADD-ONS* [online]. Sisense Inc., 2016. [cit. 2016/12/11]. Dostupné z: https://www.sisense.com/product/add-ons/.

[26] *Connect to more data* [online]. TABLEAU SOFTWARE, 2016. [cit. 2016/12/11]. Dostupné z: http://www.tableau.com/products/desktop#data-sources-professional.

[27] *Tiobe index for December 2016* [online]. TIOBE software BV, 2016. [cit. 2016/12/10]. Dostupné z: http://www.tiobe.com/tiobe-index/.

[28] *Booz Allen Hamilton and Splunk Announce Strategic Alliance to Deliver Predictive Security Analytics and Operationalize Threat Intelligence* [online]. Splunk Inc., 2015. [cit. 2016/12/10]. Dostupné z: http://www.splunk.com/view/booz-allen-hamilton-and-splunk-announce-strategic-alliance-to-deliver-predictive-SP-CAAAPB7.

[29] *Sisense* [online]. ZiffDavis, LLC. PCMag Digital Group., 2016. [cit. 2016/12/11]. Dostupné z: http://www.pcmag.com/business/directory/data-visualization/1225-sisense.