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

Projekt 5

Analýza a vizualizace dat z ubytovacích portálů

Plzeň 2016 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 11. prosince 2016

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 Nástroje na analýzu a vizualizaci dat

První kapitola práce je zaměřena na nástroje, pomocí kterých je možné analyzaovat 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.

1.1 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 jsou některé zdarma (jako např. Vizualize Free a Style Scope Free Edition) Při vývoji se především používá XML, SOAP, jazyk Java a JavaScript[1]. Hlavně díky použití jazyku Java, který je dnes celosvětovou jedničkou mezi programovacími jazyky [13] jsou aplikace snadno integrovatelné s jinými softwary, které jsou založeny na otevřených standardech.

1.1.1 Přehled

InetSoft Style Intelligence je velmi silný nástroj, který umožňuje zpracovávat a kombinovat data z různých zdrojů. Systém je nabízen jako cloudová aplikace nebo jako řešení na míru. Řešení na míru je například produkt InetSoft Style, který je implementován přímo na server subjektu, který požádal o řešení, a InetSoft garantuje minimální zatížení na chod serveru.

1.1.2 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 před připravených materializovaných pohledů.
- Zpracování dat

- Real-time interaktivní dashboardy
- Multidimenzionální grafy
- Geografické mapy
- Předpřipravené typy dashboardů pro:
  * Prodej a nákup
  * Zdravotnictví
  * Vzdělávání
  * Různé formy analýz

1.1.3 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 semistrukturovaná nebo nestrukturovaná data. InetSoft Style Intelligence podporuje export do běžných formátů, díky čemu jsou data a grafy využitelné v 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, HP-Unix, Solaris a další.[6]

1.1.4 Hodnocení a ceník

Hlavní výhodou tohoto produktu je uživatelsky přívětivé prostředí, kde uživatelé tvoří vizualizace svých dat. To je především díky přehlednosti webové stránky, ze které se nástroj ovládá, jednotlivým tlačítkům zlehčující ovladatelnost systému a také systémem drag and drop, bez nutnosti používat jakýkoliv dotazovací jazyk. Velkým pozitivem je i možnost ovládat nástroje přes mobilní telefon, ale podle mě je to vhodné jen k zobrazování již vytvořených analýz a dashboardů. Dalším pozitivem je i možnost nahlédnout přímo do raw dat pouhým kliknutím na libovolný prvek vizualizace. Výhodou taktéž je, že produkt není určen primárně pro technicky vzdělané uživatele, ale je určené pro běžné uživatele bez nutnosti podpory IT oddělení. Balíček služeb je dostupný již od 2 800 dolarů.[9]

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.[4] 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 (Firewally, Antiviry), 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 kyber útoků. Právě díky nástrojům vhodným ke zlepšení kyber bezpečnosti spolupracuje firma s americkou vládou skrze zprostředkovatelskou firmu, která zajišťuje kyberbezpečnost pro vládu.[14]

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 semistrukturovaných a nestrukturovaných dat z non-SQL databází.

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. 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, obsahuje více než 140 příkazů a podporuje korelaci přes pět oblastí (čas, transakce, join, lookup a subvyhledávání). 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 anebo kombinovaně.
- 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. [11]
  - Data se vizualizují pomocí různých uživatelsky přizpůsobitelných grafů, které se následně skládají do interaktivních dashboardů, které lze následně 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.

2.3 Kompatibilita

Splunk je zaměřen především na zpracování big dat, ale je možné ho připojit i na relační databáze, nebo ho propojit s Excelem či Tableau. 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. [4]

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.

3 Tableau Desktop

Jedná se o produkt americké softwarové společnosti Tableau Software, jenž se zaměřuje na softwary vhodné pro vizualizaci dat a na nástroje business intelligence. Firma vznikla na základě výzkumů v oblasti vizualizace dat ze Standfordské univerzity. Konkrétně se jednalo projekt Polaris z oddělení počítačových věd.[5] 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.

3.1 Přehled

Jedná se o 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 ostatními kolegy. Předností produktu Tableau Desktop je uživatelská přívětivost, protože k jeho běžnému nejsou nutné žádné pokročilé technologické znalosti, ale stačí používat systém tvorby vizualizací drag and drop.

3.2 Technologie

- Kolekce a příprava dat
  - Aby Tableau Desktop umožňovala používání dat z vícero 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 Excelu, MS Access, textových souborů, logů a podobně) a na servery, kdy se může Tableau Desktop přímo napojit na databázové servery a data se můžou 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.“[8] Výhodou je, že 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 jiného grafu.
  - Tableau nabízí mnoho analytických nástrojů a 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 dashboardu.

3.3 Kompatibilita

Tableau nabízí možnost využít základní aplikaci pro programovatelný přístup k datům a vizualizacím, ale jelikož je napsána v JavaScriptu je možné ji 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 OSX 10.10 a novější. Verzi personal je možné propojit s zdroji typu MS Access popřípadě přímo s textovými soubory ve formátu CSV popřípadě JSON. Verze professional je již propojitelné s většinou databázových systémů, které se dnes používají (např. Oracle Data-

bases, PostgresSQL, Cloudera Hadoop Hive and Impala, Cisco Information Server).[12]

3.4 Hodnocení a ceník

Cena se pohybuje od 999$ za osobní verzi, až po 1999$ za profesionální edici Tableau Desktop.[10] 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 jednotlivý zaměstnanci, tak i zákazníci. K dispozici je taktéž mobilní aplikace, která je primárně určená pro prohlížení vytvořených dashboardů.

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, které se v očích odborné veřejnosti tak i v očích uživatelů jeví jako velmi spolehlivé.[7]

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.

4.2 Technologie

- Kolekce a příprava dat
  - Sisense nevyžaduje časově nároční předzpracování dat, tak jako ostatní BI nástroje na trhu. Data z různých zdrojů jsou sjednocena a importována do jednotného repositáře, 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.

  - 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 jednotlivá 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 můžou sdílet napříč organizací a sledovat jeho vývoj a případně upravit uživatelská práva. Každý dashboard obsahuje ovládací panely s filtry a lze zde přímo upravovat query dotazy a přizpůsobovat tak vizualizace dle aktuálních potřeb. Aplikace již obsahuje některé dashboardy předpřipravené, což je určeno především pro nezkušené analytiky.

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. Software je založen na webových technologiích a je tak kompatibilní s běžnými operačními systémy jako jsou Windows, Android, 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.

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 avšak demo verze, která obsahuje již data a je tak možné si vyzkoušet funkcionalitu systému před jeho zakoupením. Výhodou je napojení systému na mobilní telefony, kdy výrobce deklaruje, a zákazníci potvrzují, že především sledování vývoje dashboardů přes mobilní zařízení je velmi praktické. 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.

5 Kibana

5.1 Přehled

5.2 Technologie

- Kolekce a příprava dat
  - –
  - –
- Zpracování dat
  - –
  - –

5.3 Kompatibilita

5.4 Hodnocení a ceník

6 Slovník pojmů

BI: „Business intelligence, nebo také BI, je rámcový termín, označující paletu softwarových aplikací využívaných k analýze syrových dat organizace.“ [2]
SOAP: Simple Object Access Protocol je protokol zajišťující přenos zpráv založených na XML pomocí protokolu HTTP.
Materializovaný pohled: jsou to databázové objekty obsahující výsledek dotazu. Přstup 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.
Big Data : „big data je termí aplikovatelný na soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými nástroji v rozumném čase.“ [3]

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] *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.

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

[4] *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.

[5] *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

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

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

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

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

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

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

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

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

[14] *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`.