



Sofistikované spracovanie dát

Dokumentácia k inžinierskemu dielu

Vedúci tímu: Ing. Michal Holub

Členovia tímu: Bc. Igor Daniš, Bc. Jakub Kmeťko, Bc. Martin Košut, Bc. Martin Lošák,
Bc. Stanislav Paľove, Bc. Alex Ostrovský, Bc. Peter uherek

Akademický rok: 2014/2015

OBSAH

1	ÚVOD.....	5
2	GLOBÁLNE CIELE PROJEKTU NA ZIMNÝ SEMESTER	6
2.1	SOFISTIKOVANÉ SŤAHOVANIE DATASETOV	6
2.2	PRVOTNÁ ANALÝZA DÁT.....	6
3	NEČAKANÁ SPOLOČNOSŤ	7
3.1	PRÍPADY POUŽITIA.....	8
3.2	DÁTOVÝ MODEL.....	9
3.3	KOSTRA GUI.....	10
3.3.1	ANALÝZA	10
3.3.2	NÁVRH	10
3.3.3	MODEL.....	11
3.3.4	IMPLEMENTÁCIA.....	12
3.3.5	TECHNOLÓGIA.....	12
3.3.6	DETAILY PODPORY	12
3.4	REGISTRÁCIA	13
3.4.1	ŠPECIFIKÁCIA	13
3.4.2	ANALÝZA	13
3.4.3	IMPLEMENTÁCIA.....	13
3.4.4	TESTOVANIE	13
3.5	PRIHLÁSENIE	13
3.5.1	ŠPECIFIKÁCIA	13
3.5.2	ANALÝZA	14
3.5.3	IMPLEMENTÁCIA.....	14
3.5.4	TESTOVANIE	14
3.6	ODHLÁSENIE.....	14
3.6.1	ŠPECIFIKÁCIA	14
3.6.2	ANALÝZA	14
3.6.3	IMPLEMENTÁCIA.....	15
3.6.4	TESTOVANIE	15
3.7	SPRÁVA PROFILU	15
3.7.1	ANALÝZA	15
3.7.2	NÁVRH	15
3.7.3	IMPLEMENTÁCIA.....	15
3.7.4	TESTOVANIE	15
3.8	VYTVORENIE OBRAZU DATASETU	16
3.8.1	ANALÝZA	16
3.8.2	IMPLEMENTÁCIA.....	16
3.8.3	TESTOVANIE	16
3.9	ZOBRAZENIE, MAZANIE A EDITÁCIA DATASETU	17
3.9.1	ANALÝZA	17
3.9.2	IMPLEMENTÁCIA.....	17
3.9.3	TESTOVANIE	17

CEZ VRCHY A POD VRCHMI	18
3.10 ZÍSKAVANIE DÁT	19
3.10.1 PLÁNOVAČ SŤAHOVANIA	19
3.10.2 SPÔSOBY IMPLEMENTÁCIE	19
3.10.3 POUŽITIE KOMPLETNÉHO PLÁNOVACIEHO SOFTVÉROVÉHO RÁMCA.....	20
3.10.4 ANALÝZA VYBRANÝCH PLÁNOVACÍCH GEMOV	20
3.10.5 MOŽNOSTI ĎALŠIEHO ROZŠÍRENIA VYBRANÝCH PLÁNOVACÍCH NÁSTROJOV	21
3.10.6 NAVRHOVANÉ RIEŠENIE	21
3.10.7 REFERENCIE:.....	22
3.10.8 ANALÝZA IMPLEMENTÁCIE SŤAHOVANIA	23
3.10.9 MOTIVÁCIA K SŤAHOVANIU	23
3.10.10 PROBLÉMY KTORÉ MÔŽU VZNIKNUŤ POČAS SŤAHOVANIA	23
3.10.11 SPÔSOBY IMPLEMENTÁCIE	23
3.10.12 NAVRHOVANÉ RIEŠENIE	24
3.10.13 REFERENCIE.....	25
3.11 UKLADANIE DATASETOV A ELASTICSEARCH.....	25
3.11.2 NÁVRHY ARCHITEKTÚR UKLADANIA DÁT DO DATABÁZ	27
3.11.3 ZDROJE	28
3.12 TYPY DATASETOV	28
3.12.1 ANALÝZA TYPOV DATASETOV.....	28
3.12.2 ÚLOHY DO ĎALŠIEHO ŠPRINTU VYPLÝVAJÚCE Z TEJTO ANALÝZY	29
3.12.3 DATASETY	29
3.13 SPRACOVANIE DATASETOV	31
3.13.1 ANALÝZA DATASETOV.....	32
3.13.2 STROJOVÉ UČENIE V RUBY	33
3.13.3 PREPOJENIE R A RUBY	33
3.14 APLIKÁCIE 3. STRÁN.....	34
3.14.1 AUTENTIFIKÁCIA	34
3.14.2 GEOLOGICKÉ DÁTA	35
3.14.3 GOOGLE MAPS.....	35
3.14.4 WOLFRAMALPHA.....	35
3.14.5 VYHĽADÁVANIE ĽUDÍ A FIRIEM	35
3.14.6 PIPL	35
3.14.7 FINSTAT	36
3.14.8 CAPTCHA	36
3.15 VYKRESĽOVANIE DÁT.....	36
3.15.1 D3JS	36
3.15.2 HIGH CHARTS	37
3.15.3 JQUERY SPARKLINES	38
3.15.4 GOOGLE CHARTS	38
3.15.5 FLOT.....	39
3.15.6 RAPHAËL JS.....	39
3.15.7 GAUGE.....	39
3.15.8 VÝSLEDNÉ HODNOTENIE	41
3.16 OBRAZOVKY GUI.....	42
3.17 PRISPÔSOBENIE GUI POUŽÍVATEĽOM	44
3.17.1 AKCIE, KTORÉ MÔŽU POUŽÍVATEĽ VYKONÁVAŤ:.....	44

3.17.2	MODEL.....	45
4	HÁDANKY V TME.....	47
4.1	PRÍPADY POUŽITIA.....	48
4.2	DÁTOVÝ MODEL.....	49
4.3	RECAPTCHA.....	50
4.3.1	ŠPECIFIKÁCIA.....	50
4.3.2	ANALÝZA.....	50
4.3.3	IMPLEMENTÁCIA.....	50
4.3.4	TESTOVANIE	50
4.4	EMAILOVÁ VERIFIKÁCIA PRI REGISTRÁCII	50
4.4.1	ŠPECIFIKÁCIA.....	50
4.4.2	ANALÝZA.....	50
4.4.3	IMPLEMENTÁCIA.....	50
4.4.4	TESTOVANIE	51
4.5	PASSWORD RESET	51
4.5.1	ŠPECIFIKÁCIA.....	51
4.5.2	ANALÝZA.....	51
4.5.3	IMPLEMENTÁCIA.....	51
4.5.4	TESTOVANIE	51
4.6	REFACTOR PROFILU.....	51
4.6.1	OPIS	51
4.6.2	ANALÝZA.....	51
4.6.3	IMPLEMENTÁCIA.....	51
4.6.4	TESTOVANIE	52
4.7	STIAHNUTIE DATASETU A PRIDANIE DO DB.....	52
4.7.1	VSTUP	52
4.7.2	VÝSTUP	52
4.7.3	ANALÝZA.....	52
4.7.4	NÁVRH	53
4.7.5	IMPLEMENTÁCIA.....	54
4.7.6	TESTOVANIE	55
4.8	CHCEM VIDIEŤ ZÁKLADNÉ TEXTOVÉ INFORMÁCIE (ATRIBÚTY, DÁTUM, VEĽKOSŤ)55	
4.8.1	ŠPECIFIKÁCIA.....	55
4.8.2	ANALÝZA.....	55
4.8.3	IMPLEMENTÁCIA.....	55
4.8.4	TESTOVANIE	55
4.9	POUŽÍVATEĽ MENÍ TYP ATRIBÚTU	55
4.9.1	ŠPECIFIKÁCIA.....	55
4.9.2	ANALÝZA.....	55
4.9.3	IMPLEMENTÁCIA.....	55
4.9.4	TESTOVANIE	55
4.10	AKO POUŽÍVATEĽ CHCEM VIDIEŤ PRVÝCH 15 RIADKOV DATASETU	56
4.10.1	ŠPECIFIKÁCIA.....	56
4.10.2	ANALÝZA.....	56
4.10.3	IMPLEMENTÁCIA	56
4.10.4	TESTOVANIE	56

1 ÚVOD

Analýza dát patrí v súčasnosti medzi základné postupy pri objavovaní znalostí a budovaní konkurencieschopnosti. V mnohých oblastiach je dnes aj napriek tomuto faktu štandardom intuitívne spracovanie dát v kancelárskych nástrojoch, ktoré neponúkajú žiaden rozšírenia na hĺbkovú analýzu. Výsledkom je, že vznikajú sety neznámych dát, alebo prípadne známych, no nepreskúmaných.

Webový nástroj DataPoints ponúka jednoduché a časovo nenáročné riešenie na spracovanie nie len týchto dát. Medzi naše devízy patrí najmä užívateľsky nenáročné prostredie a fakt, že nahrávanie a analýza dát beží na serveri, s tým, že počas tejto doby môže užívateľ nerušene pokračovať vo svojich aktivitách. Na konci spomínaného procesu je zaslaná informácia na užívateľsky e-mail.

Nástroj bude robustný, aby vedel spracovať rôznorodé datasety. Cieľom je pozbierať a integrovať najlepšie riešenia na spracovanie dát a doplniť ich tak, aby výsledný nástroj dokázal prezentovať zaujímavé zistenia, zobrazovať súvislosti, prepojí dáta s mapami a inými zdrojmi na webe.

2 GLOBÁLNE CIELE PROJEKTU NA ZIMNÝ SEMESTER

2.1 SOFISTIKOVANÉ SŤAHOVANIE DATASETOV

Prvým dôležitým míľnikom je zabezpečiť plánovanie sťahovania, na základe ktorého dostaneme dáta od užívateľa do databázy. Nad týmito dátami automaticky spustíme prvotnú analýzu.

2.2 PRVOTNÁ ANALÝZA DÁT

V rámci prvého kontaktu s dátami budeme vedieť určiť nie len ich databázový typ, ale aj niekoľko entít z reálneho sveta (osoba, adresa, email, dátum, čas, a pod.)

3 NEČAKANÁ SPOLOČNOSŤ

Číslo šprintu: 1

Začiatok šprintu: 9.10.2014

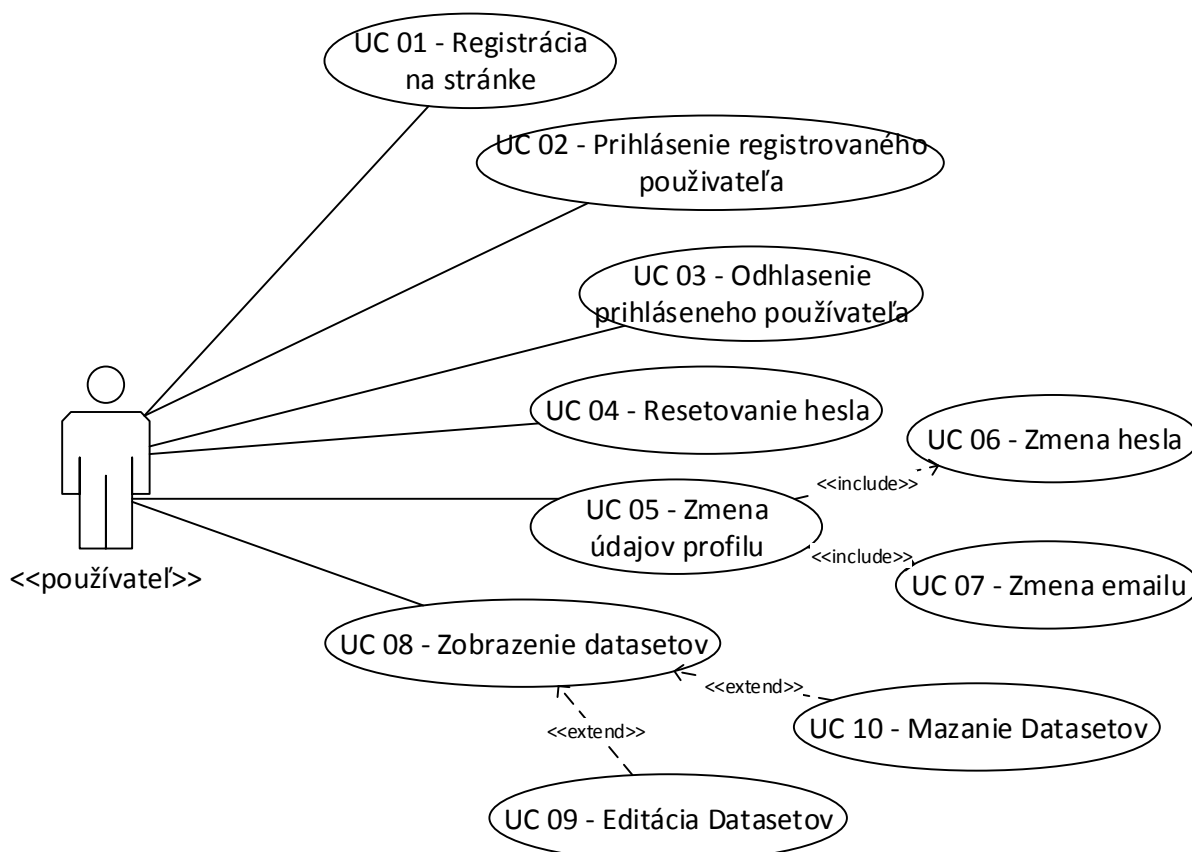
Koniec šprintu: 23.10.2014

Príbehy:

- Kostra GUI
- Registrácia
- Prihlásenie
- Odhlásenie
- Správa profilu
- Vytvorenie obrazu datasetu
- Vidieť uploadnuté datasety
- Získanie dát na server
- Zmazať/Upraviť datasety
- Vidieť výsledky analýz

3.1 PRÍPADY POUŽITIA

Na obrázku číslo 1 je uvedený diagram prípadov použitia. Obrázok sa skladá z 10 prípadov použitia, ktoré sme identifikovali v prvom šprinte. Opis jednotlivých prípadov použitia sa nachádza pod obrázkom.



Obrázok 1- Diagram prípadov použitia pre 1. šprint.

UC 01: Registrácia na stránke

Neregistrovaný používateľ bude mať možnosť vytvoriť si účet. S pomocou účtu sa bude prihlasovať do systému.

UC 02: Prihlásenie registrovaného používateľa

Používateľ, ktorý bude mať vytvorený účet na stránke bude schopný prihlásiť sa.

UC 03: Odhlásenie prihláseného používateľa

Prihlásený používateľ bude mať možnosť odhlásiť sa zo systému a tým ochrániť svoje údaje pred zneužitím.

UC 04: Resetovanie hesla

Používateľovi, ktorý zabudol svoje heslo bude umožnené jeho resetovanie s možnosťou nastavenia nového hesla.

UC 05: Zmena údajov profilu

Používateľ bude mať možnosť prezrieť si svoj profil a v ňom zmeniť jednotlivé údaje uvedené pri registrácii.

UC 06: Zmena hesla používateľa

Používateľ bude mať možnosť zmeny hesla, ktoré uviedol pri registrácii.

UC 07: Zmena emailu používateľa

Používateľ bude mať možnosť zmeny emailu, ktorý uviedol pri registrácii.

UC 08: Zobrazenie datasetov

Prihlásený používateľ bude mať možnosť zobraziť svoje nahrané datasety.

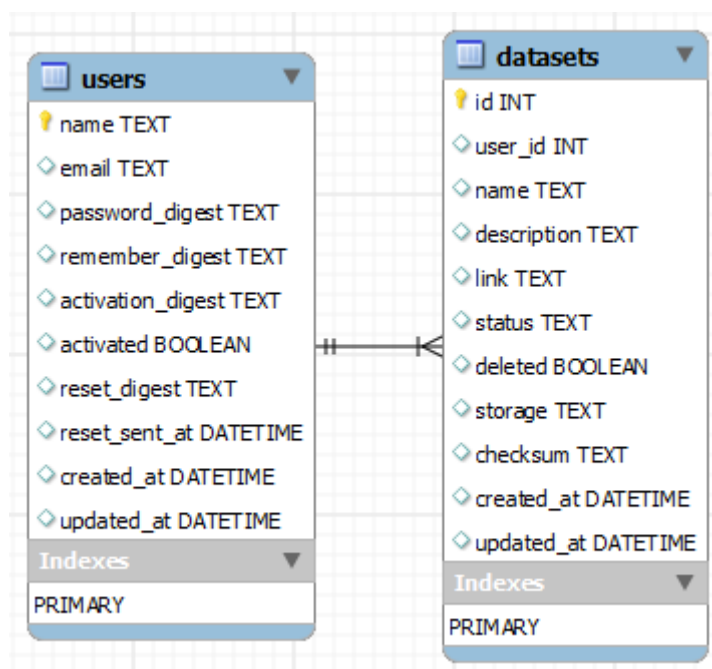
UC 09: Editácia datasetov

Pri zobrazení nahraných datasetov bude mať používateľ možnosť editovať ich popis a meno.

UC 10: Mazanie datasetov

Pri zobrazení nahraných datasetov bude možné vymazať zvolený dataset.

3.2 DÁTOVÝ MODEL



Obrázok 2 - Dátový model k 1. šprintu.

Dátový model opisuje dve tabuľky *users* a *datasets*. Vzťah medzi tabuľkami je nasledovný: jeden dataset môže patriť práve jedenému používateľovi a jeden používateľ môže mať veľa datasetov. Ako zo vzťahu vyplýva tabuľka *users* opisuje používateľov nášho systému a tabuľka *datasets* opisuje ich datasety.

Atribúty v tabuľke *users* sú samo opisné a nie je potrebný ich ďalší opis. Tabuľka *datasets* obsahuje cudzí kľúč na záznam v tabuľke *users*, meno daného datasetu, opis tohto datasetu, link odkiaľ bude dataset stiahnutý, status v ktorom sa práve dataset nachádza, flag *deleted*, ktorý vypovedá o tom, či bol daný dataset zmazaný alebo nie, hodnotu *storage*, ktorá uchováva cestu k súboru na danom serveri, hodnotu *checksum*, ktorá uchováva hash súboru.

3.3 KOSTRA GUI

Vytvoriť grafické rozhranie. Úvodná stránka, registrácia používateľov, profil používateľov, správa dokumentov.

3.3.1 ANALÝZA

Dizajn webovej stránky by mal súvisieť s celkovou identitou značky DataPoints, ktorá je zatiaľ zachytená najmä v logu (viz. Obrázok 1.4.2 – logo DataPoints). V rámci hlavnej kostry je potrebné navrhnuť dizajn a implementovať hlavičku stránky, úvodnú slideshow, stručný a zákazníčkovi jasný popis priebehu procesu, stručné vysvetlenie našej ponuky s odkazom na registráciu, a pätku stránky.

3.3.2 NÁVRH

Farebná schéma bude zodpovedať logu, a **Úvodná slideshow** bude zobrazovať:

- oranžová (#f46430)
- čierna (#272727)
- biela (#ffffff)

- obrázok hlavnej obrazovky v procese analýzy, do ktorej vstupujú z rôznych strán obrazovky s dátami, zobrazujúcimi veľké dáta, mapy, grafy, analýzy
- Veľký nápis Big Data so stručným textom a odkazom na detail procesu

Hlavička stránky bude obsahovať elementy v tomto poradí:

- logo DataPoints
- odkaz na úvodnú stránku
- odkaz na produkt / službu
- odkaz na demo verziu
- možnosť prihlásenia
- odkaz na registráciu

Opis priebehu procesu bude obsahovať elementy v tomto poradí:

- Nahrajte svoj dataset
- My ho spracujeme
-A kontaktujeme vás

Bude opisovať najmä jednoduchosť riešenia a veľkú devízu v tom, že počas procesu sťahovania datasetu a prvotnej analýzy nemusí užívateľ čakať, a my ho na konci procesu kontaktujeme

Stručné vysvetlenie našej ponuky bude obsahovať jasnú správu o tom, že užívateľom chceme pomôcť pri skúmaní ich známych a neznámych dát, a tým im sprístupniť nové vedomosti

Pätička stránky bude pre lepšiu navigáciu na stránke obsahovať odkazy z hlavičky stránky, logo STU FIIT, informácie o vývojovom tíme a kontakt

3.3.3 MODEL



Obrázok 3. Návrh webového dizajnu kostry stránky



Obrázok 4. Logo DataPoints

3.3.4 IMPLEMENTÁCIA

Dizajn webovej stránky je vytvorený v Adobe Photoshop a uložený do formátu PSD, v ktorom je naďalej editovateľný. Následne po vytvorení bol dizajn rozsekaný na potrebné časti, ktoré sú uložené v „assets/images“.

Štruktúra webovej stránky je vytvorená v HTML, ktorý obsahuje univerzálne bloky, ktoré musia byť obsiahnuté na každej stránke:

- Hlavička (s navigáciou)
- Hlavná časť, do ktorej sa v Ruby vkladá výstup z každej podstránky
- Päťka (s navigáciou)

Ďalej bloky, ktoré sú obsiahnuté v úvodnej stránke:

- Úvodná slideshow
- Stručné vysvetlenie našej ponuky

A elementy užívateľského rozhrania:

- Odkazy
- Tlačidlá
- Formuláre
- Ikony

Štýlovanie týchto elementov je spravované prostredníctvom CSS, ktoré sa nachádza v dvoch súboroch:

- Master.css – kaskádové štýly, ktoré musia byť obsiahnuté na každej stránke
- Forms.css – kaskádové štýly pre formuláre

3.3.5 TECHNOLÓGIA

Adobe Photoshop CS6 – webdizajn

HTML5 – štruktúra webovej stránky

CSS3 – kaskádové štýly

3.3.6 DETAILY PODPORY

Adobe Photoshop - <http://helpx.adobe.com/photoshop/topics.html>

HTML5 - http://www.w3schools.com/html/html5_new_elements.asp

CSS3 - http://www.w3schools.com/css/css3_intro.asp

3.4 REGISTRÁCIA

Ako neregistrovaný používateľ sa chcem registrovať aby som mohol pracovať so systémom

3.4.1 ŠPECIFIKÁCIA

Na stránku príde nový užívateľ, ktorý sa chce zaregistrovať. Na stránke by mal byť jasne viditeľný odkaz pre registráciu. Po kliknutí na odkaz sa mu zobrazí registračný formulár. Následne po úspešnom vyplnení registračného formulára ho stránka automaticky prvýkrát prihlási.

3.4.2 ANALÝZA

Pre registráciu užívateľa je nutné vytvoriť prihlasovací formulár, ktorý bude obsahovať textové polia pre zadanie emailu mena a hesla a jeho overenie. Po odoslaní formulára sa validuje platnosť a jedinečnosť emailovej adresy. Overí sa minimálna dĺžka hesla. Po úspešnom overovaní sa užívateľ uloží do databázy užívateľov.

3.4.3 IMPLEMENTÁCIA

Pre registráciu sme vytvorili samostatný formulár skladajúci sa z textových polí meno, heslo, email, heslo a potvrdenie hesla. Všetky údaje sme nastavili ako povinné, pričom pri zadávaní emailu validujeme jeho pravosť pomocou regulárneho výrazu. Pri hesle kontrolujeme jeho minimálnu dĺžku, ktorú sme určili na 6 znakov. V rámci hesla tiež overujeme či je rovnaké ako potvrdzovacie heslo. Po správnom vyplnení formulára a potvrdení sa pre používateľa vytvorí účet zapísaním používateľa do tabuľky používateľov v databáze. Po potvrdení bude používateľ automaticky prihlásený na svoj profil. V prípade chyby vyskočil používateľovi error message, kde sme nastavovali aj správny tvar výpisu chyby v prípade množného čísla chýb.



Obrázok 5. Formulár registrácie používateľa

3.4.4 TESTOVANIE

Registráciu sme testovali vytvorením viacerých typov používateľov, ktorí robia rôzne neštandardné chyby obvyklé v E-maily.

3.5 PRIHLÁSENIE

Ako registrovaný používateľ sa chcem prihlásiť do systému

3.5.1 ŠPECIFIKÁCIA

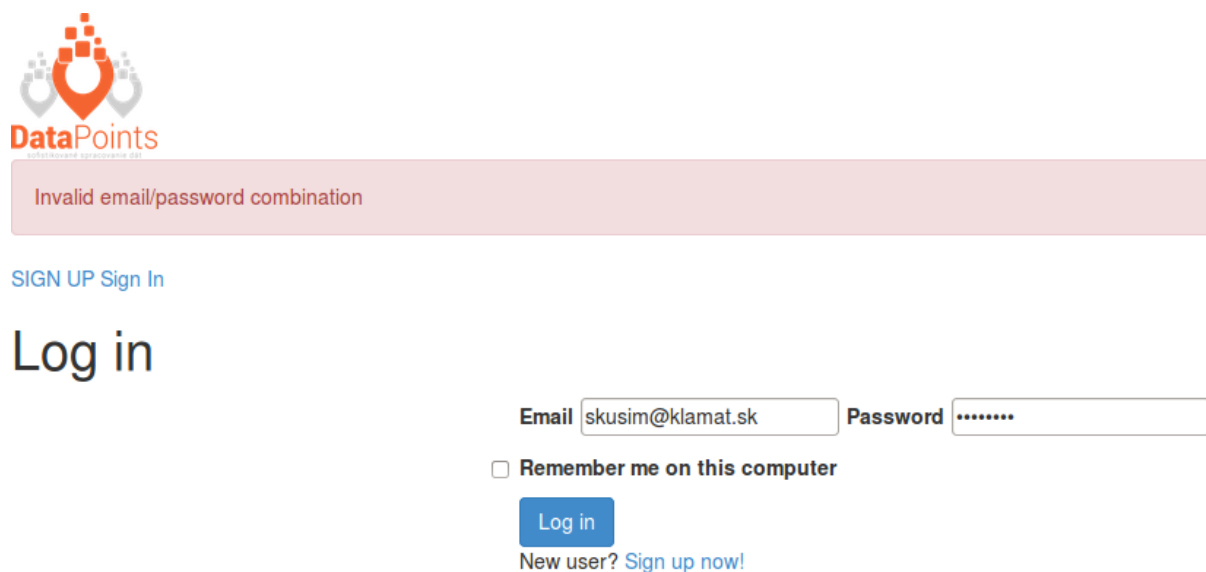
Na stránke bude jasne viditeľný odkaz pre prihlásenie registrovaných užívateľov. Po kliknutí na odkaz sa zobrazí formulár, ktorý požiada o prihlasovacie údaje. Po úspešnom overení stránka prihlási používateľa.

3.5.2 ANALÝZA

Pre prihlásenie sa pridá na hlavnú obrazovku odkaz, ktorý presmeruje používateľa na prihlasovací formulár. Prihlasovací formulár bude pozostávať z mena a hesla. Pre úspešné prihlásenie bude zadať správnu kombináciu mena a hesla. Meno a heslo sa budú overovať voči databáze na serveri.

3.5.3 IMPLEMENTÁCIA

Pre prihlásenie sme vytvorili samostatnú obrazovku s formulárom na prihlásenie, ktorý pozostáva z polí Email a heslo. Pri prihlasovaní overujeme existenciu užívateľa.



The screenshot shows the login interface for 'DataPoints'. At the top left is the logo with the text 'DataPoints' and 'efektívne spracovanie dát'. Below it, a red error message states 'Invalid email/password combination'. There are two links: 'SIGN UP' and 'Sign In'. The main heading is 'Log in'. The login form includes an 'Email' field with 'skusim@klamat.sk', a 'Password' field with masked characters, a checkbox for 'Remember me on this computer', a blue 'Log in' button, and a link for 'New user? Sign up now!'.

Obrázok 6. Prihlásenie používateľa

3.5.4 TESTOVANIE

Testovali sme zlé heslá, neexistujúcich používateľov, prípadne či sa dá obísť prístup bez prihlásenia.

3.6 ODHLÁSENIE

Ako prihlásený užívateľ chcem mať možnosť odhlásiť sa zo systému

3.6.1 ŠPECIFIKÁCIA

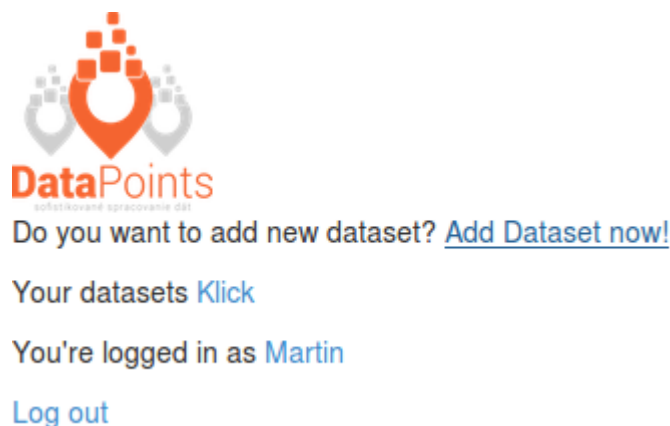
Po úspešnom prihlásení sa užívateľovi zobrazí možnosť na odhlásenie sa zo stránky. Po kliknutí na odkaz pre prihlásenie bude užívateľ automaticky odhlásený.

3.6.2 ANALÝZA

Odhlásenie bude dostupné zo všetkých obrazoviek pomocou samostatného odkazu. Po kliknutí na odkaz bude používateľ ihneď odhlásený.

3.6.3 IMPLEMENTÁCIA

Implementovali sme funkciu destroy, ktorá zruší session aktuálne prihláseného používateľa. V profile sa aktivuje po kliknutí na tlačidlo *Log out*.



Obrázok 7. Základné menu po prihlásení

3.6.4 TESTOVANIE

Testovali sme, či sa nedalo dostať do profilu po odhlásení napríklad cez špecifický odkaz profilu.

3.7 SPRÁVA PROFILU

Ako prihlásený užívateľ chcem vidieť svoj profil a mať možnosť upravovať ho (meno, e-mail, heslo)

3.7.1 ANALÝZA

Aby sa používateľ pri zadávaní hesla nepomýlil, je dôležité heslo overiť dvojnásobným zadáním. Z bezpečnostných dôvodov by mali byť znaky skryté.

3.7.2 NÁVRH

Na zmenu údajov som navrhol jednoduchý formulár so štyrmi hodnotami: Name, Email, Password a Confirm password. Pri zobrazení profilu používateľa sú vyplnené polia Name a Email hodnotami prihláseného používateľa. Na potvrdenie zmeny údajov slúži tlačidlo Update.

3.7.3 IMPLEMENTÁCIA

Správu profilu som implementoval v Ruby on Rails. Využil som preddefinované funkcie tohto frameworku. Pri stlačení tlačidla Update sa overí či sú vyplnené všetky polia, overí sa správnosť emailovej adresy, dĺžka hesla (minimálne 6 znakov) a skontroluje sa či sa zadané heslá zhodujú. Ak sú všetky údaje vyplnené správne, zapíšu sa do databázy do tabuľky users.

3.7.4 TESTOVANIE

Funkčnosť som overil editáciou svojich údajov. Vyskúšal som zadať prázdny reťazec, nesprávny tvar emailovej adresy, nedostatočnú dĺžku hesla, nesprávne potvrdzovacie heslo a program tieto vstupy vyhodnotil správne, ako nevalidované.



Do you want to add new dataset? [Add Dataset now!](#)

Your datasets [Klick](#)

You're logged in as [Martin](#)

[Log out](#)

Your profile

Name Email Password Confirm password

Obrázok 8. Zobrazenie profilu používateľa

3.8 VYTVORENIE OBRAZU DATASETU

Ako prihlásený používateľ chcem nahrať obraz datasetu na server

3.8.1 ANALÝZA

Po prihlásení sa používateľovi zobrazí možnosť nahrať nový dataset. Po kliknutí na túto možnosť sa používateľ presunie na formulár pre jeho pridanie, obrázok nižšie. Musí vyplniť povinný údaj Názov datasetu a Link na vzdialený súbor datasetu v tvare `http://`. Údaj poznámka je voliteľný. Po vyplnení údajov tlačidlom nižšie potvrdí Nahratie.

3.8.2 IMPLEMENTÁCIA

Implementovali sme nový model datasetu ako v Rails, tak aj v databáze, ktorý obsahuje atribúty *name*, *description* a *link*. Rails k nemu automaticky pridáva atribúty *created_at* a *updated_at*, ktoré nám poslúžia v ďalšej úlohe pre zobrazenie dátumov vytvorenia a úpravy. Následne sme vytvorili view formulára na Obr. č. . Pomocou funkcie *create* v controlleri sme implementovali samotnú funkcionálnosť, pričom referenciu práve prihláseného používateľa sme naviazali na práve nahrávaný dataset. Nastavili sme aby jeden používateľ mohol mať viacero datasetov.



Do you want to add new dataset? [Add Dataset now!](#)

Your datasets [Klick](#)

You're logged in as [Martin](#)

[Log out](#)

Add new Dataset

Name Description Link

Obrázok 9. Pridanie nového datasetu

3.8.3 TESTOVANIE

Testovanie prebehlo skúšaním rôznych http odkazov tak, aby bolo možné zadať iba korektný hypertextový odkaz.

3.9 ZOBRAZENIE, MAZANIE A EDITÁCIA DATASETU

Ako prihlásený používateľ chcem vidieť nahraný dataset a chcem mať možnosť zmazať a upraviť svoje nahrané datasety

3.9.1 ANALÝZA

Je potreba vytvoriť zoznam datasetov, ktoré používateľ priamo využíva. Jeho dostupnosť by mala byť z hlavného menu stránky, ktoré je vidieť na obrázku číslo 3. Zoznam datasetov by mal byť na umiestnení na samostatnej stránke. Mazanie a editácia základných údajov datasetov môže byť implementovaná v rámci zoznamu datasetov.

3.9.2 IMPLEMENTÁCIA

Do projektu bol implementovaný zoznam, ktorý sa nachádza na samostatnej stránke datasets/show. Zoznam je implementovaný pomocou knižnice bootstrap. Používateľovi sa v zozname na základe jeho id čísla zobrazujú datasety, ktoré chcel pridať. Zoznam obsahuje informácie o mene, opise, statuse, dátume pridania datasetu a dátum poslednej zmeny v datasete.

Zmena informácií o datasete je dostupná po kliknutí na tlačidlo "edit". Po kliknutí tlačidla do popredia vyskočí okno v ktorom je možné editovať len meno datasetu a opis datasetu. Vyskakovacie okno je spravené pomocou bootstrapovej knižnice modal. Mazanie datasetu je priamo dostupné zo zoznamu datasetu a to po kliknutí na tlačidlo Destroyed. Po kliknutí na toto tlačidlo sa daný dataset nezmaže len sa databáze zmení hodnota deleted na TRUE. Kvázy vymazaný dataset sa už nezobrazí viac používateľovi v tabuľke datasetov.

3.9.3 TESTOVANIE

Testovací scenár pre zobrazenie datasetov sa skladá z viacerých krokov:

1. Prvý krok pozostáva z pridania nového datasetu a následného skontrolovania zobrazenia v zozname.
2. Druhý krok pozostáva z editovania mena a opisu datasetu a následného uloženia zmien. Následne je nutné skontrolovať či sa zmeny prejavili i v zozname datasetov.
3. Tretí krok pozostáva z vymazania datasetu a kontrolovania jeho vymazania zo zoznamu datasetov.

CEZ VRCHY A POD VRCHMI

Číslo šprintu: 2

Začiatok šprintu: 23. 10. 2014

Koniec šprintu: 6. 11. 2014

Príbehy:

- Získavanie dát
- Plánovač sťahovania
- Ukladanie datasetov a Elasticsearch
- Typy datasetov
- Spracovanie datasetov
- Aplikácie 3. strán
- Vykresľovanie dát
- Obrazovky GUI
- Prispôsobenie GUI používateľom

3.10 ZÍSKAVANIE DÁT

3.10.1 PLÁNOVAČ SŤAHOVANIA

3.10.1.1 ANALÝZA IMPLEMENTÁCIE PLÁNOVANIA

3.10.1.2 MOTIVÁCIA K PLÁNOVANIU

Počas vývoja webových aplikácií je bežné, že požadujeme, aby sa niektoré typy úloh spracovávali asynchrónne. Takýmto typom úloh môže byť hocičo - plánovaná údržba, odosielanie emailov, dávkové spracovanie dát, http sťahovanie, úprava obrázkov... Ďalšou podmienkou ktorú požadujeme je, možnosť paralelného spracovania danej úlohy.

Riešením tohto problému býva presunutie spracovania úlohy z klienta na server. Samotný spôsob spracovania úlohy však nie je naprieč programovacími jazykmi rovnaký. Kým niektoré aplikačné servery poskytujú štandardizovaný spôsob tvorby nových vlákien, pri iných je nutné túto funkcionality implementovať. Potreby našej webovej aplikácie vyžadujú, aby sme boli schopní plánované úlohy uložiť a podľa priority ich spracovávať neskôr. Taktiež požadujeme, aby bolo možné prioritizáciu flexibilne upravovať, aby bolo možné obnoviť stav spracovávania aj po výpadku a aby spracovávanie jednej úlohy neblokovalo celý proces.

3.10.2 SPÔSOBY IMPLEMENTÁCIE

3.10.2.1 PLÁNOVANIE A VYKONÁVANIE ÚLOH NA SERVERI

Prvým spôsobom, ako vyriešiť problém plánovania je implementácia vlastného plánovacieho algoritmu priamo v prostredí serverovej časti webovej aplikácie. Nevýhody tohto prístupu však jasne prevyšujú jeho výhody. Vlastná implementácia je náchylnejšia na chyby, ťažko sa testuje, zlá implementácia spôsobí, že spracovanie jednej úlohy bude blokovat celý proces, jednotlivé úlohy budú úzko previazané s kontrolerom alebo dokonca pohľadom. Vylepšením tohto prístupu môže byť vykonávanie plánovaných úloh démonom.

3.10.2.2 PLÁNOVANIE ÚLOH V KÓDE A ICH VYKONÁVANIE VLASTNÝM DÉMONOM

Démon je program, ktorý beží dlhodobo na pozadí bez interakcie s používateľom. Oproti kompletnej správe a vykonávaní úloh je vykonávanie úloh démonom lepšou alternatívou. Tento prístup však v sebe zahŕňa skrytý problém: po naplánovaní úloh a ich zaradení do rady pre vykonávanie je ich ďalšia správa veľkým problémom napr. ak sa náhodou zmení prioritizácia úloh alebo treba konkrétnu úlohu presunúť medzi skupinami. Okrem uvedeného synchronizačného problému v správe vzniká problém aj so serializáciou aktuálneho stavu, čo by v konečnom dôsledku spôsobilo stratu údajov pri potenciálnom reštarte. Všetky tieto záležitosti by mohli zbytočne navýšiť čas potrebný pre korektnú implementáciu.

Existujúce gemy na vytváranie démonov:

- Daemons
- Daemon-kit

3.10.2.3 PLÁNOVANIE ÚLOH V KÓDE A ICH VYKONÁVANIE SPRAVOVANÉ EXISTUJÚCIM DÉMONOM

Tento prístup oproti predchádzajúcemu poskytuje iba malé vylepšenie, nakoľko naplnenie požadovaných vlastností závisí od funkcionality existujúcich démonov. Démoni vykonávania úloh napr. Cron však plánujú úlohy na základe času a nie priority, takže flexibilná zmena v pláne je veľmi komplikovaná.

Existujúce gemy na plánovanie:

- Whenever
- Resque-scheduler

3.10.3 POUŽITIE KOMPLETNÉHO PLÁNOVACIEHO SOFTVÉROVÉHO RÁMCA

Tento prístup vyzerá byť pre naše potreby najoptimálnejším. Väčšina rámcov je jednoducho rozširiteľných a už v základnej verzii poskytujú veľmi dobrú funkcionality. Chýbajúce vlastnosti teda vieme relatívne ľahko doimplementovať.

Existujúce gemy na kompletnú správu plánovania:

- Resque
- Sidekiq
- Delayed Jobs

3.10.4 ANALÝZA VYBRANÝCH PLÁNOVACÍCH GEMOV

RESQUEUE

1151 odnoží, 6139 obľúbení, vek 5 rokov

Je knižnica programovacieho jazyku Ruby určená na vytváranie úloh vykonávaných na pozadí, ich radenie do

skupín a neskoršie vykonávanie. Skladá sa z 3 častí:

- Knižnice na vytváranie a dopytovanie úloh,
- Rake skriptu na spustenie zamestnanca ktorý úlohy spracováva
- Aplikácie Sinatra na monitorovanie zamestnancov, úloh a skupín

Implementácia úlohy je jednoduchá, pretože ňou môže byť hocijaká trieda alebo modul.

SIDEKIQ

740 odnoží, 4452 obľúbení, vek 2 roky

Je kompletným plánovačom a vykonávačom spúšťaným z backendovej časti Rails webovej aplikácie. Výhodou

oproti konkurenčným nástrojom je jeho výkonnosť a efektívnosť. Sidekiq umožňuje jeho paralelnú integráciu s iným

plánovačom napr. Resque, pričom bude Sidekiq použitý iba ako vykonávač. Jeho open-source vývoj je však

sponzorovaný komerčnou verziou, ktorá poskytuje väčšinu zaujímavej funkcionality.

DELAYED JOBS

949 odnoží, 3222 obľúbení, vek 6 rokov

Poskytuje abstrakciu a spoločný rámec pre vytváranie plánovaných úloh. Pôvodne bol súčasťou webovej aplikácie

Shopify, ale kvôli možnosti jeho hromadnejšieho využitia bol publikovaný aj pre verejnosť.

Taktiež poskytuje

možnosť prioritizácie úloh a použitie databázy pre serializáciu stavu.

Výhody Nevýhody

- + Zrelosť kódu - Potrebná ďalšia databáza (Redis)
- + Serializácia úloh - Každá úloha vytvára nový proces
- + Monitorovanie stavu

Výhody Nevýhody

- + Plánované úlohy - Udržiava jeden človek
- + Každý nový job vytvára nový thread - Notifikácie o zmene stavu úlohy v PRO verzii
- + Profilované java profilerom v jRuby - Grupovanie úloh v PRO verzii
- + Webové rozhranie - Metriky v PRO verzii
- + Réžia jedného 300MB Sidekiq procesu je rovnako pamäťovo efektívna ako réžia desiatich 200Mb Resque procesov

3.10.5 MOŽNOSTI ĎALŠIEHO ROZŠÍRENIA VYBRANÝCH PLÁNOVACÍCH NÁSTROJOV

ABSTRAKCIA NAD PLÁNOVACÍMI RÁMCAMI

Active Jobs

Poskytuje jednotné rozhranie nad rozdielmi medzi API jednotlivých plánovacích rámcov. V praxi to znamená, že môžeme vymeniť plánovací nástroj bez toho, aby sme museli priamo meniť kód, stačí iba úprava konfigurácie

ActiveJobs.

MONITOROVANIE PLÁNOVACÍCH MECHANIZMOV

Activejob::Stats

Je rozšírením Active Jobs umožňujúcim zber štatistík z plánovacích rámcov a ich následné odosielanie na logovací a monitorovací server. Momentálne však podporuje iba server Statsd.

Výhody Nevýhody

- + Každá nová úloha vytvára nový proces - Potreba dedikovaného monitorovania
- + Plánované úlohy - Potrebná databáza
- + Serializácia úloh - Každý nová úloha vytvára proces
- + Hooky pre rozličné stavy úloh
- + Možnosť integrácie s PostgreSQL

Výhody Nevýhody

- + Ďalšia úroveň abstracie - Nutná nadbytočná konfigurácia
- + Jednotné API pre Sidekiq, Resque, Delayed Jobs - Vyššia réžia a nižšia efektivita

TÍMOVÝ PROJEKT - ANALÝZA IMPLEMENTÁCIE PLÁNOVANIA

3.10.6 NAVRHOVANÉ RIEŠENIE

Domnievam sa, že pre naše potreby by bolo vhodné použiť kompletnú knižnicu Delayed job. Dôvody sú nasledovné:

1. Kompatibilita s PostgresSql vďaka čomu odpadá nutnosť inštalovať ďalšiu databázu
2. Jednoduchá implementácia pozorovateľov stavu, sledovanie priebehu vykonávanej úlohy
3. Zrelosť kódu a podpora integrácie s ostatnými knižnicami
4. V prípade pamäťových problémov možnosť integrácia Sidekiq na vykonávanie úloh
5. Dobrá dokumentácia

3.10.7 REFERENCE:

- [1] <https://github.com/resque/resque>
- [2] <https://github.com/mperham/sidekiq>
- [3] https://github.com/collectiveidea/delayed_job

3.10.8 ANALÝZA IMPLEMENTÁCIE SŤAHOVANIA

3.10.9 MOTIVÁCIA K SŤAHOVANIU

Vývoj našej webovej aplikácie vyžaduje, aby sme boli schopní analyzovať súbory datasetov zo vzdialeného umiestnenia. Naskytujú sa nám preto dve možnosti:

- Používateľ ktorý má záujem analyzovať vybraný dataset ho vlastní a nahraje na serverovú časť našej webovej aplikácie
- Používateľ pozná URL a analyzovaný dataset bude stiahnutý serverovou časťou webovej aplikácie

V nasledujúcich riadkoch sa budem primárne venovať druhému spôsobu, prvý menovaný bude pravdepodobne podrobený ďalšej analýze v nasledujúcich šprintoch.

3.10.10 PROBLÉMY KTORÉ MÔŽU VZNIKNUŤ POČAS SŤAHOVANIA

Na stiahnutie vybraného datasetu je nutné vopred nadviazať so serverom spojenie, čo trvá určitý čas. Preto vyžadujeme, aby bolo sťahovanie asynchrónne a v ideálnom prípade plánované. Počas sťahovania taktiež môže dôjsť k chybe a preto potrebujeme ošetrovať chybové stavy a poškodené sťahovanie spustiť odznova. Datasety na vzdialenej lokácii môžu byť časom upravené a preto je nutné synchronizovať zmeny. Adresa, ktorú zadá používateľ môže obsahovať presmerovania, a preto treba vhodne zvoliť politiku, ako budeme dáta z podobných adries spracovávať. Problémovou môže byť aj situácia, keď sú dáta na vzdialenej lokácii chránené a náš sťahovač buď ani nenadviaže spojenie alebo nemá práva na čítanie datasetu.

3.10.11 SPÔSOBY IMPLEMENTÁCIE

3.10.11.1 INTEGRÁCIA EXISTUJÚCEHO SŤAHOVAČA A JEHO VOLANIE Z RUBY

Prvou variantou ako sťahovať dáta na server je využitie štandardných linuxových programov pre sťahovanie. Ako

príklad uvediem dva nástroje príkazového riadku Curl a Wget. Oba nástroje dokážu v rámci HTTP/HTTPS

protokolu sťahovať pomocou GET aj POST dopytov a oba podporujú cookies.

Curl

Curl však navyše podporuje aj automatickú dekompresiu v prípade, že by bol obsah komprimovaný pomocou

gzip. Curl taktiež podporuje protokoly viaceré protokoly FTP, FTPS, HTTPS, SCP, SFTP, LDAP, POP3, IMAP,

SMTP... Curl však nie je zamýšľaný pre rekurzívne sťahovanie.

Wget

Wget podporuje iba protokoly OpenSSL a GnuTLS a základnú HTTP autentifikáciu. Jeho hlavnou devízou je však

možnosť rekurzívneho sťahovania. Nanešťastie, my túto funkcionality pravdepodobne nevyžadujeme.

Po zvolení vhodného sťahovaču potrebujeme vykonať integráciu s webovou aplikáciou. Tu sa nám naskytujú

možnosti využitia objektu Kernel či vstavanej syntaxe pomocou reťazca “%x”. [2]

Potenciálnym riešením by mohlo byť aj vytvorenie sťahovacieho démona. Démon je program, ktorý beží dlhodobo na pozadí bez interakcie s používateľom. Jeho implementácia by však bola náročnejšia a ťažšie by sa testovala.

Nevýhodou uvedeného spôsobu je, že nás prakticky pripúta k linuxovej platforme, čím skomplikuje napríklad beh testov na vývojárskych strojoch.

3.10.11.2 IMPLEMENTÁCIA VLASTNÉHO SŤAHOVAČA

Druhým spôsobom, ako vyriešiť problém plánovania je implementácia vlastného sťahovaču priamo v prostredí serverovej časti webovej aplikácie. Ruby poskytuje aplikačné rozhranie nazvané Net::HTTP. Toto rozhranie poskytuje pomerne detailné nastavenia vytváraných dopytov. Umožňuje vytváranie vlastných hlavičiek, HTTPS dopytov, serializáciu na disk, automaticky dekomprimuje GZIP, udržiavanie spojenia či nasleduje presmerovania. Taktiež ponúka pohodlný prístup k odpovediam na dopyty. Nevýhodou tohto prístupu je samozrejme mierne náročnejšia implementácia ako v prvom prípade. Výhodou ale ostáva fakt, že uvedený prístup nevyžaduje žiadnu ďalšiu konfiguráciu na vývojárskych strojoch. Ďalšou obrovskou výhodou je možnosť monitorovania priebežného stavu sťahovania priamo v kóde resp. bežiacej aplikácii.

3.10.12 NAVRHOVANÉ RIEŠENIE

V prostredí Ruby on Rails je zaužívanou konvenciou zaradenie dlho trvajúcich úloh do radu úloh vykonávaných na

pozadí. Keďže Ruby je v zásade obmedzené na beh jedného vlákna v rámci procesu a Passenger a Nginx

obsluhujú v rovnakom čase iba jeden dopyt, považujem za vhodné využitie plánovača, ktorý bude sťahovanie

vykonávať. Vyhne sa tak problémom s nízkou odozvou webovej aplikácie. Príklady plánovačov sú uvedené v predchádzajúcej kapitole.

Navrhované riešenie je v ďalších krokoch závislé od funkcionality ktorú budeme od výslednej webovej aplikácie požadovať:

- Bude nutné flexibilne vytvárať postupnosť krokov predspracovania datasetu (odstránenie hlavičky z CSV, zmena oddeľovačov)? Bude obsah sťahovaných súborov validný vzhľadom k ich formátu (chýbajúce zátvorky v XML)?
- Budeme požadovať, aby sme podporovali sťahovanie pomocou rozličných protokolov?

Ak sme na obe otázky odpovedali nie, ideálnym bude implementácia vlastného sťahovaču. Ak sme aspoň na

jednu otázku odpovedali áno, bude potrebné zvážiť použitie existujúceho sťahovaču. Z popisu oboch sťahovačov

by som sa však skôr priklonil k nástroju Curl, ktorý poskytuje viacej funkcionality a zároveň k nemu existujú Ruby

adaptéry. Reportovanie aktuálneho stavu zo sťahovaču naspäť do webovej aplikácie však bude veľmi

problematické a preto by som skôr preferoval implementáciu vlastného riešenia.

3.10.13 REFERENCIE

- [1] <http://daniel.haxx.se/docs/curl-vs-wget.html>
- [2] <https://gist.github.com/JosephPecoraro/4069>
- [3] <http://ruby-doc.org/stdlib-2.1.4/libdoc/net/http/rdoc/Net/HTTP.html>

3.11 UKLADANIE DATASETOV A ELASTICSEARCH

Analýza vychádza z faktu, že systém bude musieť ukladať dáta rôzneho formátu, ako napríklad CSV, XML, SQL. Každý z týchto formátov má inú štruktúru a preto je potrebné ich transferovať do jednotného tvaru a následne ich uložiť do prislúchajúcej databázy. V nasledujúcich riadkoch analyzujeme rôzne spôsoby ukladania dát do databázy.

3.11.1.1 FORMÁT UKLADANÝCH DÁT

SQL súbor obsahuje príkazy SQL jazyka na modifikovanie objektovo relačných databáz. Tento súbor môže obsahovať dáta, ktoré je pomerne ľahko možné dosadiť do objektovo relačnej databázy bez pomoci manuálneho zadávania dodatočných informácií. Pri takýchto súboroch vzniká riziko, že ich vykonaním sa môžeme dostať do neželanej situácie. Preto je potrebné takéto súbory validovať, predtým ako sa vykonajú.

CSV

je jednoduchý súborový formát pre výmenu tabuľkových dát. Samotné nahranie dát tohto formátu do objektovo-relačnej databázy vyžaduje najprv vytvorenie tabuľky, z prislúchajúcimi stĺpcami, ktoré zodpovedajú jednotlivým dátam v CSV súbore. Následne je možné nahratie dát do vytvorenej tabuľky. Vytvorenie tabuľky na základe CSV súboru je možné len manuálne alebo automaticky pomocou skriptu na základe dát v súbore. Nevýhodou tohto riešenia môže byť nedostatočné rozpoznanie jednotlivých typov stĺpcov, čo by malo za následok manuálne opravovanie a kontrolovanie všetkých stĺpcov a ich dátových typov.

Alternatívnou formou je vytvorenie JSON objektu z každého riadku, ktorý je v CSV súbore a následne uloženie takéhoto formátu do jedného riadku objektovo relačnej databázy. Pred samotným uložením dát do databázy by bolo potrebné vytvoriť tabuľku JSON objektov, ktorá by mohla vyzeráť nasledovne:

PK	JSON_OBJECT
1	{ "name": "Angela Barton", "is_active": true, "company": "MagnaFone" }
2	{ "name": "Johan Bulltos", "is_active": true, "company": "MagnaFone" }

Skript transformovania dát z jedného formátu do druhého by následne v Ruby on Rails vyzeral:

```
require 'csv'
require 'json'
```

```
CSV.parse(data).to_json
```

Riešenie pomocou JSON objektov odkladá zložité generovanie tabuliek, avšak ponúka len obmedzenú funkcionálnosť pri dopytovaní informácií pomocou SQL jazyka v JSON objektoch a môže sa stať, že všetko na čo je programátor zvyknutý v SQL jazyku pri klasických dátových typoch nemusí byť podporované v dátovom type JSON. Toto obmedzenie je závislé predovšetkým na type databázy a type dát, ktoré ukladáme.

XML

XML je rozšírený značkovací jazyk. Nahratie XML súboru do databázy sa môže uskutočniť okamžite a podobne ako pri JSON objektoch sa najskôr vytvorí nová tabuľka v objektovo-

relačnej databáze a následne sa nahrajú dáta do tabuľky, ktorá by sa skladala z primárneho kľúča a dátového typu XML. Väčšina objektovo-relačných databáz síce poskytuje XML dátový typ, ale neposkytuje ďalšie operácie, ktoré by umožňovali lepší prístup k informáciám ktoré sú reprezentované v XML dátovom type.

Riešenie tohto problému môže byť pretransformovanie XML dátového typu na iný dátový typ. V Ruby on Rails je možné zmeniť XML na JSON objekt, s ktorým sa dá lepšie pracovať:

```
require 'json'
```

```
Hash.from_xml('<variable type="product_code">5</variable>').to_json
```

Ďalším riešením by mohlo byť dynamické vytváranie tabuliek pre jednotlivé XML súbory, čo by avšak nemuselo byť vždy úspešné kvôli rôznym typom údajov. Taktiež by to prinieslo veľa námahy ako zosúladiť takéto typy dát s objektovo-relačnými databázami. Iným riešením môže byť použitie špeciálnych databáz na XML, čo by avšak skomplikovalo prácu s inými dátovými typmi a celým ekosystémom aplikácie.

3.11.1.2 TYPY DATABÁZ

Výber databázy vo veľkom záleží na type dát a operáciami, ktoré chcem vykonávať nad danými dátami. Pokiaľ máme štruktúrované dáta a chceme zisťovať vzťahy medzi jednotlivými dátami a je dôležité dodržiavanie ACID vlastností, tak je vhodné použiť klasické objektovo-relačné databázy. Pokiaľ máme veľa neštruktúrovaných dát a ich formát je nám neznámy, tak je vhodné použiť NoSQL databázy [1].

NoSQL databázy sú vhodné, taktiež pre riešenia, kde je dôležitý okamžitý prístup k údajom avšak za cenu zníženia ACID vlastností. NoSQL databázy sú viac orientované na typy dát a ich prácu s nimi. Preto vznikli rôzne druhy NoSQL databáz, ako napríklad, grafové, orientované na kľúč-hodnota, dokumentové atď. Dopytovanie v týchto databázach je taktiež ťažkopádnejšie ako v bežných objektovo-relačných databázach. Kombináciou oboch databáz môžu byť databázy typu NewSQL, ktoré dosahujú rovnakú výkonnosť ako NoSQL databázy a pritom zaručujú dodržanie ACID vlastností [1], [2].

PostgreSQL

Voľne šíriteľná objektovo-relačná databáza, ktorá podporuje dotazovací jazyk SQL. PostgreSQL podporuje dátový typ JSON, ktorý je možné uložiť do databázy a taktiež použiť nad ním operácie a metódy. Najnovšia verzia 9.3 podporuje sadu niekoľkých operácií, ktorých využitie v praxi je možné nájsť tu [6]. Príklad dotazu na informácie v dátovom type JSON je nasledovný:

```
INSERT INTO books VALUES (1, '{ "name": "Book the First", "author": {  
  "first_name": "Bob", "last_name": "White" } }');
```

```
SELECT id, data->'author'->>'first_name' as author_first_name FROM books;
```

```
id | author_first_name
```

```
----+-----
```

```
1 | Bob
```

Ako je vidieť z príkladu použitie SQL nad JSON objektmi je jednoduché a veľmi podobné klasickému SQL štandardu. Dopyty podporujú filtrovanie, agregáciou pomocou GROUP_BY a ďalšie iné dopytovacie metódy [6].

V novej verzii PostgreSQL 9.4 je plánované rozšírenie o nový dátový typ JSONB. Rozdiel medzi klasickým JSON dátovým typom a typom JSONB je v ich ukladaní na dátový nosič. JSON sa ukladá v klasickom v texte a tak zaberá menej miesta ako JSONB, ktorý sa ukladá v binárnom formáte. Využitie týchto dátových typov závisí od ich použitia v databáze. V prípade, že

databáza slúži len na uchovávanie JSON objektov, tak je vhodné použiť JSON dátový typ. V opačnom prípade, keď sú vykonávané viaceré operácie nad JSON objektmi, tak je vhodné použiť dátový typ JSONB. Alternatívou pre PostgreSQL je MySQL, MongoDB.

MongoDB

Je dokumentovo orientovaná databáza, ktorá je klasifikovaná ako NoSQL databáza. Využíva sa hlavne na uchovávanie JSON objektov. Jednotlivé schémy a tabuľky sa vytvárajú genericky, preto poskytuje vysokú škálovateľnosť. Je bežne používaná všade tam, kde je potrebné rýchla dostupnosť dát. Alternatívou môžu byť rôzne iné NoSQL databázy [5].

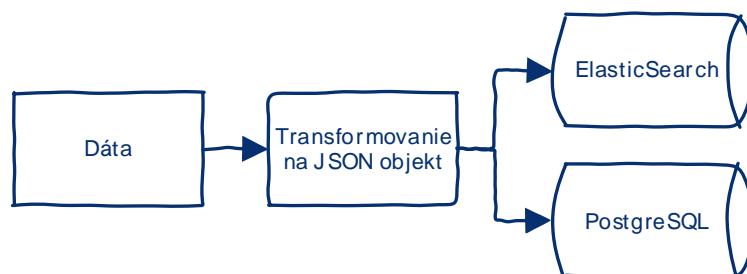
Elasticsearch

Je predovšetkým fulltextový vyhľadávač, ktorý môže byť použitý na indexovanie a ukladanie jednotlivých JSON objektov. Elasticsearch plne podporuje CRUD (create, read, update, delete). Taktiež je dobrý pre agregáciu viacerých JSON objektov a ich filtrovanie. Elasticsearch spracúva požiadavky v reálnom čase a taktiež poskytuje kontrolu preklepov, resp. funkciu automatickej kontroly chýb. Prípady použitia Elasticsearch väčšinou súvisia so spracovaním štruktúrovaných textov, napríklad statusov na sociálnych sieťach alebo spracovaním záznamov rôzneho formátu. Využitie elasticsearch ako databázy sa môže využiť všade tam, kde nás strata dát nemusí až tak trápiť, avšak pre plnohodnotné zabezpečenie vlastností ACID sa odporúča zálohovať dáta v externej databáze [3]. Alternatívou pre Elasticsearch je Apache Solr.

3.11.2 NÁVRHY ARCHITEKTÚR UKLADANIA DÁT DO DATABÁZ

Pri navrhovaní databázovej architektúry v našom projekte je potrebné si uvedomiť, že neviem presne určiť aké typy dát budeme spracovávať. Preto je vhodné navrhnuť takú architektúru, ktorá bude dostatočne flexibilná spracovať rôzne typy dát.

Na obrázku číslo 9 je vidieť prvú navrhovanú architektúru. Všetky dáta sa budú transformovať do jednotného tvaru JSON objektu, ktorý sa bude uchovávať v PostgreSQL databáze a analyzovať v Elasticsearch. Využitie takejto architektúry nám poskytne jednotnú formu dát a to v podobe JSON objektov, ktorých analýza bude prevažne závislá na vyhľadávacom engine, ktorý nie je primárne určený na analyzovanie dát a ich vzťahov medzi sebou. PostgreSQL databáza nám poskytne len malé množstvo funkcií a metód, s ktorými budeme môcť pracovať preto je vhodné túto architektúru prehodnotiť vzhľadom na typ dát, aké naša aplikácia spracuje a na funkcie, ktoré chceme aby poskytovala.

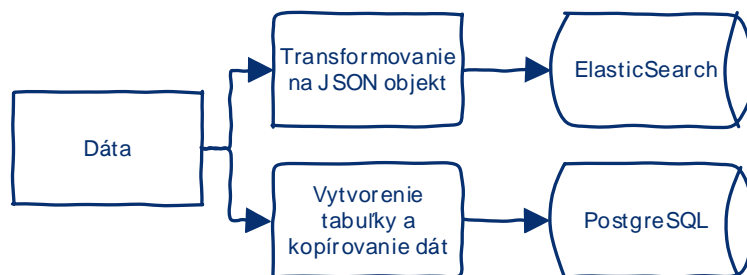


Obrázok 9. Návrh ukladania dát do databázy

Odpoveďou na prvú navrhovanú architektúru je architektúra znázornená na obrázku číslo 2. Táto architektúra poskytuje oveľa väčšie možnosti práce s dátami, pretože dáta budú jednak reprezentované štandardnými dátovými typmi, ale taktiež JSON objektmi v Elasticsearch. Nevýhodou takejto implementácie je väčšia námaha pri implementovaní takéhoto typu architektúry. Dáta, ktoré by sa transformovali z XML alebo CSV súborov do klasickej tabuľky, by potrebovali validovanie samotnými používateľmi. Celkovo by takáto architektúra poskytla

zaujímavé možnosti skúmania vzťahov medzi dátami za pomoci klasického dopytovacieho jazyka SQL.

Dáta Transformovanie na JSON objekt PostgreSQL ElasticSearch Vytvorenie tabuľky a kopírovanie dát



Obrázok 10. Návrh ukladania dát do databázy

Pri oboch riešeníach si je potrebné uvedomiť, že je potrebné vymyslieť zautomatizované indexovanie nad väčšími množstvami dát a to hlavne pri použití objektovo-relačných databáz. Pri implementácii je taktiež možné využiť alternatívne databázy na rozdiel od tých čo sú zobrazené na obrázkoch 9 a 10.

3.11.3 ZDROJE

[1] Todd Hoff, 35+ Use Cases For Choosing Your Next NoSQL Database, Jún 2011, [online]

<http://highscalability.com/blog/2011/6/20/35-use-cases-for-choosing-your-next-nosql-database.html>

[2] Todd Hoff, 101 Questions To Ask When Considering A NoSQL Database, Jún 2011, [online]

<http://highscalability.com/blog/2011/6/15/101-questions-to-ask-when-considering-a-nosql-database.html>

[3] Karussell, Jetslide uses ElasticSearch as Database, Júl 2011, [online]

<http://karussell.wordpress.com/2011/07/13/jetslide-uses-elasticsearch-as-database/>

[4] Moshe Kaplan, When to Use MongoDB Rather than MySQL (or Other RDBMS): The Billing Example, Marec 2014, [online]

<http://java.dzone.com/articles/when-use-mongodb-rather-mysql>

[5] Sarah Mei, Why You Should Never Use MongoDB, November 2013, [online]

<http://www.sarahmei.com/blog/2013/11/11/why-you-should-never-use-mongodb/>

[6] Dave Clark, What can you do with PostgreSQL and JSON?, Júl 2014, [online]

<http://clarkdave.net/2013/06/what-can-you-do-with-postgresql-and-json/>

3.12 TYPY DATASETOV

3.12.1 ANALÝZA TYPOV DATASETOV

Je potrebné analyzovať niekoľko (aspoň 5) rozličných typov datasetov a dát v nich plus je nevyhnutné vytvoriť rôzne testovacie vzorky.

Nižšie sú jednotlivé rôzne vzorky datasetov s ich popisom hlavičiek pre pochopenie obsahu.

Najčastejší a preferovaný formát je CSV, prípadne TXT, čo je v podstate rovnaký tvar dát väčšinou oddelený bodkočiarkou alebo čiarkou. Viacero datasetov bolo vyexportovaných do XML prípadne už v Excelovskom formáte XLS. Pri získavaní dát z datasetov treba dať pozor hlavne na riadky pred hlavičkou, ktoré môžu slúžiť ako poznámky prípadne boli vygenerované zároveň s datasetom. Takisto niektoré datasety obsahovali hlavičku viackrát, typicky export z PDF alebo z nejakých dokumentov, kde sa tabuľka kopíruje na viac strán.

Vzorky datasetov sú uploadnuté na stránke v adresari patterns:

<http://labss2.fiit.stuba.sk/TeamProject/2014/team03issi/datasets/>

3.12.2 ÚLOHY DO ĎALŠIEHO ŠPRINTU VYPLÝVAJÚCE Z TEJTO ANALÝZY

- Nahranie vzorky na Datapoints server
- Implementácia GEMU na parsovanie CSV do objektov
- Nahranie dát do databázy

3.12.3 DATASETY

1. domeny.txt

Dataset obsahuje názvy všetkých slovenských domén, ako aj ich registrátora a iných identifikátorov. Dataset je aktuálny ku 24.10.2014 04:40.

Hlavička:

- *domena* - názov domény
- *ID reg* – identifikátor registrátora domény
- *ID drzitela* – identifikátor držiteľa domény (ak doména nie je premigrovaná tak ako identifikátor sa použije IČO)
- *flag NEW/OLD*
 - NEW doména je premigrovaná resp. registrovaná v novom systéme
 - OLD doména nie je premigrovaná
- *Stav domeny* – stav v ktorom sa nachádza doména
- *NS1-4*- NS záznamy pre doménu
- *ICO drzitela* – IČO držiteľa domény

2. job_uchadzaci.csv

Dataset obsahuje údaje o uchádzačoch o zamestnanie, pričom analyzuje dáta medzi rokmi medzi 2001-2013

a rozdeľuje uchádzačov na viacero typov ako absolventi, mladí ľudia, ZP, dlhodobo evidovaní a pod.

Hlavička:

- *Počet uchádzačov o zamestnanie so ZP*
- *Počet uchádzačov o zamestnanie absolventi*
- *Počet uchádzačov o zamestnanie mladiství*
- *Počet dlhodobo evidovaných uchádzačov o zamestnanie*

3. zoznam_datasetov.xml

Dataset vo formáte XML obsahuje zoznam všetkých datasetov štátnej správy.

Hlavička:

- *porc* – ID datasetu
- *nazov* – názov datasetu
- *ucel* – popis účelu datasetu
- *prevadzkovatel* - rezort, ktorý prevádzkuje dataset
- *institucia* – konkrétna inštitúcia, niekedy rovnaká ako rezort
- *stav* – stav datasetu, či sa jedná o elektronickú/papierovú formu, (ne)štrukturovaný, a pod.
- *format* - formáty, v ktorých sa dataset nachádza
- *rozsah* – počet záznamov, alebo veľkosť prípadne iné.
- *cas* – ako často sa aktualizuje
- *specifikacia* – popisuje hlavičky datasetu prípadne iné
- *zverejnitelnost* – informácia o zverejniteľnosti datasetu
- odovodnenie

- *plan* – dátum datasetu
- *vyjadrenie* – či bol odsúhlasený alebo sa rokuje prípadne iné.

4. zoznam_datasetov.xls

Dataset je rovnaký s predchádzajúcim pričom pre porovnanie je vo formáte xls.

5. volby_prezident.csv

Dataset obsahuje informácie v jednotlivých regiónoch o hlasovaní v oboch kolách prezidentských volieb 2014. Do hlavičky sme vybrali druhé kolo.

Hlavička pre druhé kolo:

- *municipality_uid* - ID obce
- *ward* - okres
- *municipality* - obec
- *r2_precincts* – počet okrskov
- *r2_voters_eligible* – počet oprávnených voličov
- *r2_ballots_given_out* – počet rozdáných lístkov
- *r2_ballots_cast* – odovzdaných hlasov
- *r2_ballots_valid* – platných hlasov
- *r2_voter_turnout_pct* – volebná účasť v %
- *r2_ballots_cast_pct* – odovzdaných hlasov v %
- *r2_ballots_valid_pct* – platných hlasov v %
- *r2_count_fico* – počet hlasov Fico
- *r2_pct_fico* – percentuálne vyjadrenie Fico
- *r2_count_kiska* – počet hlasov Kiska
- *r2_pct_kiska* – percentuálne vyjadrenie Kiska

6. dlznici_zdravotna.csv

Dataset obsahuje zoznam dlžníkov z radov firiem na zdravotnom poistení k 20.10.2014.

Hlavička:

- *Obchodné meno*
- *PSČ*
- *Ulica*
- *Mesto / Obec*
- *IČO*
- *Výška pohľadávky*
- *Typ platiteľa* - informácia, či sa jedná o SZČO alebo Zamestnávateľa

7. medzinarodna_doprava.csv

Dataset obsahuje medzinárodné autobusové trasy zo slovenska.

Hlavička:

- *Číslo* – ID trasy
- *Odkiaľ* – mesto odkiaľ sa začína
- *Kam* – mesto kde končí
- *Číslo - rozh.* Číslo rozhodnutia o trase
- *Spoločnosť* – názov spoločnosti
- *Dátum - platnosti* dokedy platí trasa
- *Štát nástupu*
- *Štát výstupu*

8. kriminalita_na_mladezi.csv

Dataset obsahuje kriminalitu spáchanú na mládeži za rok 2012. Je špecifický tým, že ako oddelovač používa klasickú čiarku. Je krátky, takže by sa dal považovať za už spravenú štatistiku, kde oddeľuje jednotlivé druhy kriminality podľa jednotlivých typov osôb.

9. zoznam_faktur.xml

Dataset obsahuje všetky faktúry štátu za rok 2012. Je to XML formát podľa hlavičky exportovaný z PDF. Tento dataset je špecifický aj tým, že hlavička sa objavuje znovu po každej osmici dát.

Hlavička:

- *Identifikačný údaj faktúry*
- *Popis fakturovaného plnenia*
- *Celková hodnota plnenia*
- *Identifikácia zmluvy*
- *Identifikácia objednávky*
- *Dátum doručenia faktúry*
- *Identifikačné údaje dodávateľa*

10. evidencia_hosp_zvierat.csv

Dataset obsahuje evidenciu všetkých hospodárskych zvierat v SR. Generovaný bol 5.2.2013.

Hlavička:

- *dátum aktualizácie*
- *číslo farmy*
- *názov farmy*
- *ulica*
- *číslo*
- *obec*
- *okres*
- *kraj*
- *druh zvierat* - obsahuje skratky ako HD pre hydinu a pod.
- *majiteľ*
- *ulica*
- *číslo*
- *obec*
- *PSČ*

3.13 SPRACOVANIE DATASETOV

Tento dokument obsahuje analýzu k možnostiam spracovania datasetov ich analýzy, strojového učenia v Ruby. Tiež sa zaoberá možnosťami prepojenia ruby s jazykom R.

Pre spracovanie datasetov som nenašiel žiadne vhodné gemy preto všetky úpravy a narábanie s datasetmy bude nutné vytvoriť. Pri spracovaní datasetov vystávajú dva hlavné problémy a to vyčistenie datasetu od nepotrebných alebo zdvojených dát pre korektné zobrazovanie štatistík a grafov. Druhou úlohou je identifikácia dát a to v zmysle ich významu, či ide o mestá, čísla reprezentujúce roky, percentá alebo iný údaj. Identifikácia dát je dôležitá pre automatizovanie procesu pri vytváraní prvého náhľadu na dataset.

3.13.1 ANALÝZA DATASETŮV

Pre spracovanie dát nachádzajúcich sa v datasetoch som našiel nasledovné gemy pre Ruby:

Statsample

Dostupná na : <https://github.com/clbustos/statsample>
<https://github.com/sciruby/statsample-glm>
<http://www.rubydoc.info/gems/statsample-timeseries/0.0.3/frames>

Tento gem poskytuje ako základné tak aj pokročilé štatistické funkcie. Funkcie sú uvedené na stránkach gemu. Gem som úspešne nainštaloval a otestoval na vybranom príklade zo stránok.

Pri inštalácii nedošlo k žiadnym chybám. Pre tento gem boli vytvorené ďalšie rozširujúce gemy a to **Statsample TimeSeries** a **Statsample GLM**. Tieto gemy rozširujú Statsample o funkcie autokorelácie, ktorá umožňuje hľadanie opakujúcich sa vzorov, autoregresívne modely využívané na opis náhodných procesov, ktorých správanie sa dá predpovedať na základe správania z minulosti. Ďalej poskytujú poissonovu regresiu využívanú pri modelovaní kontingenčných tabuliek a logistickú regresiu umožňujúcu napr. odhad ako bude niekto voliť na základe demografických údajov.

Výhody

- Gem je aktuálny posledný release bol 11.10.2014
- Podporuje nové verzie ruby
- Poskytuje rozsiahle štatistické funkcie
- Umožňuje priame čítanie a zápis do databázy, CSV a Excel súborov

Nevýhody

- Slabá dokumentácia
- Nutnosť podrobnejšieho sa oboznámenia sa s poskytovanými funkciami

Descriptive statistics

Dostupná na : https://github.com/thirtysixthspan/descriptive_statistics

Gem umožňujúci základné štatistické funkcie ako priemer, medián, modus, štandardná odchýlka a percentil.

Výhody

- Jednoduchosť gemu a práca sním

Nevýhody

- Žiadna dokumentácia len príklady
- Poskytuje len štatistické minimum

Descriptive-statistics

Dostupná na : <https://github.com/jitescher/descriptive-statistics>

Gem umožňujúci základné štatistické funkcie ako priemer, medián, modus, rozsah, minimum, maximum, percentil pre danú hodnotu alebo hodnotu pre daný percentil. Gem som odskúšal. Počas skúšky nenastali žiadne problémy.

Výhody

- Jednoduchosť gemu a práca sním

Nevýhody

- Žiadna dokumentácia len príklady
- Poskytuje len štatistické minimum

3.13.2 STROJOVÉ UČENIE V RUBY

Pre využitie strojového v ruby existuje viacero prístupov.

WEKA

Prvým prístupom je využitie populárneho softvéru WEKA napísaného v JAVE. WEKA je silný nástroj so širokou paletou ponúkaných algoritmov pre strojové učenie a data mining.

Weka je dostupná na : <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

Pre sprístupnenie funkcionality WEKY v ruby je nutné použiť gem, ktorý prepojí ruby s JAVOU. Na tento účel slúži gem *rjb*.

Rjb je dostupný na: <https://github.com/arton/rjb>

Na stránke : <http://www.tylerclemons.com/weka-and-ruby/> je popísaný krátky príklad ako pracovať s *rjb* a WEKOU v Ruby

Apache mahout

Druhou možnosťou je využitie Apache mahout knižnice ktora je tiež napísaná v JAVE. Pre využitie tejto knižnice by sa mohol dať využiť rovnaký postup ako v prípade WEKY. Druhou možnosťou implementácia funkcionality v inom jazyku napr. JAVA, JRuby. Tieto jazyky fungujú na báze JVM a preto je možné v nich priamo využívať spomenuté knižnice. Po implementácii v inom jazyku by bolo potrebné zabezpečiť aj komunikačný kanál medzi hlavnou aplikáciou a modulom pre strojové učenie.

Tretou možnosťou je využitie Ruby gemu, ktorý poskytuje strojové učenie priamo v Ruby. Pre prácu s AI a strojovým učením v ruby existujú viaceré gemy, ktoré sa zaoberajú problémami klasifikácie, klustrovania npr Ruby Band . Ďalším gemom poskytujúcim funkcionality strojového učenia a AI je gem *AI4R*. Tento gem poskytuje rôzne algoritmy z oblasti strojového učenia a AI. Výhodou oboch knižníc je že sú aktívne spravované . Ich nevýhodou je slabá dokumentácia.

Gemy sú dostupné na : <https://github.com/SergioFierens/ai4r>
<https://github.com/arrigonalberto86/ruby-band>

3.13.3 PREPOJENIE R A RUBY

RinRUBY

Dostupné na: <https://github.com/clbustos/rinruby>

Je knižnica napísaná v Ruby. Je to jeden skript, ktorý sprístupňuje funkcie jazyka R priamo z Ruby. Skúšobná inštalácia neúspešná pri spustení Ruby servera boli hlásené chyby.

Výhody:

- Nevyžaduje R
- Pomalšie ako RsRuby ale robustnejšie

Proti:

- Pomalé pri priradovaní
- Limitované na dátové typy vektor a matica
- Projekt je neaktívny
- Slabá dokumentácia

RsRuby

Dostupné na: <https://github.com/alexgutteridge/rsruby>

Premosťujúca knižnica pre ruby umožňujúca prístup ku všetkým funkciám R priamo z Ruby scriptu. Rsruby predstavuje čiastočnú konverziu knižnice RPy.

Výhody:

- Super rýchle 5-10x rýchlejšie ako Rserve a 100-1000x ako RinRuby
- Bezproblémová integrácia s ruby každá metóda a objekt sú zaobchádzané ako Ruby objekt.

Nevýhody:

- Naposledy aktualizované v novembri 2011
- Závisle od operačného systému, implementácie ruby a verzie R
- Slabá dokumentácia

RSERVE

Dostupné na: <https://github.com/clbustos/Rserve-Ruby-client>

100% ruby

Používa TCP/IP sokety pre výmenu dát a príkazov. Vyžaduje Rserve inštaláciu.

Výhody:

- Relácie umožňujú asynchrónne spracovanie dát
- Rýchle 5-10 rýchlejšie ako RinRuby

Nevýhody:

- Vyžaduje rserve
- Slabá dokumentácia
- Neaktívny projekt

3.14 APLIKÁCIE 3. STRÁN

3.14.1 AUTENTIFIKÁCIA

Na našej webovej aplikácii chceme mať možnosť pre používateľa registrovať sa následne na to používať naše webové služby pomocou prihlásení rôznych 3. strán. S najväčšou určite to bude napríklad Facebook a Google+. Na túto autentifikáciu je najlepšie použiť knižnicu Omniauth. Omniauth je knižnica ktorá štandardizuje autentifikáciu viacerých poskytovateľov webových

aplikácií. Bola vytvorená aby bola výkonná, bezpečná a flexibilná. Táto knižnica podporuje možnosť autentifikácie Facebook aj Google+ takže nebude potrebné používať viacero systémov.

API: <http://intridge.github.io/omniauth/>

3.14.2 GEOLOGICKÉ DÁTA

Jednou z možností dát ktoré môžu obsahovať používateľom nahraté datasety sú napríklad geologické dáta ako napríklad súradnice alebo adresy. Na spracovanie s takýmito dátami je najlepší nástroj Geocoder, ktorý poskytuje veľké množstvo pracovania geologickými dátami. Jednou z týchto funkcií je napríklad prevedenie geologických súradníc na adresu. Táto knižnica má taktiež priamu integráciu s Google maps API. Bohužiaľ je táto integrácia spravená takým spôsobom, že výstup z Google maps je možné zobrazíť iba ako statický obrázok s ktorým sa nedá ďalej pracovať a preto na túto funkciu použije iný nástroj.

API: <http://www.rubygeocoder.com>

3.14.3 GOOGLE MAPS

Na integráciu našej aplikácie s Google maps existuje gem s názvom Gmaps4rails. Tento nástroj poskytuje všetky možnosti ktoré potrebujeme a mnoho ďalších ktoré my zatiaľ nepotrebujeme ale boli by možnosťou rozšírenia. Gmaps4rails je vyvinutý tak aby bol jednoducho vytvoril Google mapu s vrstvami (značky, informatické okná). Napriek tomu je založený na flexibilnom kódovom základe.

API: <https://github.com/apneadiving/Google-Maps-for-Rails>

3.14.4 WOLFRAMALPHA

Táto multifunkčná služba by sa dala použiť na mnoho vyhľadávaní informácií ktoré sa nachádzajú v datasetoch. Dokáže totižto pracovať napríklad priamo s Wikipédiou. Ako príklad si môžeme uviesť vyhľadávanie informácií o ľuďoch kde nájde o človeku jeho základné informácie ako celé meno, rok narodenia, miesto narodenia ale aj napríklad povolanie, krajinu pôsobenia ako aj informácie z Wikipédie. Tento nástroj je taktiež možné použiť na vyhľadanie informácií o mestách kde dokáže zistiť veci ako napríklad počet obyvateľov alebo najbližšie veľké mestá. Myslím si, že tento nástroj bude mať v našej aplikácii veľmi široké využitie pretože dokáže pracovať s veľmi veľa užitočnými informáciami.

API: <http://products.wolframalpha.com/api/>

3.14.5 VYHĽADÁVANIE ĽUDÍ A FIRIEM

Ďalšou možnou informáciou nachádzajúcou sa v datasete. Na tento typ informácie existuje mnoho aplikácií ale väčšina z nich je platená čo pre náš projekt neprichádza do úvahy. Preto pre naše potreby bude najlepšie použiť Facebook a WolframAlpha. Facebook je najrozšírenejšou sociálnou sieťou a nachádza sa na nej takmer každý. Nanešťastie na Facebook sa nachádzajú iba informácie ktoré tam daný človek zavesil a zdieľa s ľuďmi. Na druhú stranu WolframAlpha pracuje s informáciami ktoré sú verejne dostupné na napríklad na Wikipédii. Obe tieto služby sa dajú využiť na vyhľadanie konkrétnej osoby alebo firmy.

3.14.6 PIPL

Táto webová aplikácia poskytuje nástroje na vyhľadávanie informácií o ľuďoch kde prehľadáva mnoho známych sociálnych sietí ako napríklad Twitter alebo Facebook. Toto vyhľadávanie je

možné realizovať podľa mena, emailu, username alebo telefónneho čísla a voliteľným parametrom je mesto alebo krajina. API Pipl je spoplatnená ale majú možnosť pre neziskové organizácie poskytnutie svojich knižníc bez poplatne. Treba im ale poslať formulár s požiadavkou a opísaním projektu pre ktorý bude ich knižnica použitá.

API: <http://dev.pipl.com>

3.14.7 FINSTAT

Táto webová aplikácia poskytuje API ktoré si môže integrovať do svojej aplikácie bezplatne každý vývojár a poskytuje mu prístup k dátam ktoré sa na FinStat nachádzajú. Základné FinStat API obsahuje napríklad, IČO, DIČ spoločnosti, adresu sídla, informácie o tržbách a zisku alebo strate podniku. Firmy sa dajú vyhľadávať buď podľa názvu alebo ich IČO. Táto služba ale bohužiaľ funguje iba na slovenské firmy a informácie o nich.

API: <http://www.finstat.sk/hromadny-export-dat>

3.14.8 CAPTCHA

CAPTCHA (skratka pre “Completely Automated Public Turing test to tell Computers and Humans Apart”) je druh testu výzva-odpoveď používaný v aplikáciach na zistenie či je používateľ človek. Funguje takým spôsobom, že poskytne obrázok na ktorom sa väčšinou nachádzajú písmená a čísla a používateľ musí tento obrázok prepísať. Zatiaľ existuje je veľmi málo technológií ktoré dokážu prelomiť tento systém ochrany pred automatizovanými botmi. Samozrejme existujú aj rôzne iné CAPTCHE ako len text a čísla, a to také ktoré používajú napríklad obrázky zvierat alebo zvuk na rozpoznanie človeka.

Implementácia: <http://richonrails.com/articles/recaptcha-and-rails>

3.15 VYKRESĽOVANIE DÁT

3.15.1 D3JS

Táto Java Script-ová knižnica, je veľmi jednoduchá na integráciu, ktorá sa realizuje jednoduchým includom v HTML, s odkazom na:

- Interný súbor, stiahnutý z <https://github.com/mbostock/d3/releases/download/v3.4.13/d3.zip>
- Externý zdroj <http://d3js.org/d3.v3.min.js> , kde sa nachádza vždy najnovšia verzia

Podporuje zobrazovanie nie len základných typov grafov ako Line chart, Pie Chart, Bar chart, Grouped bar chart, Donut charts, ale aj omnoho sofistikovanejších, postavených na základoch D3JS. Knižnice sú dostupné na GIT Hub, pod MIT licenciou (voľne šíriteľné, upravovateľné, použiteľné na komerčné účely, no bez zodpovednosti za funkčnosť). Nás môžu zaujímať najmä nasledujúce:

- **Crossfilter** (<http://square.github.io/crossfilter/>) – knižnica určená na prehľadávanie a filtrovanie v datasetoch obsahujúcich viacero premenných. Filtre je možné vykresliť ako graf, pričom x-ová os zobrazuje premennú, a y-nová os sumu výskytu. V týchto grafoch je možné vyznačovať určitú čas x-ovej osy, pričom sa mení obsah nie len v tabuľke zodpovedajúcich výsledkov, ale aj ostatných filtrov (viz. Príklad v linku hore). Je navrhnutý na prácu s pomerne veľkými datasetmi, pričom tvrdí, že dokáže reagovať na interakciu s používateľom v reálnom čase (a to pod 30 milisekúnd).

- **Sortable Bar Chart** (<http://bl.ocks.org/mbostock/3885705>) - ponúka zobrazenie dát do jednoduchého Bar chart, s možnosťou rýchleho a animovaného zotriedenia podľa y-novej osy, s logikou od najväčšieho po najmenší.
- **Process Map** (<http://nylen.tv/d3-process-map/graph.php?dataset=les-mis>) - interaktívna a samo-zoradujúca sa procesná mapa. Link na sťahnutie sa nachádza na <https://github.com/nylen/d3-process-map>

Plusy	Mínusy
Jednoduchá inštalácia	Sofistikované nástroje obsahujú slabú, prípadne žiadnu dokumentáciu
Je na ňom postavené kvantum knižníc poskytujúcich sofistikované zobrazenie dát, na MIT licencií	Základná (defaultná) verzia slabú, alebo žiadnu interaktivitu s užívateľom
	Základná (defaultná) verzia slabú, alebo žiadnu animáciu

Tabuľka 1. Hodnotenie D3JS

3.15.2 HIGH CHARTS

Extenzívny a moderný nástroj pre zobrazovanie grafov. Jeho použitie síce nie je bezplatné, no na súkromné a neziskové účely je dostupný pod Creative Commons Attribution-NonCommercial 3.0 License, a stiahnuteľný z <http://www.highcharts.com/download>.

Obsahuje:

- Basic line chart
- Area chart
- Column Chart
- Bar chart
- Pie chart
- Scatter and bubble chart
- Dynamic chart
- Combinations
- 3D chart
- Gauges
- Heat map
- Polar chart
- Spiderweb
- Wind rose
- Box plot
- Error bar
- Waterfall
- Funnel chart
- Pyramid chart
- General drawing

Všetky grafy sú prepracované, interaktívne, animované, a dostupné v štyroch rôznych dizajnoch (Default theme, Dark Unica, Sand, Signika, Grid Light).

Inštalácia je extrémne jednoduchá, nakoľko je knižnica vo verzii 4.0.4 vytvorená aj ako Ruby Gem.

Plusy	Mínusy
Veľmi jednoduchá inštalácia	Platená pri komerčnom použití
Výborná úroveň animácie	In-line grafy nie sú stavané na minimalistické zobrazenie
Výborná úroveň interaktivity	
Výborná dokumentácia	
Rozumná cena pri komerčnom použití	
Vysoká dôveryhodnosť na základe silných referencií	
Možnosť zobraziť grafy in-line	

Tabuľka 2. Hodnotenie High Charts

3.15.3 JQUERY SPARKLINES

Služby ponúkané knižnicou jQuery Sparklines sú zamerané na minimalistické zobrazovanie rôznych typov grafov. V našom projekte má veľký potenciál pri zobrazovaní dôležitých výstupov analýzy datasetov už v tabuľkovom náhľade, čím môžeme užívateľom ponúknuť kvalitný prehľad a možnosť porovnania datasetov bez nutnosti otvárania v separátnych oknách.

Grafy sú interaktívne a reagujú na kurzor myši tak, ako sme zvyknutý pri tradičnom zobrazení. Podporuje nasledujúce zobrazenia:

- Basic line chart
- Area chart
- Bar chart
- Pie chart
- Tristate chart
- Box plot
- Pre-computed box plot
- Bullet chart

Na stiahnutie je dostupná na: <http://omnipotent.net/jquery.sparkline>

Plusy	Mínusy
Jednoduchá inštalácia	Platená pri komerčnom použití
Ideálne na in-line zobrazenie grafov	Grafy nie sú stavané na zobrazenie v tradičnej veľkosti
Dobrá úroveň animácie	Najnovšia verzia z Júna 2013
Dobrá úroveň interaktivity	
Dobrá dokumentácia	
Dobrá úroveň dôveryhodnosť na základe referencií (napr. Pekingské letisko)	

Tabuľka 3. Hodnotenie jQuery Sparklines

3.15.4 GOOGLE CHARTS

Ako aj iné služby a produkty od Google, tak aj Java Script-ová knižnica Google Charts ponúka množstvo kvalitných riešení. Inštalácia prebieha prostredníctvom jednoduchého include v headeri HTML rozhrania:

```
<script type="text/javascript" src="https://www.google.com/jsapi"></script>
```

Grafy sú zobrazované vo flat dizajne, s tým, že ich vizuálna kustomizácia je buď náročná, alebo nerealizovateľná. Úroveň dizajnu je aj napriek tomu dobrá, vďaka interaktivite, animáciám a faktom, že sú ľudia naň zvyknutý z iných produktov od spoločnosti Google.

Okrem štandardných grafov, ponúka veľké množstvo pokročilých, a taktiež možnosť tvorby vlastných šablon.

Ponúka nasledovné typy grafov:

- Annotation Charts
- Area Charts
- Bar Charts
- Bubble Charts
- Calendar Charts
- Candlestick Charts
- Column Charts
- Line Charts
- Maps
- Org Charts
- Pie Charts
- Sankey Diagrams
- Scatter Charts
- Stepped Area Charts

- Combo Charts
- Diff Charts
- Gauge Charts
- Geo Charts
- Histograms
- Intervals
- Table Charts
- Timelines
- Tree Map Charts
- Trendlines
- Word TreesNew!

Plusy	Mínusy
Jednoduchá inštalácia	Zle kustomizovateľný vzhľad
Výborná dokumentácia s návodmi a ukážkami	
Výborná úroveň animácie	
Výborná úroveň interaktivity	
Dobrý dizajn	
Rozumné a progresívne spoplatnenie pri komerčnom použití	

Tabuľka 4. Hodnotenie Google Charts

3.15.5 FLOT

Java Script-ová knižnica ponúka základné typy grafov, no je založená na plugin repozitári, kde je potrebné vybrať a nainštalovať všetky knižnice samostatne. Tento prístup môže spôsobiť značný chaos ako v zdrojovom kóde, tak aj v organizácii priečinkov. Dizajn je na podpriemernej úrovni, interakcia s užívateľom je minimálna, a rozsah ponúkaných grafov veľmi základný. Je stiahnuteľný z <http://www.flotcharts.org/>.

Plusy	Mínusy
Zdarma aj na komerčné použitie	Zložitá inštalácia
Výstup je možné uložiť v PNG	Zlá úroveň animácie
	Zlá úroveň interaktivity

Tabuľka 5. Hodnotenie Float

3.15.6 RAPHAËL JS

Knižnica na vykresľovanie grafov, prácu s obrázkami a textom, jednoduchú mapu, a color picker. V základnej verzii ponúka iba štyri typy grafov, ktoré sú síce priemerne interaktívne, no neumožňujú zobrazovať dostatočné množstvo informácií. Grafický dizajn zaostáva, no je založený na SVG W3C štandardoch. Je stiahnuteľná z <http://dmitrybaranovskiy.github.io/raphael/>.

Plusy	Mínusy
Jednoduchá inštalácia	Neponúka nič zaujímavé
Podpora SVG	Priemerná úroveň animácie
	Nízka úroveň interaktivity

Tabuľka 6. Hodnotenie Raphael JS

3.15.7 GAUGE

Podporuje zobrazovanie grafu v tvare polkruhovej elipsy. Je založený na knižnici Raphaël JS. Inštalácia, tak ako aj použitie je jednoduché. Animácia nevyžaduje knižnicu jQuery – stačí Java

Script. Knižnica je použiteľná napríklad pri zobrazovaní vyťaženia servera. Je stiahnuteľná z <http://justgage.com/>.

Plusy	Mínusy
Jednoduchá inštalácia	Neponúka nič zaujímavé
Podpora SVG	Žiadna interaktivita
Dobrá úroveň animácie	
Jednoduché použitie	

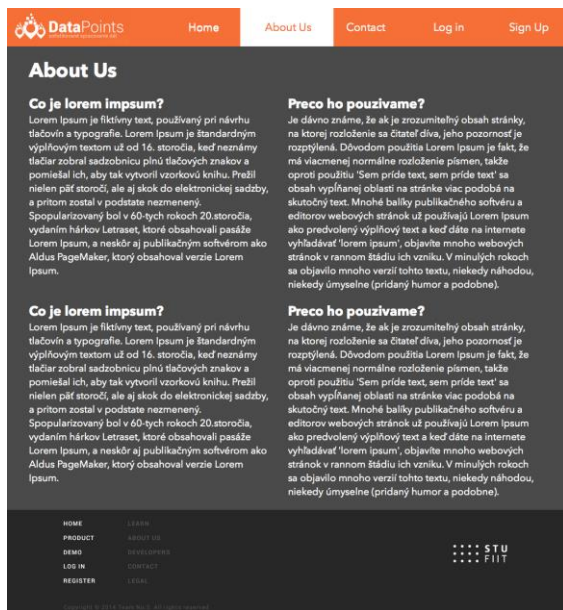
Tabuľka 7. Hodnotenie Gauge

3.15.8 VÝSLEDNÉ HODNOTENIE

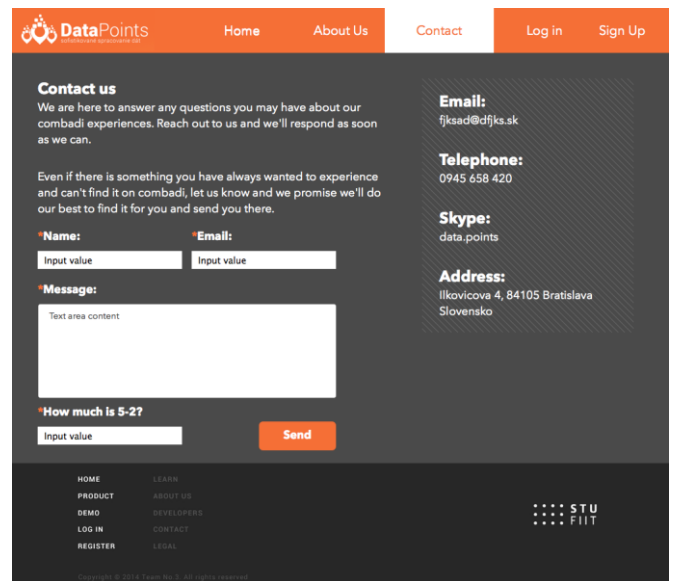
Názov	Zložitosť inštalácie	Použitelnosť	Dokumentácia	Úroveň animácie	Úroveň interaktivity	Dizajn	Spoplatnenie	Výsledok
D3JS	7	7	7	8	9	8	10	8
High Charts	10	9	9	8	9	9	8	8.9
jQuery Sparklines	7	7	8	6	7	7	10	7.4
Google Charts	7	9	10	7	9	7	8	8.1
Flot	3	2	3	3	2	2	10	8.1
Raphael JS	7	1	2	3	3	3	10	4.1
Gauge	7	5	9	8	0	5	10	6.3

Tabuľka 8. Výsledné hodnotenie

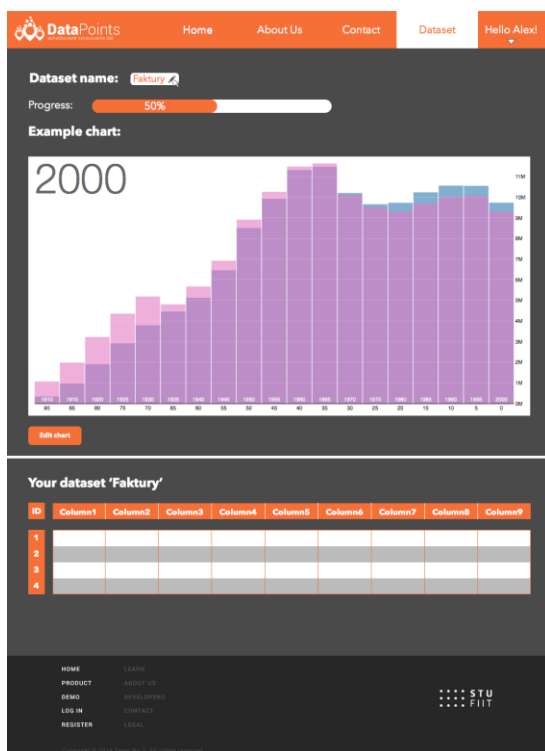
3.16 OBRAZOVKY GUI



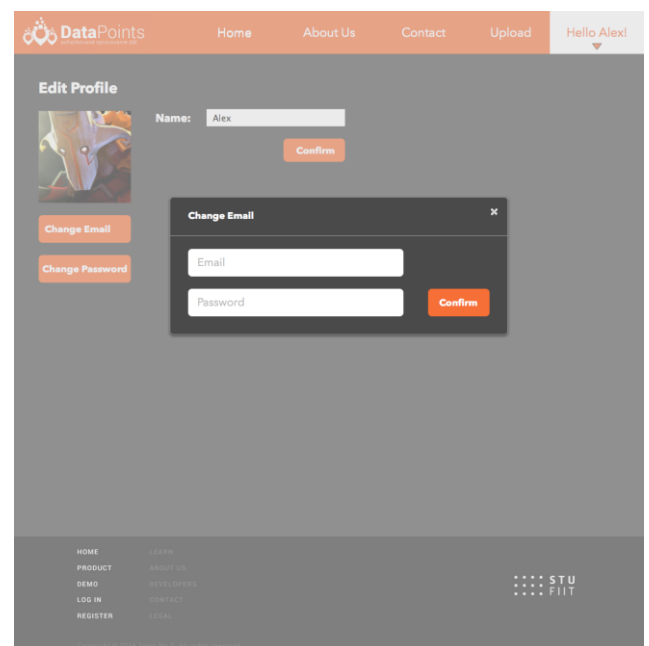
Obrázok 10. O nás



Obrázok 11. Kontakt



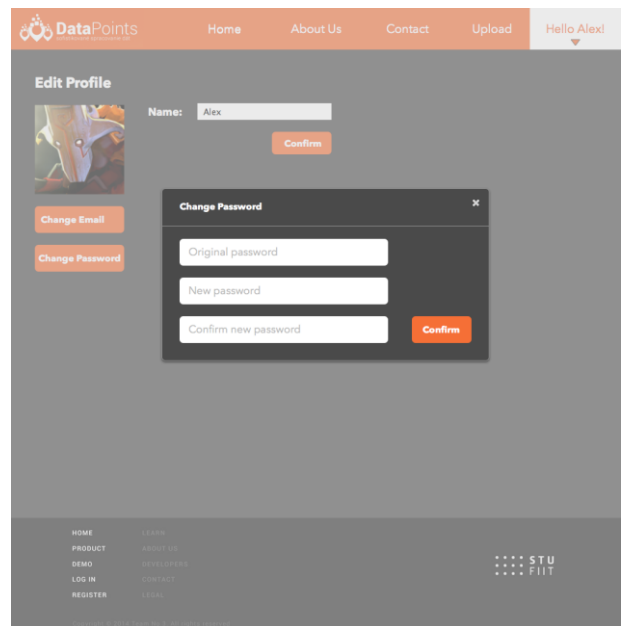
Obrázok 12. Dataset



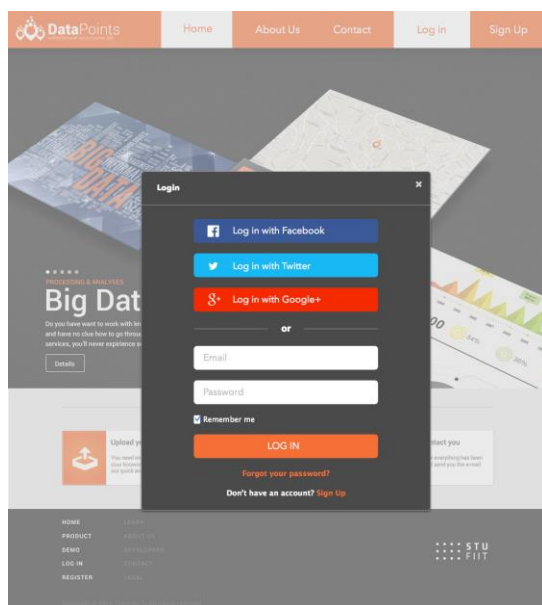
Obrázok 13. Zmena E-mailu



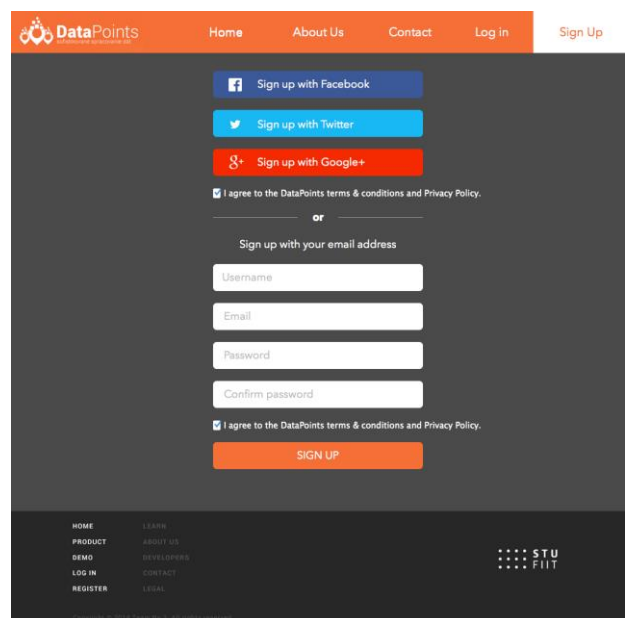
Obrázok 14. Úvodná stránka



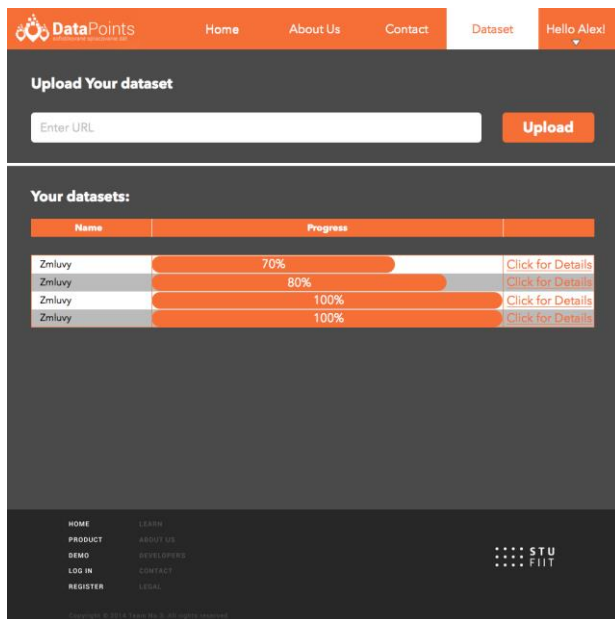
Obrázok 15. Zmena hesla



Obrázok 16. Prihlásenie pomocou tretích strán



Obrázok 17. Registrácia pomocou tretích strán



Obrázok 18. Náhľad datasetu

3.17 PRISPÔSOBENIE GUI POUŽÍVATEĽOM

3.17.1 AKCIE, KTORÉ MÔŽU POUŽÍVATEĽ VYKONÁVAŤ:

I. Stĺpcový diagram

1. Používateľ môže zmeniť osi X a Y na príslušné stĺpce datasetu.
2. Používateľ môže zmeniť farbu stĺpcového diagramu.
3. Používateľ môže zmeniť názov stĺpca datasetu.
4. Používateľ si môže diagram uložiť v podobe obrázka.

II. Geografický diagram

1. Používateľ môže zvoliť, ktorý stĺpec zobrazíť na mape.
2. Používateľ môže zmeniť názov stĺpca datasetu.
3. Používateľ si môže diagram uložiť v podobe obrázka.

III. Koláčový diagram

1. Používateľ môže zvoliť, ktorý stĺpec zobrazíť ako diagram.
2. Používateľ môže zvoliť, ktorého stĺpca dáta sa budú počítať.
3. Používateľ môže meniť farbu pre konkrétnu časť koláčového diagramu.
4. Používateľ môže zmeniť názov stĺpca datasetu.
5. Používateľ si môže diagram uložiť v podobe obrázka

3.17.2 MODEL

Prehodenie stĺpcov a riadkov tabuľky a
Zmena zobrazenia metrik (v stĺpcoch / v riadkoch)

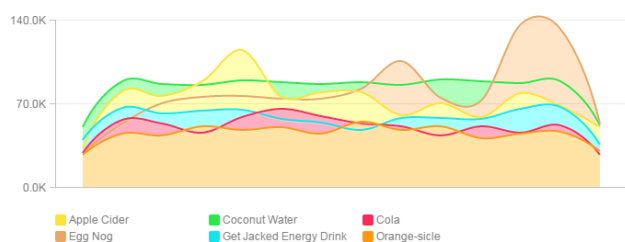
Mktg Channel	Angry	Worried	Upset	Thoughtful	Stressed	Satisfied
Facebook	54.7%	47.9%	38.1%	38.2%	46.3%	44.1%
Instagram	57.5%	65.6%	44.3%	55.9%	39.5%	65.9%
Pinterest	45.1%	40.8%	45.9%	53.1%	47.7%	45.4%
Tumblr	49.9%	50.8%	47.3%	45.7%	49.1%	47.2%
Twitter	38.6%	48.8%	55.5%	52.7%	43.9%	34.1%
Youtube	42.9%	37.4%	53.5%	47.9%	43.7%	53.7%

Mktg Channel	Facebook	Instagram	Pinterest	Tumblr	Twitter	Youtube
Angry	54.7%	57.5%	45.1%	49.9%	38.6%	42.9%
Worried	47.9%	65.6%	40.8%	50.8%	48.8%	37.4%
Upset	38.1%	44.3%	45.9%	47.3%	55.5%	53.5%
Thoughtful	38.2%	55.9%	53.1%	45.7%	52.7%	47.9%
Stressed	46.3%	39.5%	47.7%	49.1%	43.9%	43.7%
Satisfied	44.1%	65.9%	45.4%	47.2%	34.1%	53.7%
Sad	50.7%	53.5%	50.1%	49.4%	43.7%	60.0%
Playful	52.5%	51.1%	52.3%	55.7%	40.0%	38.7%
Optimistic	64.2%	53.1%	46.5%	53.3%	57.2%	56.4%
Happy	57.8%	45.5%	39.1%	62.7%	51.7%	61.9%
Excited	50.7%	55.9%	56.9%	49.3%	41.1%	52.6%
Confused	46.5%	50.1%	47.8%	45.5%	43.6%	54.3%
Confident	34.5%	46.5%	52.5%	43.1%	64.1%	40.9%
Appreciative	58.0%	55.9%	38.9%	48.7%	46.1%	45.4%
Annoyed	49.8%	51.7%	40.8%	47.9%	48.9%	55.3%

Mktg Channel	Facebook	Instagram	Pinterest	Tumblr	Twitter	Youtube
Angry	54.7%	57.5%	45.1%	49.9%	38.6%	42.9%
Worried	47.9%	65.6%	40.8%	50.8%	48.8%	37.4%
Upset	38.1%	44.3%	45.9%	47.3%	55.5%	53.5%
Thoughtful	38.2%	55.9%	53.1%	45.7%	52.7%	47.9%
Stressed	46.3%	39.5%	47.7%	49.1%	43.9%	43.7%
Satisfied	44.1%	65.9%	45.4%	47.2%	34.1%	53.7%
Sad	50.7%	53.5%	50.1%	49.4%	43.7%	60.0%
Playful	52.5%	51.1%	52.3%	55.7%	40.0%	38.7%
Optimistic	64.2%	53.1%	46.5%	53.3%	57.2%	56.4%
Happy	57.8%	45.5%	39.1%	62.7%	51.7%	61.9%
Excited	50.7%	55.9%	56.9%	49.3%	41.1%	52.6%
Confused	46.5%	50.1%	47.8%	45.5%	43.6%	54.3%
Confident	34.5%	46.5%	52.5%	43.1%	64.1%	40.9%
Appreciative	58.0%	55.9%	38.9%	48.7%	46.1%	45.4%
Annoyed	49.8%	51.7%	40.8%	47.9%	48.9%	55.3%

Obrázok 18 – Prehľad tabuliek

Trendy

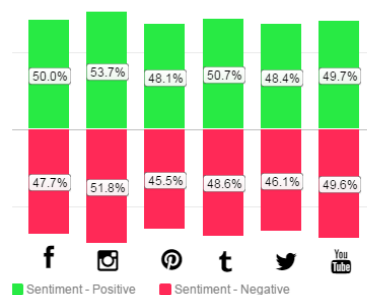


Obrázok 20 – Trendy v datasete



Obrázok 19 – Úvodná stránka ku datasetu

Zobrazenie pozitívnych a negatívnych
hodnot podľa užívateľom zadanej hodnoty



Obrázok 21 – Zobrazenie pozitívnych
a negatívnych hodnôt

Filtrovanie

Číslo	Odkiaľ	Kam	Číslo rozh.	Spoločnosť	Dá
102 701 Bratislava	Zoradiť od A po Z		4860-150/02	SAD Bratislava, a.s.	
102 703 Bratislava	Zoradiť od Z po A		1241-2100-05	SAD Trnava, a.s.	
102 802 Bratislava	Zoradiť podľa farby		4535-150/2002	SAD Bratislava, a.s.	
102 814 Bratislava	Vymazať filter od „Kam“		5292-150/2002	SAD Bratislava, a.s.	
102 814 Bratislava	Filtrovat' podľa farby		211-97/2000	SAD BBDS š. p. Banská Bystrica	
301 701 Bánovce	Filtrovat' podľa farby		375-100/2005	SAD Prievidza, a. s.	
307 701 Prievidza	Filtrovat' podľa farby		1463-2100/05	SAD š. p. Prievidza	
309 701 Trenčín	Filtrovat' podľa farby		659-150/03	SAD Trenčín, a.s.	
309 703 Drietom	Filtrovat' podľa farby		3196-2100/05	SAD Trenčín, a.s.	
403 701 Nitra	Filtrovat' podľa farby		8654-150/02	SAD Nitra, a.s.	
507 703 Turzovka	Filtrovat' podľa farby		1428-2100/05	SAD š. p. Žilina	
502 705 Korňa	Filtrovat' podľa farby		221-565/99	SAD š. p. Žilina	
502 707 Čadca	Filtrovat' podľa farby		211-52/2000	SAD Žilina, a. s.	
502 708 Kľočno	Filtrovat' podľa farby			SAD š. p. Žilina	
503 701 Dolný Kľočno	Filtrovat' podľa farby			SAD š. p. Žilina	
507 702 Námestovo	Filtrovat' podľa farby			SAD š. p. Žilina	
511 801 Žilina	Filtrovat' podľa farby			SAD š. p. Žilina	
601 702 Banská Bystrica	Filtrovat' podľa farby			SAD š. p. Žilina	
601 705 Banská Bystrica	Filtrovat' podľa farby			SAD š. p. Žilina	
601 801 Banská Bystrica	Filtrovat' podľa farby			SAD š. p. Žilina	
601 803 Banská Bystrica	Filtrovat' podľa farby			SAD š. p. Žilina	

Obrázok 22 – Filtrovanie v datasete

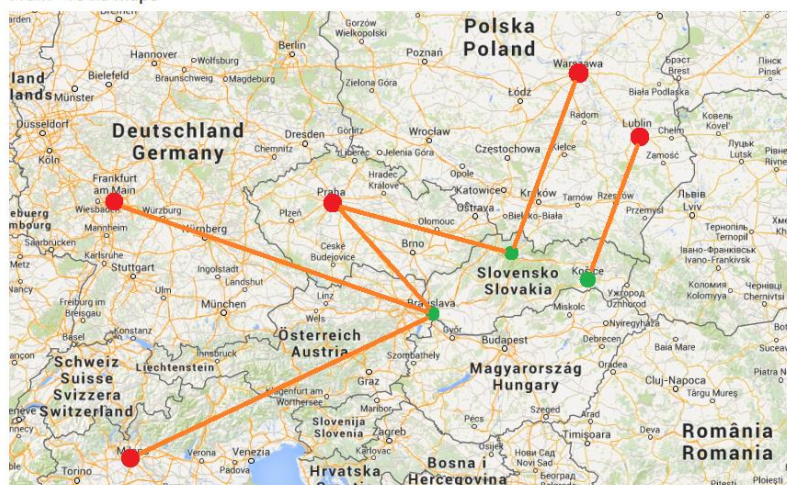
Drag & Drop stĺpcov

Date	Low	High	Open
1/31/2013	27.97	27.76	27.79
1/30/2013	27.76	28.19	28.01
1/29/2013	27.6	28.13	27.82
1/28/2013	27.76	28.23	28.01

20px tolerance

Obrázok 22 – Drag & Drop stĺpcov v datasete

From - To na mape



Obrázok 23 – Zobrazenie geolokácie na mape

4 HÁDANKY V TME

Číslo šprintu: 3

Začiatok šprintu: 23. 10. 2014

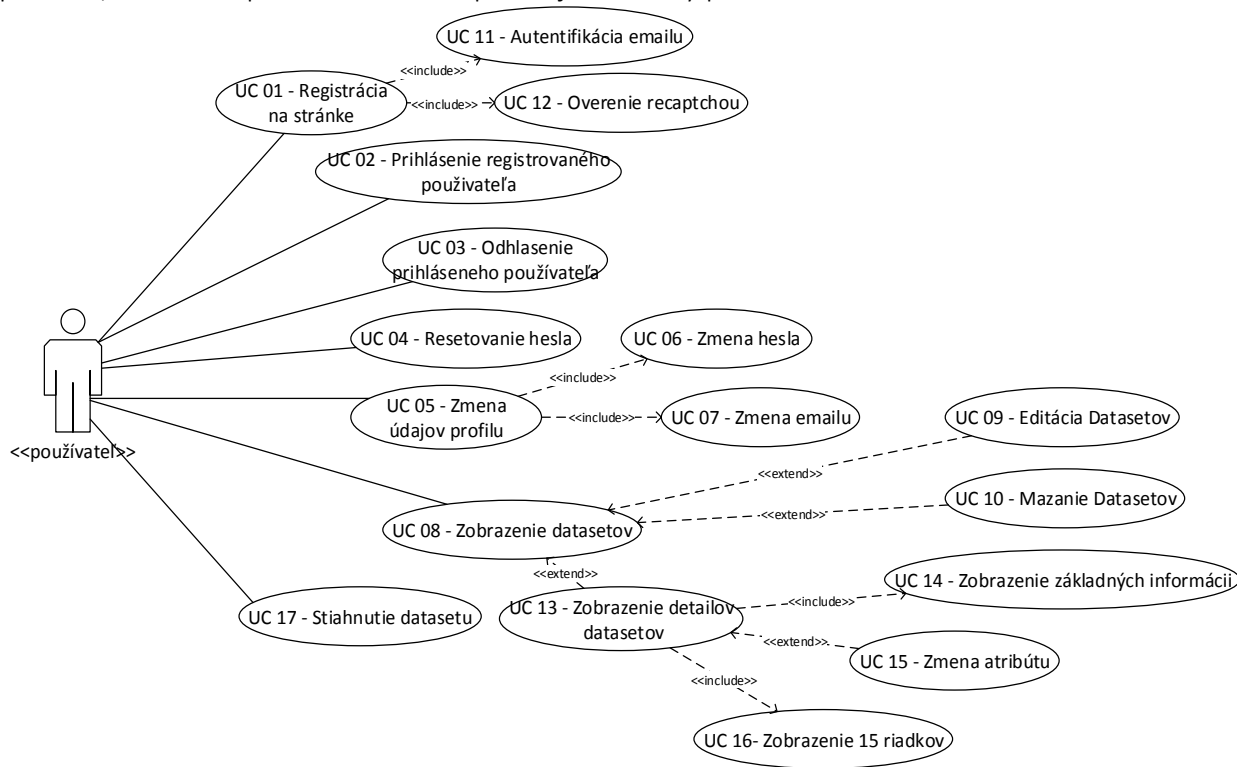
Koniec šprintu: 6. 11. 2014

Príbehy:

- Recaptcha
- Emailová verifikácia pri registrácii
- Password reset
- Refactor profilu
- Stiahnutie datasetu a pridanie do DB
- Chcem vidieť základné textové informácie (atribúty, dátum, veľkosť)
- V zozname datasetov sa zobrazia ich atribúty
- Zobrazíť typy atribútov v zozname
- Používateľ mení typ atribútu
- Vymyslieť 6 funkcií manipulácie s dátami + obrazovky
- Ako používateľ chcem vidieť prvých 15 riadkov datasetu

4.1 PRÍPADY POUŽITIA

V nasledujúcich riadkoch je na obrázku 24 uvedený rozšírený diagram prípadov použitia z prvého šprinu. Diagram bol rozšírený o 7 prípadov použitia. Opis jednotlivých prípadov použitia, ktoré boli pridané v tomto šprinte je uvedený pod obrázkom.



Obrázok 24. Diagram prípadov použitia v 3. Šprinte.

UC 11: Autentifikácia emailu

Pri registrácii bude používateľov email overený pre jeho pravosť a aktívne použitie aktivačným emailom. Po aktivácii bude používateľ schopný prihlásiť sa na stránku.

UC 12: Overenie pomocou captche

Používateľ bude musieť preukázať že nie je stroj prejdením jednoduchého vizuálneho turningovho testu.

UC 13: Zobrazenie detailu datasetu

Pri nahraných datasetoch bude mať používateľ možnosť zobraziť detail datasetu obsahujúci podrobnejšie informácie o datasete.

UC 14: Zobrazenie základných informácií

V detailu datasetu budú zobrazené detailné informácie o zvolenom datasete.

UC 15: Zmena atribútov

Používateľ bude mať v detaile datasetu možnosť upravoť typy atribútov, ktoré sa vyskytujú v datasete.

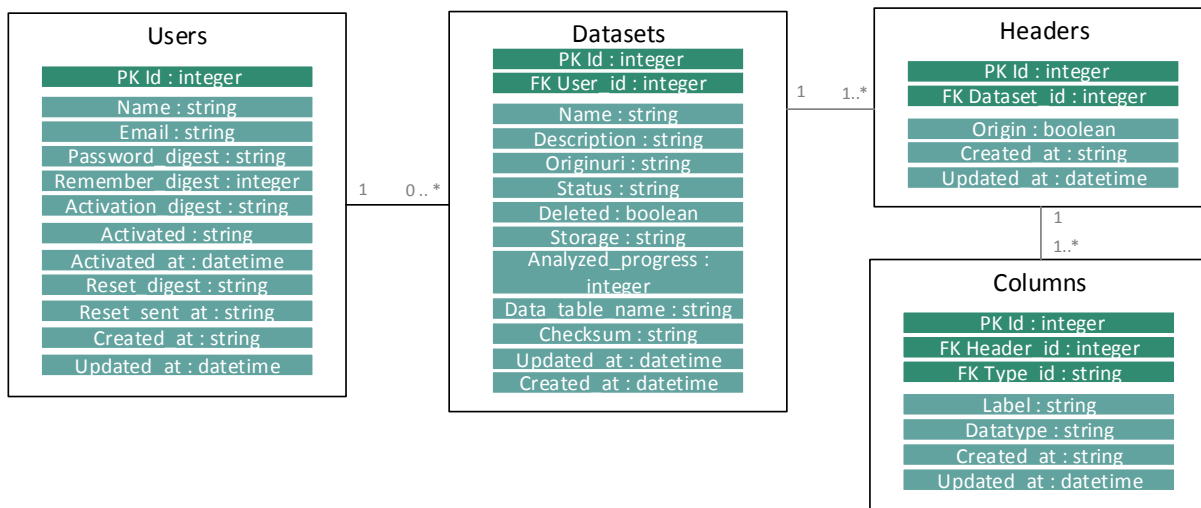
UC 16: Zobrazenie 15 riadkov datasetu

V detaile datasetu bude zobrazená ukážka prvých 15 riadkov datasetu pre odstránenie nutnosti použitia externej aplikácie pre zobrazenie dát.

UC 17: Stiahnutie datasetu

Používateľ bude mať možnosť sťahovať rôzne datasety.

4.2 DÁTOVÝ MODEL



Obrázok 23. Diagram dátového modelu

Pre vyriešenie všetkých úloh bolo potrebné rozšíriť pôvodný dátový model z prvého šprintu. Rozšírenie tohto modelu je vidieť na obrázku číslo 23. Ako je vidieť z obrázku dátový model obsahuje 4 tabuľky, z ktorých dve sú pôvodne, *users* a *datasets*, a dve sú nové *headers* a *columns*. Pôvodná tabuľka *datasets* sa rozšírila o nové atribúty:

- Data_table_name – Atribút slúžiaci na uchovanie názvu genericky vytvorenej tabuľky.
- Originuri – Je premenovaný atribút link, slúžiaci na uchovanie url adresy datasetu.

Novo vytvorená tabuľka *headers* slúži na prepojenie tabuliek *datasets* a *columns*. Jeden *datasets* môže mať viacej hlavičiek a to v takom prípade ak používateľ edituje hlavičku *datasetu*. Riadky môžu patriť práve jednej hlavičke. Atribút *Origin* označuje pôvodnú hlavičku, ktorá bola vytvorená pri procese nahratia dát z datasetu do databázy hodnotou TRUE. Všetky ostatné hlavičky vytvorené počas práce s datasetom budú mať túto hodnotu FALSE.

Tabuľka *columns* slúži na uchovávanie metadát o datasete. Tabuľka *columns* odkazuje na tabuľku *headers* a tabuľku *type*, ktorá je plánovaná v budúcich šprintoch. Tabuľka *columns* obsahuje dva atribúty:

- Datatype – Atribút, v ktorom je zapísaný reálny dátový typ, pod ktorým je uchovávaný stĺpec v databáze. Pre všetky riadky v tabuľke *columns* je to momentálne dátový typ "string".
- Label – Meno stĺpca zobrazované používateľovi v tabuľke. Meno tohto stĺpca sa môže líšiť od mena pod ktorým je stĺpec zapísaný v databáze.

4.3 RECAPTCHA

4.3.1 ŠPECIFIKÁCIA

Pri registrácii používateľa sa zobrazí výzva na vyplnenie captche. Používateľ musí pre úspešnú registráciu prejsť overením captchou.

4.3.2 ANALÝZA

Recaptcha je overenie, či používateľ nie je stroj pokúšajúci sa o prihlásenie sa na stránku. Rozhodli sme sa pre využitie captche od google pre jej jednoduchú implementáciu a jej bezpečnosť.

4.3.3 IMPLEMENTÁCIA

Pre implementáciu captche sme využili gem recaptcha, ktorý sprostredkúva recaptcha API. Recaptcha predstavuje web servis. Naša aplikácia sa overuje voči web servisu vopred vygenerovanými kľúčmi od googlu. Web servis slúži na generovanie a následné overenie captche.

4.3.4 TESTOVANIE

Pri registrácii používateľa sme testovali dva scenáre. Prvým scenárom je registrácia používateľa so správne vyplnenou capchou. Druhým scenárom je prihlásenie používateľa s nesprávne vyplnenou captchou. V oboch prípadoch sme dostali predpokladaný výsledok kedy sa používateľ pri správne vyplnenej capchi úspešne registroval. Pri nesprávnej captchi registrácia neprebehla. Výsledok registrácie sme overovali voči záznamom z tabuľky používateľov.

4.4 EMAILOVÁ VERIFIKÁCIA PRI REGISTRÁCII

4.4.1 ŠPECIFIKÁCIA

Pre úspešné ukončenie registrácie bude nutné potvrdiť link, ktorý bude poslaný na príslušný email uvedený pri registrácii.

4.4.2 ANALÝZA

Pre overenie používateľovho emailu je nutné zaslať mu na emailovú adresu token uložený v linku, ktorý po kliknutí na link bude overený voči tokenu uloženému u nás v aplikácii. Pred aktiváciou je používateľ považovaný za neaktívneho po potvrdení aktivácie sa prepne do aktívneho stavu.

4.4.3 IMPLEMENTÁCIA

Pre potreby overenia autentifikácie sme vytvorili dva nové atribúty a to activation_digest v tabuľke používateľa pre kontrolu pravosti tokenu a atribút activated ktorý je predvolený na hodnotu false. Po vytvorení registrácie sa používateľovi pomocou mail handera odošle mail, ktorý obsahuje odkaz s tokenom. Token sa overí voči activation_digestu daného usera a profil používateľa sa pomocou atribútu activated zmenou na hodnotu true zmení na aktívny a umožní prihlásenie používateľa.

4.4.4 TESTOVANIE

Pri testovaní sme registrovali nových používateľov a overovali sme či používateľ, ktorý nepotvrdil aktivačný email je schopný prihlásiť sa. Tento test dopadol úspešne používateľ, ktorý nepotvrdil registráciu nebol schopný prihlásiť sa. Potom pre rovnakého používateľa sme potvrdili aktivačný email a skúsili sme, či je schopný prihlásiť sa. Užívateľ po potvrdení aktivačného emailu bol schopný prihlásiť sa.

4.5 PASSWORD RESET

4.5.1 ŠPECIFIKÁCIA

Registrovaný používateľ bude mať možnosť resetovania hesla v prípade, že ho zabudol. Po vyresetovaní hesla mu bude zaslaný odkaz, pomocou ktorého si môže zmeniť svoje heslo.

4.5.2 ANALÝZA

Používateľ pri resetovaní hesla zadá email, na ktorý mu bude odoslaný aktivačný odkaz s tokenom. Pomocou tokenu a emailu sa overí, že ide o daného používateľa a umožní sa mu zadať nové heslo.

4.5.3 IMPLEMENTÁCIA

Pre potreby resetu hesla sme pridali nové atribúty pre tabuľku používateľa prvý atribút `reset_digest` slúži na overenie požiadavky o zmenu hesla s tokenom. A druhý atribút `reset_sent_at` slúži na overenie časového limitu pre zmenu hesla. Po vyplnení žiadostí o zmenu hesla sa pomocou email handleru odošle odkaz s tokenom. Po otvorení odkazu a úspešnom overení tokenu ako aj úspešnom overení časového limitu je používateľ presmerovaný na stránku kde je mu umožnené zadať nové heslo a potvrdenie nového hesla.

4.5.4 TESTOVANIE

Do systému sme registrovali používateľa. Následne sme ho odhlásili a požiadali sme o zmenu hesla. Po doručení emailu sme zmenili heslo a vyskúšali sa prihlásiť s novým heslom. Používateľ bol úspešne prihlásený.

4.6 REFACTOR PROFILU

4.6.1 OPIS

Zmena profilu používateľa. Možnosť zmeniť meno, e-mail a heslo. Každá zmena na osobitnej stránke. Zmeny zabezpečené potvrdením aktuálnym heslom.

4.6.2 ANALÝZA

Pôvodný návrh správy profilu (na jednej stránke) bol zamietnutý a bola daná požiadavka na zmenu. Nová správa profilu má obsahovať stránku na každú zmenu z dôvodu dopĺňania ďalších informácií do profilu (Facebook account, Twitter account, atď.)

4.6.3 IMPLEMENTÁCIA

Podľa požiadavky na zmenu som vytvoril ďalšie dve stránky. V hlavnej do hlavnej stránky správy profilu som pridal tlačidlá na zmenu e-mailu a hesla. Po kliknutí na tlačidlo presmeruje

používateľa na ďalšiu stránku kde je formulár so zmenou e-mailu alebo hesla. Oba formuláre obsahujú pole na potvrdenie zmeny aktuálnym heslom.

4.6.4 TESTOVANIE

Zmeny e-mailu a hesla som testoval pre nevyplnené polia, nesprávne heslo, nesprávny tvar hesla, nesprávny tvar e-mailovej adresy. Pre zistenie či zmena údajov úspešne prebehla a či sa nové údaje uložili do databázy som priamo skontroloval dáta v databáze a zmenu hesla som testoval odhlásením a opätovným prihlásením novým heslom.

4.7 STIAHNUTIE DATASETU A PRIDANIE DO DB

4.7.1 VSTUP

- Vloženie odkazu na súbor datasetu používateľom na stránke

4.7.2 VÝSTUP

- CSV Súbor datasetu nahraný na serveri
- Dáta datasetu nahrané v databáze

4.7.3 ANALÝZA

Vývoj našej webovej aplikácie vyžaduje, aby sme boli schopní analyzovať súbory datasetov zo vzdialeného umiestnenia.

4.7.3.1 INTEGRÁCIA EXISTUJÚCEHO SŤAHOVAČA A JEHO VOLANIE Z RUBY

Prvou variantou ako sťahovať dáta na server je využitie štandardných linuxových programov pre sťahovanie, ktoré sťahujú súbory prostredníctvom HTTP/HTTPS protokolu.

4.7.3.2 IMPLEMENTÁCIA VLASTNÉHO SŤAHOVAČA

Druhým spôsobom, ako vyriešiť problém sťahovania je implementácia vlastného sťahovaču priamo v prostredí serverovej časti webovej aplikácie. Ruby poskytuje aplikačné rozhranie nazvané Net::HTTP. Toto rozhranie poskytuje pomerne detailné nastavenia vytváraných dopytov. Umožňuje vytváranie vlastných hlavičiek, HTTPS dopytov, serializáciu na disk, automaticky dekomprimuje GZIP, udržiavanie spojenia či nasleduje presmerovania. Taktiež ponúka pohodlný prístup k odpovediam na dopyty. Nevýhodou tohto prístupu je samozrejme mierne náročnejšia implementácia ako v prvom prípade. Výhodou ale ostáva fakt, že uvedený prístup nevyžaduje žiadnu ďalšiu konfiguráciu na vývojárskych strojoch. Ďalšou obrovskou výhodou je možnosť monitorovania priebežného stavu sťahovania priamo v kóde resp. bežiacej aplikácii.

4.7.3.3 PARSOVANIE CSV SÚBORU

Po uložení súboru na server vznikli dve možnosti následnej analýzy CSV súboru. Prostredníctvom už existujúceho gemu, ktorý dokáže parsovať CSV alebo naprogramovať vlastnú metódu na spracovanie. Výhoda použitia gemu je jednoduchosť použitia ale za cenu obmedzenejších funkcionalít práce s CSV súborom. Naopak vlastná metóda by vyžadovala viac času na implementáciu.

4.7.3.4 NAHRATIE DÁT DO DATBÁZY

Analýza vychádza z faktu, že systém bude musieť ukladať dáta rôzneho formátu, ako napríklad CSV, XML, SQL. Každý z týchto formátov má inú štruktúru a preto je potrebné ich transferovať do jednotného tvaru a následne ich uložiť do príslušajúcej databázy. V nasledujúcich riadkoch analyzujeme rôzne spôsoby ukladania dát do databázy.

CSV je jednoduchý súborový formát pre výmenu tabuľkových dát. Samotné nahranie dát tohto formátu do objektovo-relačnej databázy vyžaduje najprv vytvorenie tabuľky, z príslušajúcimi stĺpcami, ktoré zodpovedajú jednotlivým dátam v CSV súbore. Následne je možné nahratie dát do vytvorenej tabuľky. Vytvorenie tabuľky na základe CSV súboru je možné len manuálne alebo automaticky pomocou skriptu na základe dát v súbore. Nevýhodou tohto riešenia môže byť nedostatočné rozpoznanie jednotlivých typov stĺpcov, čo by malo za následok manuálne opravovanie a kontrolovanie všetkých stĺpcov a ich dátových typov.

Ukladanie dát do databázy je možné dvoma spôsobmi. Prvý spôsob poskytuje možnosť vytvorenia ukladania dát do databázy vo forme dátových typov JSON. Využitie takejto architektúry nám poskytne jednotnú formu dát a to v podobe JSON objektov, ktorých analýza bude prevažne závislá na vyhľadávacom engine, ktorý nie je primárne určený na analyzovanie dát a ich vzťahov medzi sebou. PostgreSQL databáza nám poskytne len malé množstvo funkcií a metód, s ktorými budeme môcť pracovať preto je vhodné túto architektúru prehodnotiť vzhľadom na typ dát, aké naša aplikácia spracuje a na funkcie, ktoré chceme aby poskytovala.

Odpoveďou na prvú navrhovanú architektúru je architektúra dva, ktorá ukladá dáta do štandardných dátových typov. Táto architektúra poskytuje oveľa väčšie možnosti práce s dátami, pretože dáta budú reprezentované štandardnými dátovými typmi. Nevýhodou takejto implementácie je väčšia námaha pri implementovaní takehoto typu architektúry. Dáta, ktoré by sa transformovali z XML alebo CSV súborov do klasickej tabuľky, budú potrebovať validovanie samotnými používateľmi. Celkovo by takáto architektúra poskytla zaujímavé možnosti skúmania vzťahov medzi dátami za pomoci klasického dopytovacieho jazyka SQL.

4.7.4 NÁVRH

Navrhované riešenie je závislé od funkcionality ktorú budeme od výslednej webovej aplikácie požadovať:

- Bude nutné flexibilne vytvárať postupnosť krokov predspracovania datasetu (odstránenie hlavičky z CSV, zmena oddeľovačov)? Bude obsah sťahovaných súborov validný vzhľadom k ich formátu (chýbajúce zátvorky v XML)?

TÍMOVÝ PROJEKT - ANALÝZA IMPLEMENTÁCIE SŤAHOVANIA

- Budeme požadovať, aby sme podporovali sťahovanie pomocou rozličných protokolov?

Po dôkladnom zvážení navrhujeme implementáciu vlastného riešenia pomocou knižnice Net::HTTP nakoľko vyžadujeme precíznu kontrolu nad spôsobom sťahovania.

Následne po stiahnutí súboru na server sme navrhli jeho prečítanie. Použitím CSV gemu, ktorý už základný Ruby on Rails balík obsahuje by sme vytiahli dáta aj hlavičku do určitého objektu a ten následne poslali do metódy na nahranie do databázy.

Samotné nahratie dát do databázy vyžaduje vytvorenie novej tabuľky. Vytvorenie tabuľky sa bude vykonávať genericky podľa hlavičky spracovaného súboru. Dáta sa budú nahrávať poriadkoch a budú sa priamo mapovať na stĺpce databázy.

4.7.5 IMPLEMENTÁCIA

Implementáciu sťahovacieho modulu sme rozdelili do nasledujúcich krokov:

1. Implementácia samotného sťahovaču súborov
2. Implementácia preprocesora datasetov
3. Vloženie atribútov datasetu do tabuľky

Sťahovač v prvom kroku skontroluje, či bol niekedy do databázy uložený identický dataset. Ak nebol, spúšťa sa samotný proces sťahovania. Do databázy je zapísaná informácia o začatí sťahovania formou príznaku. Po stiahnutí je zapísaná adresa kam bol súbor uložený, dátum jeho poslednej modifikácie, kontrolný súčet súboru a zdroj odkiaľ bol stiahnutý. Cieľová adresa kam súbor uložiť je načítavaná z konfiguračného súboru. V tomto stave je sťahovanie ukončené a sme pripravená na procesing.

Implementácia preprocesingu spočívala v prečítaní dát zo súboru do premennej s tým, že sme museli dbať na správne kódovanie. Následne sme premennú ešte dodatočne formátovali, ktorú CSV gem nedokázal spraviť. Po tejto úprave sme použitím CSV gemu vložili dáta s hlavičkami štruktúrované do dynamického poľa, ktoré sme následne poslali do metódy na nahratie do databázy.

Nahratie do databázy sa koná v triede TableFactory. Vytvorením objektu a zavolaním metódy bulider sa začína proces načítania, vytvorenia a nahratia údajov do tabuľky databázy.

Celý proces sa začína nahraťím dát zo súboru, ktorý bol stiahnutý a uložený. CSV súbor sa otvorí a jednotlivé riadky sa nahrajú do viacrozmerneho poľa. V prípade, že súbor je poškodený alebo nejde otvoriť proces sa ukončí a metóda builder vráti hodnotu 1. V prípade správneho nahratia CSV súboru sa za pomoci metódy create_table vytvorí generická tabuľka, kde všetky stĺpce budú typu string. Meno tabuľky je v tvare id_číslo_používateľa:id_číslo_datasetu, čo zaručuje, že každá genericky vytváraná tabuľka bude mať unikátne meno. V nasledujúcich dvoch metódach fill_headers a fill_columns sa naplnia tabuľky columns a headers metadátami o práve vytvorenej tabuľke. Posledným krokom je nahratie samotných dát z CSV súboru do datasetu, na toto slúži metóda fill_storage. Metóda genericky vytvorí novú triedu priamo za behu aplikácie:

```
new_class = Class.new(ActiveRecord::Base) { self.table_name =
name_of_dataset }
cols = new_class.columns.map(&:name)
```

Pomocou tejto triedy sa namapujú mená stĺpcov z databázy do premennej cols, pomocou ktorých môžeme hodnoty jednotlivých stĺpcov z CSV súboru priamo mapovať na prislúchajúce stĺpce novej tabuľky. Výhodou takejto implementácie je nevytváranie modelov alebo dodatočných súborov o tabuľke v priečinkoch aplikácie.

V prípade akýchkoľvek problémov s vytvoreným alebo nahraťím dát do tabuľky sa proces ukladania dát so tabuľky ukončí s chybou 1 a v konzole sa vypíše chybová správa.

Známe chyby:

Predspracovanie dát nedokáže spracovať súbory, ktoré sú oddelené inak než čiarkou. Pri takýchto CSV súboroch sa proces ukladania dát do databázy ukončí neúspešne

4.7.6 TESTOVANIE

Testovali sme pridaním linku na súbor datasetu priamo na front-ende systému a následne sme overovali funkčnosť nahratia všetkých potrebných dát v príslušných tabuľkách databázy.

4.8 CHCEM VIDIEŤ ZÁKLADNÉ TEXTOVÉ INFORMÁCIE (ATRIBÚTY, DÁTUM, VEĽKOSŤ)

4.8.1 ŠPECIFIKÁCIA

V obrazovke detail datasetu sa zobrazia základne textové údaje o datasete.

4.8.2 ANALÝZA

Dáta budú zobrazené v riadkoch pod sebou ako krátky popis zobrazovaného datasetu.

4.8.3 IMPLEMENTÁCIA

Zobrazované dáta sa čerpajú z dvoch tabuliek. Prvou je tabuľka datasetov, ktorá poskytuje dátum, kedy bol dataset vytvorený jeho meno a krátky popis. Druhá tabuľka je tabuľka s aktuálnymi dátami, z ktorej sa vyťahuje počet riadkov datasetu a jeho atribúty. Atribúty sú zobrazené nad príslušnými stĺpcami pri zobrazovaní prvých 15 riadkov z datasetu.

4.8.4 TESTOVANIE

Zobrazovanie sme testovali na testovacích dátach ako na datasete, ktorý bol pridaný cez nahrávanie. V oboch prípadoch zobrazené dáta zodpovedali reálnym dátam.

4.9 POUŽÍVATEĽ MENÍ TYP ATRIBÚTU

4.9.1 ŠPECIFIKÁCIA

V obrazovke detail datasetu bude používateľ schopný zmeniť typ atribútu pre aktuálny atribút z datasetu.

4.9.2 ANALÝZA

Pre uvedenú funkcionality bude potrebné vytvorenie tabuľky s dostupnými typmi atribútov, ktoré sa budú mapovať na aktuálne atribúty z headera pochádzajúceho z originálnych dát.

4.9.3 IMPLEMENTÁCIA

Pre zmenu typu atribútu sme implementovali dva dropdown boxy. V prvom si používateľ zvolí atribút a v druhom mu priradí typ z dostupnej ponuky. Typy atribútov sa momentálne vyberajú len z dvoch staticky nastavených atribútov. Dostupné atribúty sa vyberajú s tabuľky columns kde sa mapujú aktuálne názvy atribútov na typy atribútov.

4.9.4 TESTOVANIE

Zmena typu atribútu bola overená na testovacích dátach z tabuľky columns ako aj na dátach vytvorených z pridaného datasetu. V oboch prípadoch bol typ atribútu zmenený úspešne. Zmena bola overená v tabuľke columns, ktorá mapuje typ atribútu na aktuálny atribút z datasetu.

4.10 AKO POUŽÍVATEĽ CHCEM VIDIEŤ PRVÝCH 15 RIADKOV DATASETU

4.10.1 ŠPECIFIKÁCIA

Používateľovi bude umožnené z obrazovky, na ktorej ma zobrazené datasety umožnené kliknúť na odkaz detail datasetu. V tomto odkaze sa mu okrem iného zobrazí 15 riadok zo zvoleného datasetu.

4.10.2 ANALÝZA

Pre zobrazenie datasetov sa naskytli dve možnosti a to použitie existujúceho gemu alebo ručné zobrazenie údajov z datasetu.

4.10.3 IMPLEMENTÁCIA

Dáta potrebné pre vypísanie prvých 15 riadkov sa vyberajú pomocou záznamu `data_table_name` z tabuľky datasetov, ktorý predstavuje názov tabuľky obsahujúcej dáta. Následne sa pomocou SQL dopytu vyberú všetky riadky z tabuľky overí sa ich počet ak je ich viac ako 15 vypíše sa len prvých 15 ak je ich menej zobrazia sa všetky. Riadky z tabuľky sa zobrazujú do HTML tabuľky.

4.10.4 TESTOVANIE

Zobrazovanie riadkov bolo testované na vytvorenej tabuľke dát ako aj na aktuálne pridanom datasete priamo v aplikácii. Riadky boli zobrazené správne.