

Extrakcia dát z webu

[WebExtraction]

Modul Extraction management

Tím:	č. 16, WebX
Vedúci tímu:	Ivan Srba
Členovia tímu:	Ján Brechtľ, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Akademický rok:	2016/2017
Autor:	Ján Brechtľ, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Verzia číslo:	1.1
Dátum poslednej zmeny:	14.12.2016

1 Moduly systému	2
2 Extraction management	2
2.1 Analýza	2
2.2 Návrh	2
2.3 Implementácia	3
2.4 Testovanie	4

1 Moduly systému

Vychádzajúc z predchádzajúcej kapitoly sa dekomponoval systém na menšie časti (moduly), ktoré sa následnej ešte podrobnejšie rozdeľujú na US (User stories) a tie na úlohy.

Identifikované sú tieto moduly (podrobnejšie info v jednotlivých podkapitolách):

- User management
- Project management
- Browser extension
- Extraction management
- Data management

Jednotlivé moduly postupne prechádzajú 4 základnými procesnými štádiami. Sú dekomponované na jednoduchšie US a tie prechádzajú procesom analýzy a návrhu zväčša priamo na stretnutí, vo fáze prideľovania US do šprintu.

Nasleduje samostatná fáza implementácie v šprinte, počas ktorej sa návrh môže meniť. Aby boli jednotlivé US uznané ako hotové, musí prebehnúť aj úspešné testovanie.

2 Extraction management

2.1 Analýza

V predchádzajúcom module bol opísaný manažment projektov a skriptov. Tie tvoria akúsi predlohu pre extrakciu dát, ale samotná extrakcia sa vykonáva v moduli extrakcií. Obsahuje najmä extrahovanie dát z webu, t.j. crawlovanie, a spúšťač jednotlivých skriptov, t.j. scheduler.

2.2 Návrh

Z výsledku analýzy sme dospeli k návrhu modulu pre extrakciu dát. Hlavnou súčasťou je crawler, ktorý vykoná extrakciu podľa zadaného skriptu. Crawler je univerzálny pre všetky typy skriptov. Výsledok extrakcie bude uložený v databáze spolu s informáciou o správnom dokončení extrakcie.

Samotným prvkom je plánovanie extrakcií. Podľa požiadaviek zákazníka si môže používateľ nastaviť viacero časových spúšťaní skriptov. Na začiatku si používateľ zdefiniuje čas prvého spustenia a interval, teda časové obdobie za ktoré sa extrakcia zopakuje a periódu teda počet intervalov.

Príklad nastavenia extrakcie:

- čas prvého spustenia: 21.12.2016 15:00
- perióda: deň
- interval: 3

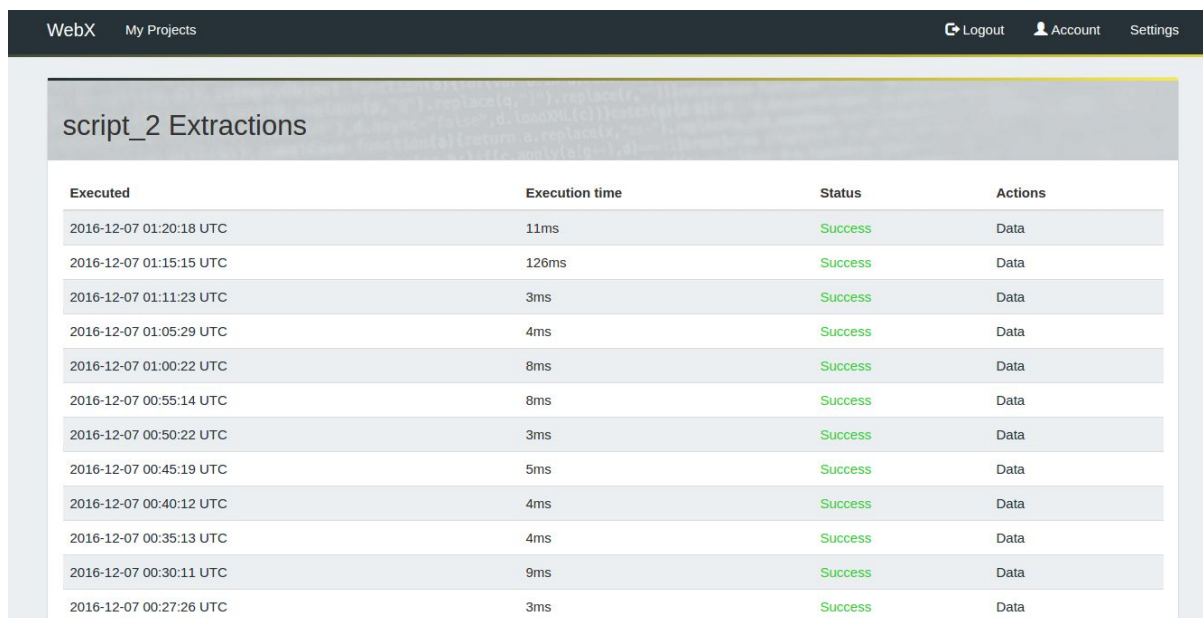
Pri takomto nastavení sa budú extrakcie spúšťať:

- 21.12.2016 15:00; 24.12.2016 15:00; 27.12.2016 15:00; ...

2.3 Implementácia

Vytvorili sme vo webovej aplikácii službu na spustenie konkrétneho skriptu, čím sa stiahnu požadované údaje podľa dátovej schémy projektu z vopred definovanej URL. Využili sme pri tom gem Nokogiri, ktorý stiahne zo zadanej URL HTML dokument a následne naň aplikuje Xpath dopyty.

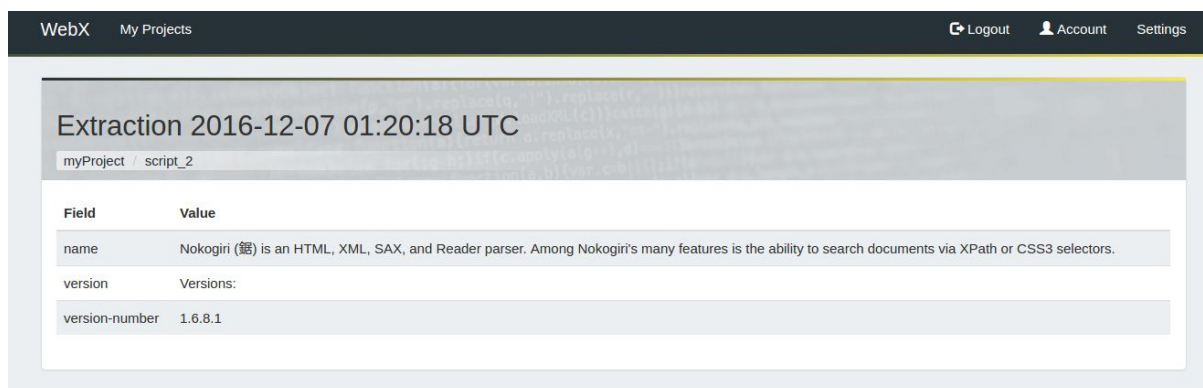
Spustením skriptu sa vytvorí záznam v tabuľke extrakcií (obr. 1), ktorý predstavuje logy zo spúšťania skriptov. Záznam v tejto tabuľke obsahuje čas spustenia, trvanie vykonávania a informáciu o tom, či bola extrakcia úspešná.



Executed	Execution time	Status	Actions
2016-12-07 01:20:18 UTC	11ms	Success	Data
2016-12-07 01:15:15 UTC	126ms	Success	Data
2016-12-07 01:11:23 UTC	3ms	Success	Data
2016-12-07 01:05:29 UTC	4ms	Success	Data
2016-12-07 01:00:22 UTC	8ms	Success	Data
2016-12-07 00:55:14 UTC	8ms	Success	Data
2016-12-07 00:50:22 UTC	3ms	Success	Data
2016-12-07 00:45:19 UTC	5ms	Success	Data
2016-12-07 00:40:12 UTC	4ms	Success	Data
2016-12-07 00:35:13 UTC	4ms	Success	Data
2016-12-07 00:30:11 UTC	9ms	Success	Data
2016-12-07 00:27:26 UTC	3ms	Success	Data

Obr. 1 - Zobrazenie tabuľky extrakcií

Z tejto tabuľky sa dá prekliknúť na zoznam extrahovaných dát, ktorý obsahuje páry - pole dátovej schémy a extrahované dáta (obr. 2).



Field	Value
name	Nokogiri (鋸) is an HTML, XML, SAX, and Reader parser. Among Nokogiri's many features is the ability to search documents via XPath or CSS3 selectors.
version	Versions:
version-number	1.6.8.1

Obr. 2 - Detail extrakcie

Spúšťanie skriptov vykonáva scheduler resque, ktorý sa spúšťa v 5 minútových intervaloch. Pri každom spustení sa vytvorí zoznam skriptov, ktoré sa majú spustiť - podľa

frekvencií skriptu sa vypočíta čas jeho nasledujúceho spustenia. Skript sa pridá do zoznamu, ak čas nasledujúceho spustenia prekročil čas spustenia schedulera.

Skript, ktorý má byť spustený v daný čas sa pridá do inej resque fronty, z ktorej sú skripty následne pridelované crawleru na spracovanie. V tejto fronte vieme zachytiť prípadné zlyhanie crawlera, kedy crawler nezapíše výsledok o neúspešnosti extrakcie.

2.4 Testovanie

Správne vykonávanie skriptov sme testovali pomocou reálnej webstránky rubygems.org, keďže extrahovať dáta z lokálneho HTML súboru sme nevedeli bež vážnejšieho zásahu do produkčného kódu. Uvedomujeme si riziko, že stránka môže byť nedostupná alebo sa môže zmeniť, a teda bude potrebné tento test časom opravovať.

V module je potrebné testovať aj samotný scheduler a úlohy na pozadí (background joby). Na ich testovanie budú použité gemy ako `resque_spec` a `timecop`.