Aplikacja Web Scraper – projekt. Proces ETL

Autorzy dokumentu: Aleksandra Niezgoda,
Piotr Piędel, Kamil Tomsia
Data utworzenia: 26.11.2019
Data ostatniej modyfikacji: 05.01.2019

1. Wstęp

1.1. Cel i przeznaczenie dokumentu oraz opis aplikacji

Dokument ma za zadanie przedstawić specyfikację aplikacji Web Scraper.
Głównym zadaniem aplikacji jest przeprowadzenie procesu ETL – Extract, Transform, Load.
Aplikacja pobiera dane wskazane przez użytkownika za pomocą identyfikatora, następnie przetwarza je i zapisuje do bazy danych. Następnie możliwe jest wyświetlenie tych danych jak również wyeksportowanie ich do plików csv.

2. Konstrukcja aplikacji - architektura

2.1. Język programowania, środowisko uruchomieniowe

- Język w którym tworzona jest aplikacja to JavaScript ES6
- Środowisko uruchomieniowe: Node.js w wersji 12+ https://nodejs.org/en/
- Aplikacja po stronie serwera wykorzystuje framework aplikacji internetowych - Express 4.6.0+ https://expressjs.com/en/starter/installing.html
- MySQL Community Server 8.0.18 https://dev.mysql.com/downloads/mysql/
- Aplikacja po stronie klienta jest utworzona na podstawie framework'u Vue.js https://vuejs.org/

2.2. Biblioteki użyte do stworzenia aplikacji

- apify: 0.16.0 – web scraper, https://apify.com/
- body-parser: 1.19.0, parsowanie url w aplikacji,
- https://www.npmjs.com/package/body-parser
- express: 4.17.1 framework aplikacji internetowych, https://expressjs.com/
- mysql2: 2.0.1, biblioteka umożliwiająca komunikację z bazą danych https://www.npmjs.com/package/mysql2
- jsdoc: 3.6.3 – dokumentacja kodu(funkcje i klasy), https://github.com/jsdoc/jsdoc
- ag-grid-vue: 22.1.1 – biblioteka użyta do tworzenia tabeli https://www.ag-grid.com/

2.3. System zarządzania bazą danych

Systemem zarządzania bazą danych użytym w projekcie jest *MySql* - system do zarządzania relacyjnymi bazami danych.

2.4. Model danych i narzędzia do modelowania bazy danych

2.4.1. Model danych

Do realizacji projektu użyty został relacyjny model danych.

2.4.2. Narzędzia użyte do modelowania bazy danych

https://dbdiagram.io/d

2.5. Minimalne wymagania sprzętowe

- 64-bit wersja systemu Microsoft Windows 10, 8, 7 (SP1
- Minimum 2 GB RAM , zalecane 8 GB RAM
- 2.5 GB wolnego miejsca na dysku HDD, zalecany dysk SSD
- Minimalna rozdzielczość ekranu 1024x768
- Zalecana rozdzielczość ekranu 1920x1080

2.6. Dokumentacja klas oraz funkcji

Dokumentacja kodu aplikacji znajduje się w „backend\docs\index.html” i należy ją otworzyć w przeglądarce internetowej

3. Instrukcja obsługi

3.1. Kroki potrzebne do uruchomienia naszej aplikacji na komputerze.

- Pobranie i instalacja node'a. (w czasie instalacji nie trzeba nic zmieniać, wszystkie domyślne opcje są poprawne)
- Pobranie i instalacja mySql Server, z nazwą użytkownika „*root”* i hasłem „*admin”.*
- Następnie należy stworzyć *schema* o nazwie „*web_scraper_schema”.*
- Utworzyć w niej tabele wklejając i odpalając *query* znajdujące się w pliku *database_schema.sql.*
- Następnie należy wejść do folderu z projektem i otworzyć tam *cmd*.
- Należy wpisać w konsoli(*cmd*) komendę „*npm i”* i poczekać, aż zostaną pobrane wszystkie potrzebne moduły.
- Aplikacja jest już gotowa do uruchomienia! Aby to zrobić wystarczy wpisać w otwartej wcześniej konsoli komendę *„npm start”*. Po chwili powinien zostać wyświetlony komunikat *„Your application is running here: http://localhost:8080”.*
- Wchodzimy na podaną stronę i gotowe! Możemy już korzystać z aplikacji! Udanego scrape'owania!

3.2. Opis funkcjonalności

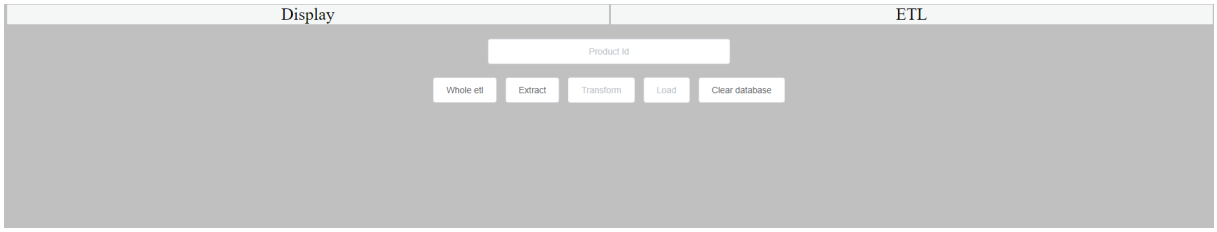
Nasza aplikacja umożliwia użytkownikowi:

- Przeprowadzenie całego procesu *ETL*
- Przeprowadzenie samego kroku *Extract* – pobrania danych na temat produktu z ceneo.pl
- Przeprowadzenie samego kroku *Transform* – obrobienia pierwotnie pobranych danych by były one zdatne do zapisania do bazy danych (pod warunkiem, że wcześniej został wykonany krok *Extract*)
- Przeprowadzenie samego kroku *Load* – zapisania wcześniej obrobionych danych do bazy danych (pod warunkiem, że wcześniej został wykonany krok *Transform*)
- Zobaczenie statystyk po każdej z powyższych operacji (m.in. ilość pobranych poszczególnych danych, ilość transformacji czy też ilość wstawień bądź aktualizacji danych w bazie.
- Usunięcie wszystkich danych z bazy

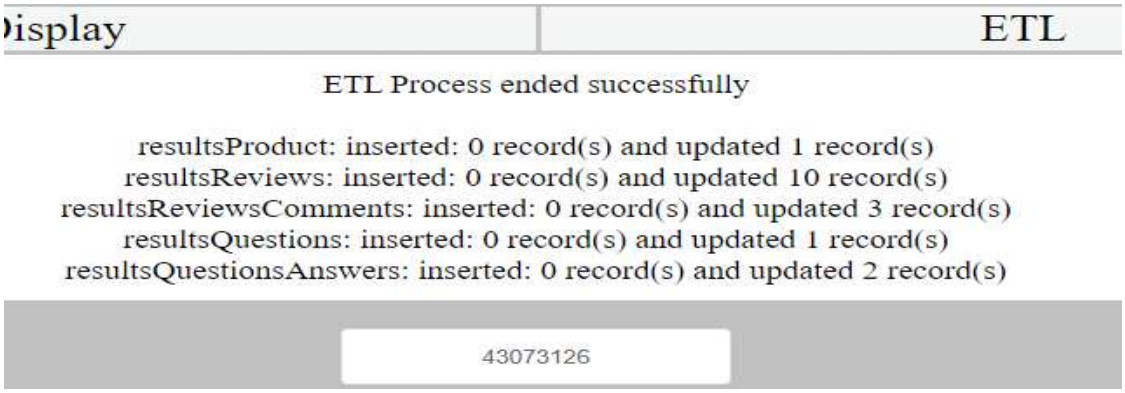
- Wyświetlenie danych wcześniej zapisanych do bazy:
  - Opinii
  - Komentarzy do nich
  - Pytań do produktu
  - Odpowiedzi na nie
- Możliwość sortowania wyświetlanych danych
- Możliwość pobrania danych konkretnego produktu do plików csv

3.3 Opis aplikacji

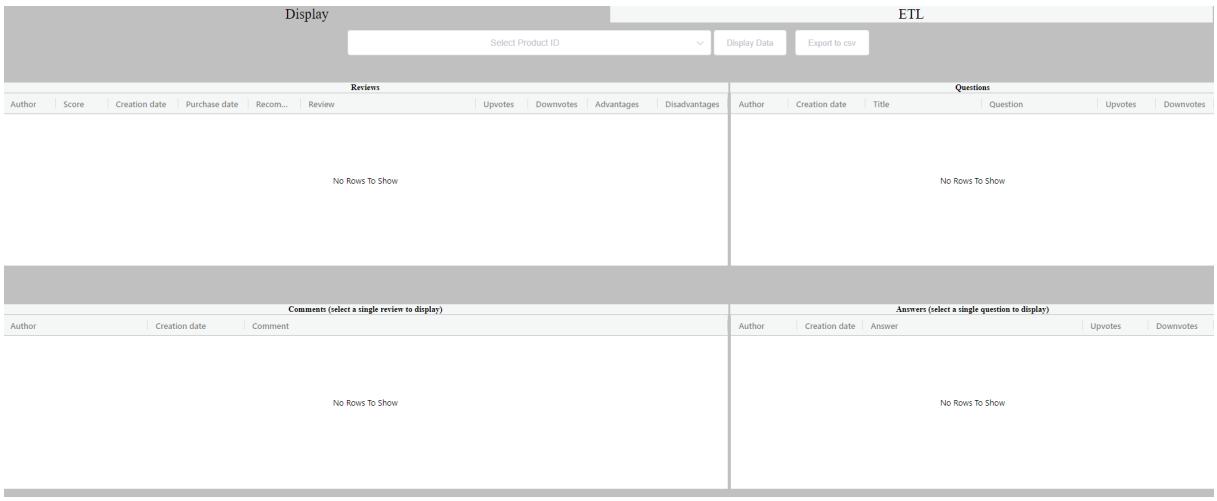
Nasza aplikacja jest aplikacją webową, co oznacza, że do używania jej będzie potrzebna przeglądarka internetowa. Cały interfejs składa się z dwóch stron: ETL - strony do pobierania i przetwarzania danych oraz Display - strony do wyświetlania danych. Zaczniemy od przedstawienia strony do przetwarzania danych. Jej wygląd jest następujący:



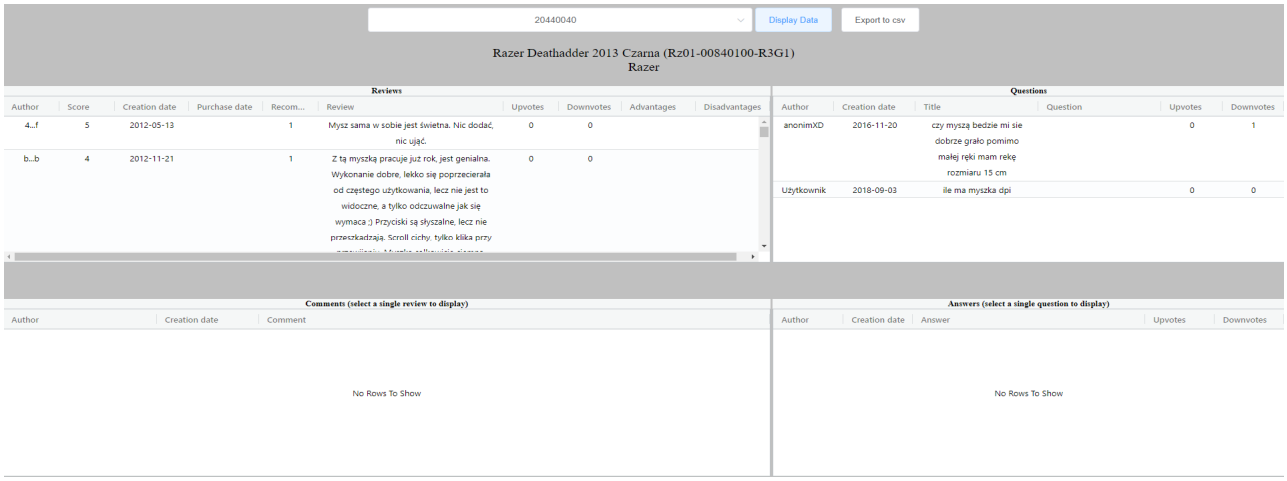
Dwa przyciski na samej górze są częścią paska nawigacji. Poniżej nich na środku jest pole, w którym należy wpisać ID produktu znajdującego się na ceneo.pl. Poniżej mamy 4 przyciski odpowiadające procesowi ETL i 5. przycisk, który służy usunięciu wszystkich danych z bazy danych. Skupmy się najpierw na pierwszych 4 przyciskach. Przycisk *Whole etl,* jak nazwa wskazuje, przeprowadza cały proces ETL, na koniec zaś wyświetla informacje o tym ile danych zostało do bazy dodanych bądź zaktualizowanych.



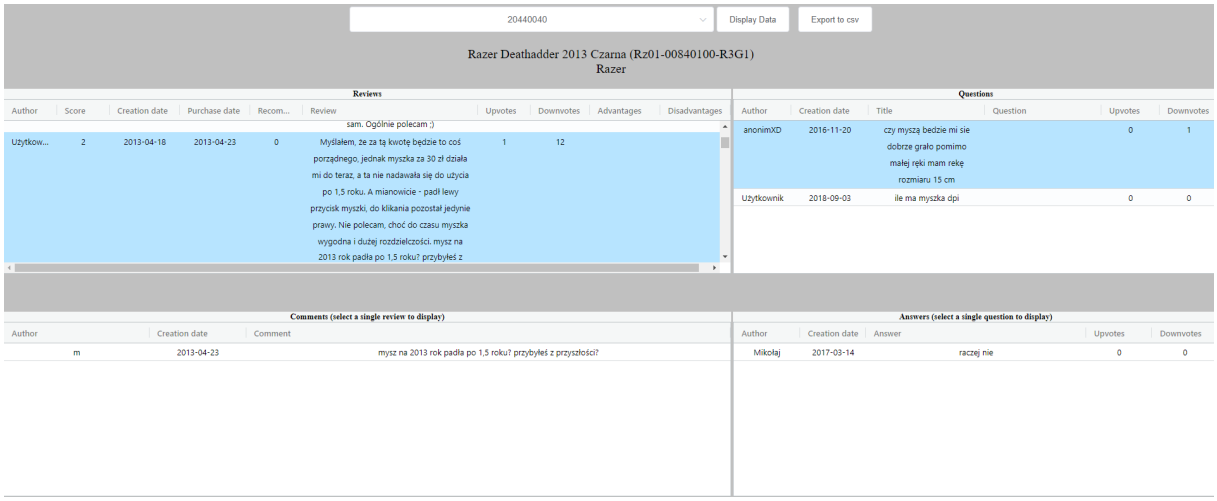
Każdy z pozostałych przycisków związanych z procesem ETL wyświetla statystyki na temat operacji, które musiały zostać wykonane. Na szczególną uwagę zasługują przyciski *Transform* oraz *Load* – mogą być one kliknięte tylko w momencie, gdy ostatnią akcją był krok je poprzedzający – *Extract* w przypadku *Transform* oraz *Transform* w przypadku *Load*. Przycisk *Clear database* usuwa wszystkie dane znajdujące się w bazie danych. By przynieść się do strony umożliwiającej odczytywanie danych, należy wcisnąć przycisk *Display* znajdujący się w lewym górnym rogu. W tym momencie powinno się nam ukazać:



Na górze ekranu mamy pole rozwijane, za pomocą którego możemy wybierać spośród wszystkich produktów znajdujących się w bazie danych. Po wybraniu jednego z produktów, przyciski *Display data* oraz *Export to csv* stają się klikalne. Po wciśnięciu Display data dwie górne tabelki zostaną wypełnione odpowiednimi danymi:



Każda z opinii może posiadać komentarz, który zostanie wyświetlony jeśli wybierzemy daną opinię klikając na nią. Analogicznie jest z pytaniami.



Wszystkie wiersze można posortować po dowolnej kolumnie.

Oprócz wyświetlania danych, możliwe jest wyeksportowanie ich do plików csv za pomocą przycisku *Export to csv.* Po naciśnięciu powinny pobrać nam się na komputer 4 pliki: reviews.csv, questions.csv, answers.csv oraz comments.csv. Może się zdarzyć za pierwszym razem, że nasza przeglądarka zablokuje pobieranie kilku plików na raz ze względów bezpieczeństwa, wtedy trzeba ręcznie dać zezwolenie naszej stronie na takową czynność.