Proces ETL – dokumentacja techniczna

1. Niniejsza aplikacja służy do pobierania z serwisu gumtree.pl ofert pracy pochodzących z kategorii „Programiści, informatyka i Internet” (lub jakiejkolwiek innej kategorii). Najważniejszą funkcjonalnością aplikacji jest proces ETL, przez który pozyskuje ona najnowsze oferty pracy wraz ze związanymi z nimi atrybutami(takie jak lokalizacja, rodzaj pracy, rodzaj umowy itd.) , następnie przekształca te dane, a na końcu zapisuje je w bazie danych.
2. Użyte technologie:
    a. PHP (5.5.38)
    b. MySQL
    c. JavaScript
3. Środowisko potrzebne do uruchomienia aplikacji na serwerze to komputer z zainstalowanym:
    a. Apache HTTP Server
    b. PHP (najlepiej w wersji > 5.4)
    c. System bazodanowy MySQL
4. Biblioteki programistyczne użyte w aplikacji:
    a. Biblioteka PHP „PHP Simple HTML DOM Parser”
    b. Biblioteka JavaScript – jQuery
    c. Biblioteka JavaScript - DataTable
5. Dodatkowe narzędzia przydatne w instalacji aplikacji na serwerze:
    a. phpMyAdmin – do stworzenia tabeli w bazie danych lub jakikolwiek klient bazodanowy isntalowany na komputerze (np. HeidiSQL, MySQL Workbench, itd.)
    b. klient FTP (Total Commander, FileZilla, itd.)
6. W aplikacji wykorzystywane jest wiele modelów danych.
    a. Na początku, w procesie Extract, bazujemy na danych pobranych ze strony – strukturze DOM (Document Object Model). Wybrane elementy z drzewa DOM są zapisywane do pliku CSV.
    b. W procesie Transform dane z pliku CSV są pobierane, zamieniane na tablicę asocjacyjną języka PHP, przekształcane, a na końcu zapisywane do innego pliku CSV.
    c. W ostatnim etapie procesu ETL, czyli procesie Load, dane z pliku CSV pobierane są do tablicy PHP, a następnie wysyłane do bazy danych MySQL za pomocą zapytania „INSERT”.
7. Aplikacja jest napisana w sposób strukturalny.
   W głównym katalogu aplikacji znajduje się między innymi plik confi.php, przechowujący za pomocą stałych, ważne parametry konfiguracyjne. Są to:

URL_TO_GUMTREE	Link do domeny serwisu Gumtree
URL_TO_GUMTREE_CATEGORY	Ścieżka url do interesującej nas kategorii
HOW_MANY_PAGES_TO_EXTRACT	Ilość stron (paginacja), z których dane będą pobrane na etapie Extract
DB_HOST	Host bazy danych MySQL
DB_NAME	Nazwa bazy danych MySQL
DB_USER	Nazwa użytkownika bazy danych MySQL
DB_PASS	Hasło użytkownika MySQL

Najważniejsze funkcje znajdują się w pliku etl/etlFunctions.php. Ich spis oraz funkcjonalności prezentuje poniższa tabela:

Funkcja	Opis
setLastEtlStatus($status);	Funkcja ustawia tzw. Status aplikacji. Robi to porzez zapisanie odpowiedniej informacji do pliku etlStatus.json. Jako parametr przyjmuje jedną z trzech liter: E, T lub L. Litera ta ma oznaczać etap procesu ETL, który obecnie w danej chwili jest wykonywany. Prócz literki zapisywana do pliku .json jest również data momentu wykonania tejże funkcji.
getLastEtlStatus();	Analogicznie do funkcji powyżej, pobiera status ostatniego etapu procesu ETL, zwracając w postaci tablicy jego literę oraz datę wykonania.
mysqlConnect();	Służy do połączenia z bazą danych. Nie przyjmuje żadnych parametrów, gdyż korzysta ze stałych globalnych, zadeklarowanych w pliku config.php. Zwraca obiekt funkcji mysqli_connect().
extractStep();	Funkcja etapu Extract
transformStep();	Funkcja etapu Transform.
loadStep();	Funkcja etapu Load.