WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Usługi Sieciowe w Biznesie

Procesy ETL z wykorzystaniem Apache Airflow oraz ich wizualizacja w Apache Superset

Autor:

Adrian Szmyd
163914

Rzeszów 2022

Spis treści

1 Wstęp

W ostatnich latach firmy gromadzą coraz więcej danych. Aby je przetwarzać potrzebna jest do tego odpowiednia technologia. W tym wypadku jest nią Apache Airflow. Za wizualizację danych przetworzonych przez Airflow odpowiedzialny będzie Apache Superset, za bazę/hurtownię danych - PostgreSQL, a za konteneryzację (wirtualizację na poziomie systemu operacyjnego) oraz utworzenie serwera - Docker.

2 Oprogramowanie

2.1 Apache Airflow - ETL

ETL (Extract, Transform and Load) - trójfazowy proces, w którym dane są najpierw pobierane, później transformowane, a następnie wprowadzane do bazy danych.

Apache Airflow pozwala na zorganizowanie przepływu danych dzięki DAGom (Directed Acyclic Graphs). Same dane zostają pobierane z bazy danych PostgreSQL, a później zapisywane do hurtowni danych po ich przetworzeniu. W Apache Airflow w pierwszej kolejności użytkownik określa zależności oraz zadania do wykonania z wykorzystaniem języka programowania Python, a następnie Airflow zajmuje się ich zorganizowaniem (uruchomieniem w odpowiedniej kolejności, DAGi) oraz wykonaniem. W rezultacie otrzymujemy pożądane przez nas tabele z danymi, które możemy później zwizualizować w Apache Superset.

2.2 Apache Superset - Wizualizacja

Apache Superset - aplikacja pozwalająca na przegląd danych oraz ich wizualizację - np. tworzenie m.in. paneli (dashboardów) z wieloma funkcjami (różne typy wykresów, mapy itp.). Wykorzystuje on ponadto pamięć podręczną do pobierania danych co skutkuje krótszym czasem odpowiedzi od aplikacji. Potrafi on także radzić sobie z bardzo dużą ilością danych (skala petabajtów).

2.3 Docker - Konteneryzacja

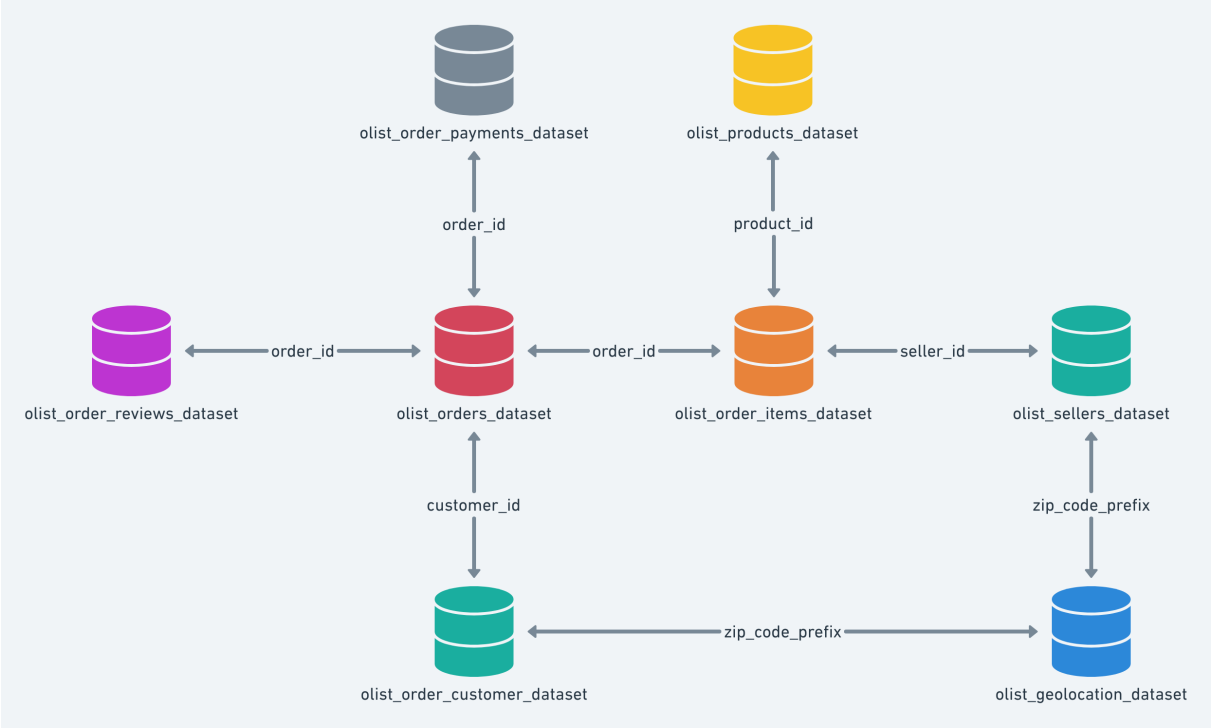
Docker - oprogramowanie wykorzystywane do wirtualizacji na poziomie systemu operacyjnego (tzw. konteneryzacji), dzięki czemu można uruchamiać raz stworzone i skonteneryzowane oprogramowanie na wielu urządzeniach. Innymi słowy - pozwala ono umieścić oprogramowanie, biblioteki oraz pliki konfiguracyjne w łatwym do przeniesienia kontenerze. Każdy kontener jest odizolowany od innych, lecz mogą komunikować się ze sobą przez odpowiednie kanały. Dzięki wykorzystywaniu tego samego jądra systemu operacyjnego, kontenery wymagają mniejszej ilości zasobów niż maszyny wirtualne.

2.4 PostgreSQL - Baza/Hurtownia Danych

PostgreSQL - system zarządzania relacyjnymi bazami danych. W bazie danych PostgreSQL przechowywane będą tabele z danymi oraz później w hurtowni danych przechowywane będą tabele utworzone przez procesy ETL.

3 Dane

Dane wykorzystane w celu pokazania działania procesów ETL z wykorzystaniem Apache Airflow pobrane zostały ze strony internetowej `https://www.kaggle.com/datasets`. Dotyczą one informacji o 100 tys. zamówień z lat 2016-2018 złożonych w Brazylii. Pozwalają one na przeglądanie informacji o zamówieniach w wielu aspektach: od statusu zamówień, cen, wartości płatności i ceny za transport, po lokalizację klienta, atrybuty produktu i recenzje napisane przez klientów. Znajduje się tu także zbiór danych geolokalizacyjnych, który łączy brazylijskie kody pocztowe ze współrzędnymi geograficznymi (`https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce`). Na poniższym rysunku (Rys. 1) przedstawione są relacje między tabelami.

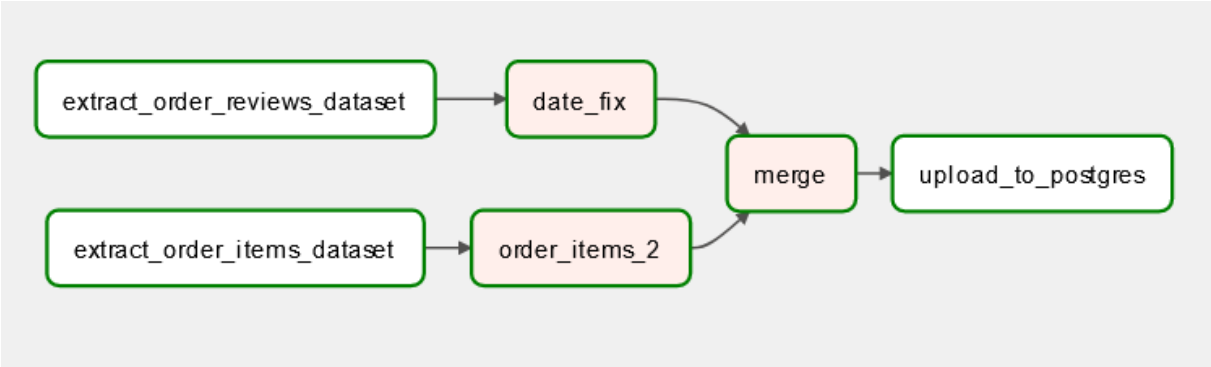


Rysunek 1: Schemat zbioru danych

4 DAGi

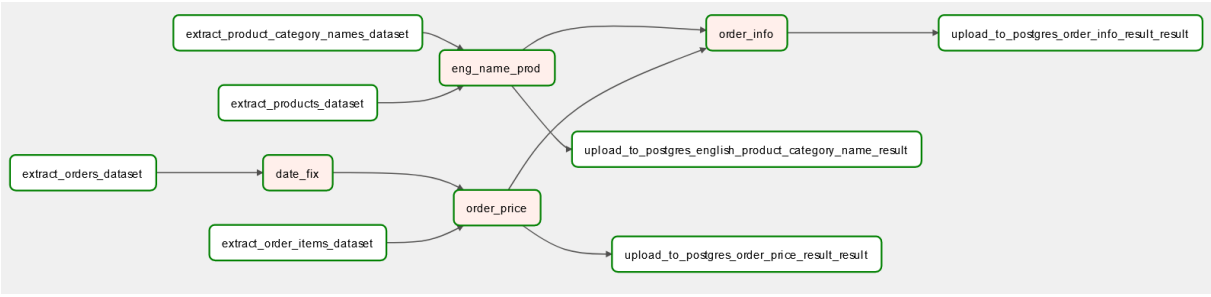
Za pomocą Apache Airflow utworzone zostały DAGi, są to skierowane acykliczne grafy pokazujące zadania, przez które przechodzą dane w celu ich przetworzenia. Dane po przetworzeniu wysyłane są do hurtowni danych w PostgreSQL. Wizualna prezentacja DAGów i relacji między procesami w nich zawartych są automatycznie generowane przez Apache Airflow.

DAG z zamówionymi przedmiotami i ich recenzjami (Rys. 2) pobiera dane z tabel *Recenzje zamówień* oraz *Zamówione przedmioty*. Naprawia daty w *Recenzjach zamówień* oraz łączy otrzymane dane do jednej tabeli, dzięki czemu otrzymujemy tabelę z recenzjami i zamówionymi przedmiotami.



Rysunek 2: DAG z zamówionymi przedmiotami i ich recenzjami

Poniższy DAG (Rys. 3) z informacjami dot. zamówień wykorzystuje informacje z tabel *Zamówienia, Przedmioty w zamówieniu, Produkty oraz Angielskie nazwy kategorii produktów* łącząc je w tabele z większą ilością informacji (ewentualnie zmiana nazw kategorii produktów z portugalskich na angielskie). Zwraca on tabelę z zamówieniami i ich ceną wraz z ceną transportu, tabelę produktów z poprawionymi nazwami kategorii oraz ogólne informacje dotyczące zamówień.



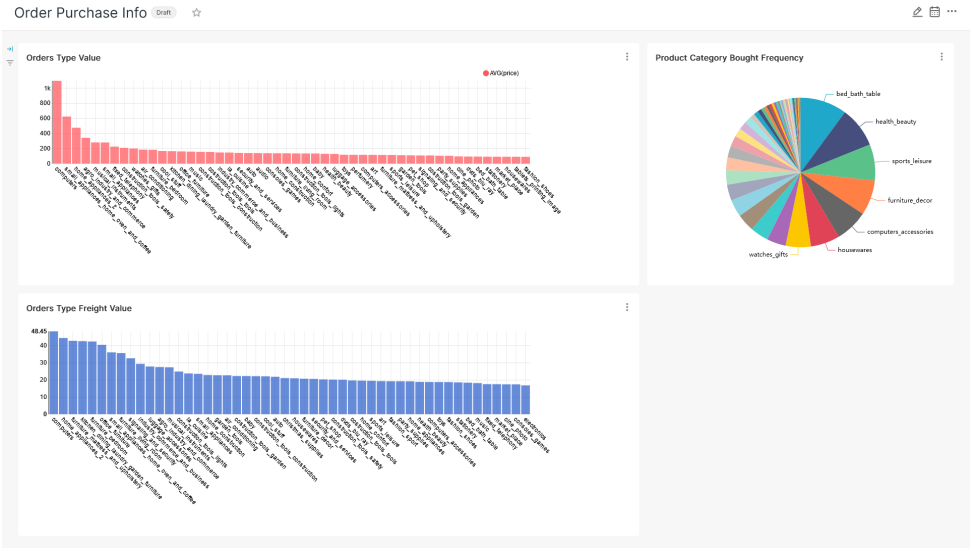
Rysunek 3: DAG z informacjami dot. zamówień

5 Wizualizacje w Apache Superset

W celu zwizualizowania otrzymanych tabel Apache Superset łączy się z hurtownią danych PostgreSQL. Aby wykonać wizualizację należy utworzyć nowy panel, wybrać dane, które ma on przedstawiać, wybrać typ wykresu oraz jakie kolumny z danej tabeli mają zostać zwizualizowane na danym wykresie. Istnieje ogromna ilość możliwych konfiguracji oraz opcji dostosowania paneli, jest wiele różnych typów wykresów możliwych do wykorzystania, dane do wykresów można przedstawić na wiele sposobów w obrębie jednego typu wykresu.

5.1 Panel kategorii zamówień

Na tym panelu pokazane są średnie wartości zamówień produktów, średnia wartość transportu oraz częstość zakupów z danych kategorii.



Rysunek 4: Panel kategorii zamówień

5.2 Panel recenzji

Panel ten porównuje ilości recenzji o danej ocenie z ilością recenzji z dołączonym komentarzem o danej ocenie.

Rysunek 5: Panel recenzji

6 Podsumowanie

W tym projekcie pokazane zostało zastosowanie Apache Airflow, PostgreSQL oraz Apache Superset w praktyce. Apache Airflow pozwala na przetwarzanie danych oraz wizualną prezentację tegoż procesu. PostgreSQL umożliwia przechowywanie dużych ilości danych oraz umożliwia do nich prosty dostęp. Dzięki Apache Superset możliwe jest zaprezentowanie rezultatów wykonanej przez Aiflow pracy na wiele różnych sposobów.