**UNIWERSYTET RZESZOWSKI**

**Wydział: Matematyczno-przyrodniczy**

**Kierunek: Informatyka**

**Hurtownie Danych.**

Hurtownia dotycząca sprzedaży książek w księgarni.

**Wykonał**

Adrian Szuptarski

Informatyka UR

Rok IV

RZESZÓW 2017.01.25

Spis treści

[1. Opis programu 1](#_Toc473121027)

[2. Wstęp 1](#_Toc473121028)

[3. Instrukcja obsługi 3](#_Toc473121029)

# Opis programu

Opracowana aplikacja służy do eksploracji hurtowni danych dotyczącej sprzedaży książek w poszczególnych księgarniach. Korzystając z biblioteki jar „weka” w wersji 3\_8 oraz bazy danych w formacie Postgress, która łączy się do programu za pomocą konektora PostgreSQL w wersji 9.4.1207 aplikacja jest w stanie na załączonej bazie danych wyliczyć statystyki, agregować dane oraz oceniać jakość klasyfikatora

Aplikacja łącząc się z baza danych „Ksiegarnie” korzysta z portu 5432, nazwy użytkownika „postgresql” oraz hasła „szupek”.

Okienkowy interfejs ułatwia komunikację z użytkownikiem, który ma do dyspozycji 3 zakładki:

# Wstęp

Hurtownia danych (ang. data warehouse) – rodzaj bazy danych, która jest zorganizowana i zoptymalizowana pod kątem pewnego wycinka rzeczywistości.

Hurtownia danych jest wyższym szczeblem abstrakcji niż zwykła relacyjna baza danych (choć do jej tworzenia używane są także podobne technologie). W skład hurtowni wchodzą zbiory danych zorientowanych tematycznie (np. hurtownia danych klientów). Dane te często pochodzą z wielu źródeł, są one zintegrowane i przeznaczone wyłącznie do odczytu.

W praktyce hurtownie są bazami danych integrującymi dane z wszystkich pozostałych systemów bazodanowych w przedsiębiorstwie. Ta integracja polega na cyklicznym zasilaniu hurtowni danymi systemów produkcyjnych (może być tych baz lub systemów dużo i mogą być rozproszone).

Architektura bazy hurtowni jest zorientowana na optymalizację szybkości wyszukiwania i jak najefektywniejszą analizę zawartości. Stąd bywa, że hurtownie danych nie są realizowane za pomocą relacyjnych baz danych, gdyż takie bazy ustępują szybkością innym rozwiązaniom.

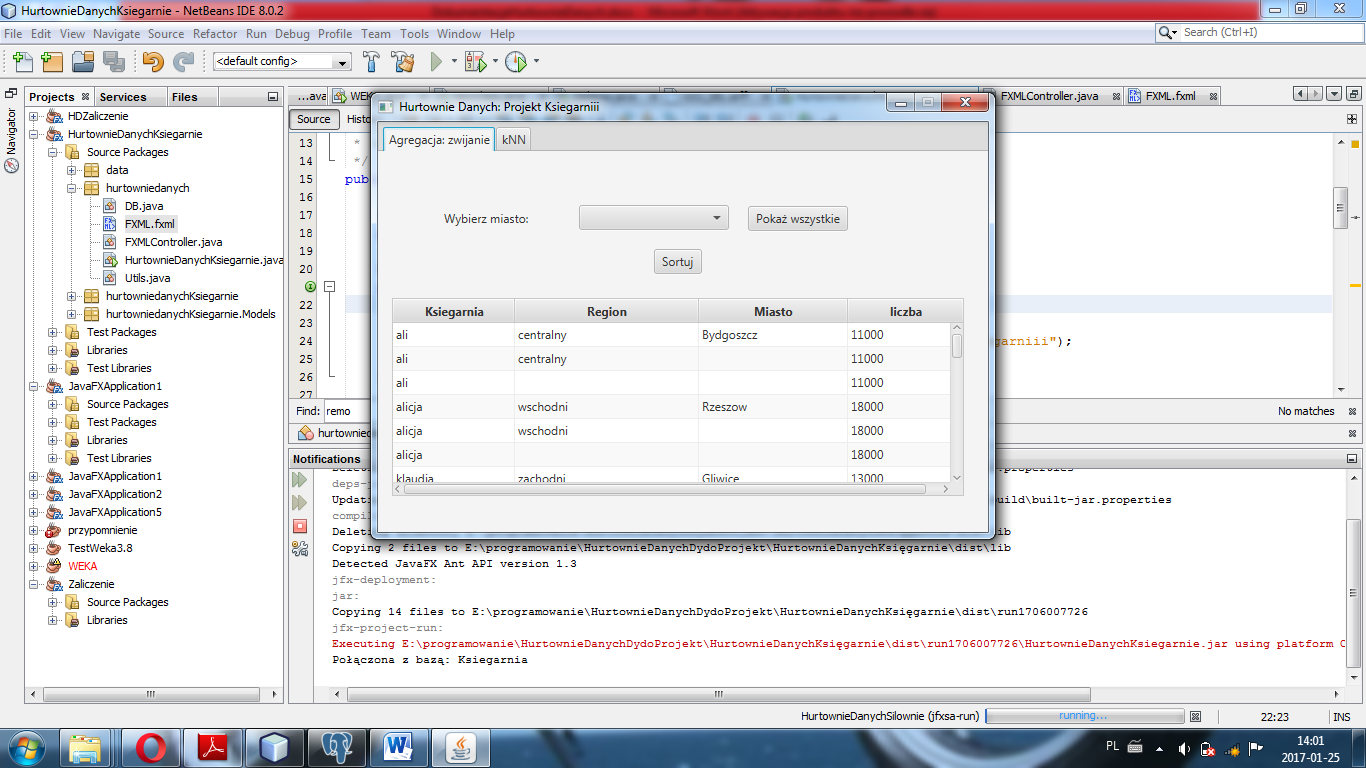
W praktyce w ramach architektury hurtowni wyróżniany jest poziom danych detalicznych oraz warstwa agregatów/kostek tematycznych.

Użytkownicy końcowi hurtowni, czyli najczęściej zarząd firmy, korzystają z danych hurtowni poprzez różne systemy wyszukiwania danych (np. Online Analytical Processing (OLAP)).

Hurtownia danych stanowi zatem rozbudowaną bazę danych, przechowującą olbrzymią ilość danych zbieranych w czasie. Dodatkowo zakłada się, że przeprowadzane na danych operacje mają charakter analityczny (wyszukiwanie i analiza). Nie stosuje się więc typowych transakcji. Ze względu na ilość i tematykę danych przeprowadzane analizy mogą polegać na szukaniu trendów, zależności, wzorców itp. Stosowana tu eksploracja danych (ang. data mining) wyszukuje ogólne formy wiedzy z olbrzymiej ilości danych. Przeprowadzane na hurtowniach danych wyszukiwania mają najczęściej charakter wielowymiarowy – nie ograniczają się bowiem tylko do jednej tabeli, lecz korzystają z wielu relacji. Przechowywane dane w hurtowni są tematycznie spójne (dotyczą konkretnego problemu, lub instytucji – np. szpitala) oraz zintegrowane, co najczęściej cechuje się centralizacją przechowywania danych (wszelkie dane są przechowywane w jednym miejscu). Istnieją również zawężone tematycznie hurtownie danych (np. do jednego oddziału szpitala, lub kliniki) nazywane minihurtowniami danych (z ang. data mart). Minihurtownie mogą istnieć jako oddzielne jednostki lub stanowić element hurtowni danych.

# Instrukcja obsługi

Po uruchomieni aplikacji, jako pierwsze wyświetla nam się okno obsługujące agregacje danych. Tabele umożliwiają sortowanie i wyświetlanie danych w zależności od miejscowości, regionu ilości sprzedanych książek. Wyświetlany na dole combo box pozwala na zwijanie danych (ograniczenie wyświetlanych danych tylko do jednego producenta)



Trzecia zakładka skrywa klasyfikator, dzięki któremu użytkownik może uzyskać ocenę jakości klasyfikacji dla tabeli testowej. Klasyfikacja wykonywana jest przez klasyfikator KNN (k najbliższych sąsiadów). Aplikacja zezwala na wybranie sposobu liczenia odległości (metryki) oraz ustawienie liczby sąsiadów. Niska dokładność klasyfikatora spowodowane jest wprowadzonymi do bazy losowymi wartościami, gdzie klasyfikator nie może znaleźć odpowiednio dobrej zależności między atrybutami.

