|  |  |  |
| --- | --- | --- |
| Indeks | 260338 |  |
| Grupa | E | |

W miarę możliwości/potrzeby proszę o dodatkowy komentarz do diagramu oraz opisów – w szczególności, warto zaznaczyć zrealizowane dodatkowe elementy, czy skomentować przyjętą interpretację pewnych elementów składowych.

Celem zadania jest przygotowanie modelu danych, korzystając z danych z bazy AdventureWorks, do analizy sprzedaży produktów na przełomie dni (miesięcy, kwartałów, półroczy oraz lat) przez różnych sprzedawców, w różnych terytoriach działalności firmy i dla różnych klientów. Do analizy chcielibyśmy także dołączyć zewnętrzne dane zebrane w systemie webowym firmy dotyczące ocen produktów.

# Zadanie 1

Jako rozwiązanie zadania prześlij wypełnione tabele.

## Uproszczony domenowy słownik danych:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Plik | Nazwa atrybutu | Typ wysokopoziomowy | Opis |
| 1 | rating.csv | id | liczbowy | Wewnętrzny identyfikator recenzji, unikalny dla poszczególnych recenzji |
| 2 | rating.csv | productid | liczbowy | Identyfikator produktu, unikalny dla poszczególnego produktu |
| 3 | rating.csv | ip | tekstowy | IP urządzenia, z którego została dodana recenzja |
| 4 | rating.csv | date | tekstowy | Data wystawienia recenzji w formacie MM/DD/YYYY |
| 5 | rating.csv | ratingWebsite | tekstowy | Ocena strony internetowej w skali 0-10 wystawiona przez użytkownika |
| 6 | rating.csv | ratingShipping | tekstowy | Ocena dostawy w skali 0-10 wystawiona przez użytkownika |
| 7 | rating.csv | ratingProduct | tekstowy | Ocena kupionego produktu w skali 0-10 wystawiona przez użytkownika |
| 8 | rating.csv | ratingOverall | tekstowy | Ocena ogólna w skali 0-10 wystawiona przez użytkownika |
| 91 | rating.csv | gender | tekstowy | Płeć użytkownika |
| 10 | rating.csv | email | tekstowy | Email użytkownika |
| 11 | rating.csv | Job | tekstowy | Stanowisko pracy użytkownika |
| 12 | rating.csv | postCode | tekstowy | Kod pocztowy adresu użytkownika |
| 13 | rating.csv | Source | tekstowy | Żródło, z którego użytkownik dowiedział się o stronie |
| 14 | rating.csv | didPurchase | tekstowy | Informacja czy użytkownik zakupił produkt, 0 – nie zakupił, TRUE – zakupił |
| 15 | rating.csv | didRecommend | liczbowy | Informacja czy użytkownik poleca produkt, 0 – nie poleca, 1 - poleca |
| 16 | rating.csv | isUsefull | liczbowy | Liczbowa miara przydatności recenzji w skali 0-25 |
| 17 | rating.csv | userAgent | tekstowy | Informacje na temat urządzenia z którego wystawiono recenzję (np. Nazwa przeglądarki, system operacyjny, wersja systemu) |
| 18 | location.csv | Column1 | tekstowy | Nazwa miasta, z którego odnotowano aktywność urządzenia |
| 19 | location.csv | Column2 | tekstowy | Nazwa państwa, z którego odnotowano aktywność urządzenia |
| 20 | location.csv | Column3 | tekstowy | Kod państwa, z którego odnotowano aktywność urządzenia |
| 21 | location.csv | Column4 | tekstowy | Nazwa kontynentu, z którego odnotowano aktywność urządzenia |
| 22 | location.csv | Column5 | tekstowy | Adres IP urządzenia, unikalny |

## Ocena jakości źródła danych

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Plik | Liczba wierszy | Liczba atrybutów | Rozmiar | Zmienność | Ziarno | Ocena |
| 1 | rating.csv | 1000 | 17 | 218 KB,  ~0.2 na wiersz | Wraz z procesem recenzji  ~3 dziennie | 1 recenzja | Łatwo dostępny, dobrze uporządkowana struktura pliku CSV, kodowanie UTF-8 |
| 2 | location.csv | 501 | 5 | 23.2  ~0.04 na wiersz | Wraz z dodaniem kolejnego IP (brak informacji o cykliczności) | 1 ip | Słabo uporządkowana struktura (nazwy atrybutów jako 1 rekord zamiast w nagłówkach) pliku CSV, kodowanie UTF-8, błędy w danych (literówki) |

## Ocena jakości atrybutów:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Plik | Nazwa atr. | Skala pomiarowa  (typ atrybutu) | Typ danych | Liczba wart. o zduplik. danych | Unikalnych wartości | Pustych wartości | Ocena (jakości/zastosowania) |
| 1 | rating.csv | id | Porządkowy | Integer | 0 | 1000 | 0 | Dane nie wymagają dod. przetwarzania; niska wartość zastosowania – pozwala unikalnie zidentyfikować recenzje |
| 2 | rating.csv | productid | Porządkowy | Integer | 255 | 33 | 1 | Dane wymagające dodatkowego obsłużenia wartości pustych, niska wartość zastosowania – pozwala unikalnie zidentyfikować produkty |
| 3 | rating.csv | ip | Porządkowy | Text | 293 | 98 | 1 | Dane niewymagające przetwarzania, niosące małą informację bezpośrednią ale pozwalające określić w analizie, jak aktywne jest dane urządzenie (np. liczba recenzji) |
| 4 | rating.csv | date | Interwałowy | Text | 268 | 70 | 1 | Dane wymagające zmiany typu danych na datę oraz obsłużenia wartości pustych, niosące dużą wartość informacji (dane można szeroko analizować pod kątem czasu) |
| 5 | rating.csv | ratingWebsite | Ilorazowy | Text | 101 | 0 | 0 | Dane wymagające zmiany typu danych na liczbowy, o wysokiej jakości – strona opiera się na recenzjach więc analizowanie ich będzie przydatne |
| 6 | rating.csv | ratingShipping | Ilorazowy | Text | 101 | 0 | 0 | Dane wymagające zmiany typu danych na liczbowy, o wysokiej jakości – strona opiera się na recenzjach więc analizowanie ich będzie przydatne |
| 7 | rating.csv | ratingProduct | Ilorazowy | Text | 90 | 1 | 1 | Dane wymagające zmiany typu danych na liczbowy oraz zaokrąglenia wartości wzorem innych atrybutów ilościowych opisujących oceny, o wysokiej jakości – strona opiera się na recenzjach więc analizowanie ich będzie przydatne |
| 8 | rating.csv | ratingOverall | Ilorazowy | Text | 74 | 10 | 1 | Dane wymagające zmiany typu danych na liczbowy, o wysokiej jakości – strona opiera się na recenzjach więc analizowanie ich będzie przydatne |
| 91 | rating.csv | gender | Porządkowy | Text | 2 | 1 | 1 | Dane nie wymagają dod. przetwarzania; niska wartość zastosowania – pozwalają zidentyfikować płeć użytkownika, ale nie umożliwi to bardzo szerokiego zakresu analizy danych |
| 10 | rating.csv | email | Porządkowy | Text | 0 | 1000 | 1 | Dane nie wymagają dod. przetwarzania; niska wartość zastosowania – pozwalają zidentyfikować użytkownika, ale nie umożliwi to bardzo szerokiego zakresu analizy danych |
| 11 | rating.csv | Job | Porządkowy | Text | 59 | 24 | 97 | Dane wymagają obsłużenia wartości pustych; niska wartość zastosowania – pozwalają zidentyfikować zawód użytkownika, ale nie umożliwi to bardzo szerokiego zakresu analizy |
| 12 | rating.csv | postCode | Porządkowy | Text | 0 | 0 | 1000 | Atrybut zbędny, mający puste wartości dla wszystkich rekordów |
| 13 | rating.csv | Source | Porządkowy | Text | 7 | 1 | 1 | Dane wymagają obsłużenia wartości pustych; w praktyce wysoka wartość zastosowania – strony tego typu często analizują, skąd użytkownicy dowiedzieli się o ich stronie |
| 14 | rating.csv | didPurchase | Porządkowy | Text | 2 | 0 | 292 | Dane wymagające zmiany typu danych na True/False, obsługi wartości pustych i ustandaryzowania – albo wartości 0/1 albo True/False, wysoka wartość zastosowania, strona opiera się na recenzjach więc jest to kluczowa informacja |
| 15 | rating.csv | didRecommend | Porządkowy | Integer | 3 | 0 | 544 | Dane wymagające zmiany typu danych na True/False, obsługi wartości pustych i ustandaryzowania – albo wartości 0/1 albo True/False, wysoka wartość zastosowania, strona opiera się na recenzjach więc jest to kluczowa informacja |
| 16 | rating.csv | isUsefull | Porządkowy | Integer | 27 | 0 | 801 | Dane wymagające obsługi wartości pustych I może zmodyfikowania skali (0-25 to troche na wyrost), wysoka wartość zastosowania - strona opiera się na recenzjach więc jest to kluczowa informacja |
| 17 | rating.csv | userAgent | Porządkowy | Text | 304 | 116 | 0 | Dane zbędne w analizie danych, nie wnoszące wartościowych zastosowań |
| 18 | location.csv | Column1 | Porządkowy | Text | 83 | 290 | 0 | Dane wymagające sformatowania nagłówka (pierwszy rekord powinien być nagłówkiem) oraz weryfikacji pod kątem poprawności danych (literówki). Dane niosą wysoką wartość zastosowania, zazwyczaj można szeroko analizować dane pod kątem lokalizacji |
| 19 | location.csv | Column2 | Porządkowy | Text | 11 | 0 | 0 | Dane wymagające sformatowania nagłówka (pierwszy rekord powinien być nagłówkiem) oraz weryfikacji pod kątem poprawności danych (literówki). Dane niosą wysoką wartość zastosowania, zazwyczaj można szeroko analizować dane pod kątem lokalizacji |
| 20 | location.csv | Column3 | Porządkowy | Text | 6 | 0 | 0 | Dane wymagające sformatowania nagłówka (pierwszy rekord powinien być nagłówkiem. Dane niosą wysoką wartość zastosowania, zazwyczaj można szeroko analizować dane pod kątem lokalizacji |
| 21 | location.csv | Column4 | Porządkowy | Text | 2 | 0 | 0 | Dane wymagające sformatowania nagłówka (pierwszy rekord powinien być nagłówkiem). Dane niosą wysoką wartość zastosowania, zazwyczaj można szeroko analizować dane pod kątem lokalizacji |
| 22 | location.csv | Column5 | Porządkowy | Text | 0 | 500 | 0 | Dane niewymagające przetwarzania, niosące małą informację bezpośrednią ale w kontekście posiadania danych o recenzjach z poprzedniego pliku będzie to kluczowy atrybut łączący tabelę, samych zastosowań w analizie jest mało |

\*Przykładowo ciąg (1,2,3,4,4,5, ,4, ,5) ma 2 wartości zduplikowane (liczba wart. o zduplik. danych), 3 wartości unikalne i 2 puste wartości

## Konkluzje dot. zadaniA 1

Obydwa pliki (rating.csv i location.csv) zaimportowaliśmy w ten sam sposób, ale domyślnie plik location.csv nie załadował wartości z pierwszego rekordu do nagłówków, więc nie modyfikowaliśmy tego, a nazwy kolumn pozostawiliśmy domyślne (Column1, Column2 itd.), co ma swoje odzwierciedlenie w tabelkach. Przez „obsługę wartości pustych) w drugiej tabelce rozumiemy nadanie tym wartościom odpowiednich nazw, na przykład adekwatnych do atrybutów, którym odpowiadają lub ogólnych – typu „Not Given”.

# Zadanie 2

Zmodyfikuj model danych z listy zadań 4 (dot. sprzedaży), aby uwzględnić dane dot. oceny produktów (patrz zadanie 4 z tej listy zadań). Wyjściowy model danych (z listy zadań 4) przedstawiony został na diagramie (w treści zadania), naturalnie jest to model nieco rozszerzony, gdyż zawiera informacje o indywidualnym kliencie zlecającym dane zamówienie (nie musisz rozszerzać modelu o ten element):

Jako rozwiązanie przedstaw odpowiednie opisy, zapytania, zrzut/prezentacje modelu danych oraz krótki komentarz/opis dotyczący przyjętej przez Ciebie metodologii realizacji tego modelu.

## Opis użytkownika

Przykładowym użytkownikiem tego systemu BI może być np. nowozatrudniony specjalista ds. marketingu i sprzedaży, chcący wstępnie przeanalizować obecną sytuację marketingową w firmie.

## Zapytania OLAP (słownie) + możliwe zastosowanie

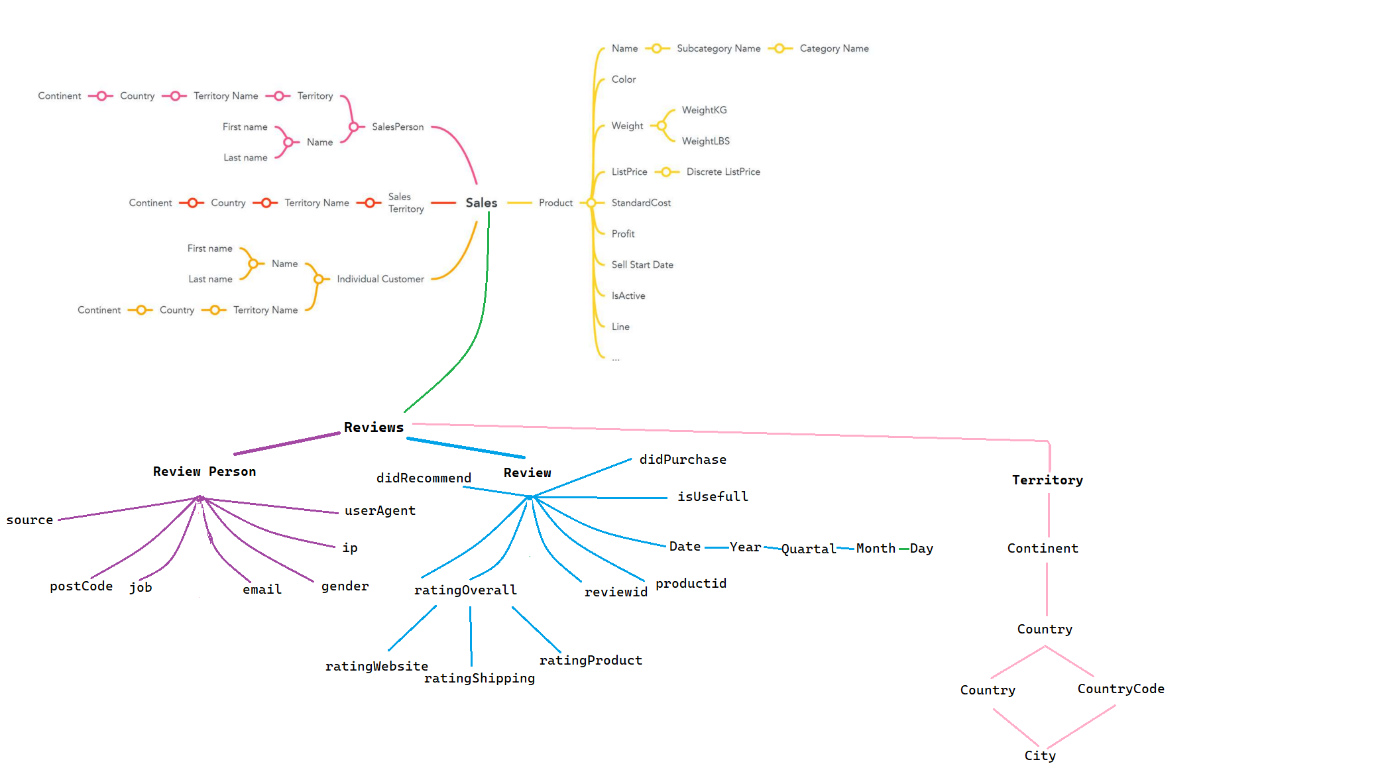
Pytania:

1. Czy płeć użytkownika ma wpływ na wystawioną ocenę strony?
2. Z jakiego kraju serwis odnotował najmniej wystawionych recenzji w ostatnim kwartale?
3. Czy w Stanach Zjednoczonych jakość dostawy jest zależna od miasta, w którym była realizowana?
4. Jaka jest średnia ocena ogólna użytkowników, którzy wystawili więcej niż jedną recenzję?
5. W którym miesiącu roku 2011 zostało utworzonych najwięcej recenzji?

Zastosowania:

1. Specjalista może ocenić, czy strona nie jest zbyt nastawiona (np. wizualnie) na jedną konkretną płeć, co mogłoby zawężać grupę docelową strony
2. Specjalista mógłby ocenić, w którym kraju strona wymaga dodatkowych nakładów finansowych na reklamę lub współprace reklamowe
3. Specjalista może ocenić, czy usługi świadczone w konkretnym regionie odbiegają od standardów świadczonych przez firmę i gdzie należy te standardy poprawić
4. Specjalista może ocenić, jaki jest potencjalny próg ogólnej oceny użytkownika, by ten chciał ponownie skorzystać ze strony.
5. Specjalista może ocenić w którym okresie użytkownicy byli najbardziej aktywni, a następnie przeanalizować tego przyczynę, by przykładowo w tym samym miesiącu przyszłego roku być już na to gotowym.

## diagram – Wysokopoziomowy model Danych



## OPIS – Wysokopoziomowy model Danych

Do realizacji zadania przyjęliśmy koncepcję modelu płatka śniegu (w miarę możliwości). Główną tabelą rozszerzającą model jest tabela Reviews. Zawiera ona wszystkie informacje z pliku review.csv. Według nas najlepszym rozwiązaniem byłoby podzielenie tej tabeli na trzy inne tabele, reprezentujące inne typy informacji o recenzjach. Pierwsza tabela – Review Person zawierająca dane o użytkowniku wystawiającym recenzję (jego zawód, płeć, email itd.). Druga tabela – Review zawierająca bezpośrednie informacje o samej recenzji (reviewid, isUsefull itd.). Ponadto w tej tabeli sugerowalibyśmy utworzenie hierarchii daty w sposób przedstawiony na diagramie. Trzecia tabela – Territory zawierające dane lokalizacyjne dla każdej recenzji, najlepiej zdefiniowana w sposób hierarchiczny zaczynając od kontynentu, a kończąc na mieście.

## Konkluzje dot. zadaniA 2

Przedstawiony przez nas model danych jest naszym zdaniem najbardziej intuicyjny i najlepiej odseparowywujący od siebie niezależne dane, a łączący dane mające wspólne czynniki (jak na przykład oceny). Reprezentacja modelu jest też dosyć konceptualna, gdyż na przykład w przypadku ratingOverall rozdzieliliśmy to na ratingShipping, ratingWebsite oraz ratingProduct, pomimo iż nie jest to zdefiniowana hierarchia. Zdecydowaliśmy się jednak na taką wizualizację, aby pozostać konsekwentnym względem innych części modelu (np. w tabeli SalesPerson na Name składa się First Name oraz Last Name).

### PODSUMOWANIE – Konkluzje do całej listy zadań

Cała lista zadań polegała na bardziej konceptualnym zrozumieniu modelu danych, aniżeli implementacji go w programie. Pomimo tego nie było łatwe zidentyfikowanie niektórych własności atrybutów lub tabel. Ostatecznie utworzyliśmy jasny i przejrzysty model, który powinien spełniać podstawowe założenia i umożliwić przykładowemu użytkownikowi analizę danych.