Projekt końcowy - Hurtownie Danych

Zadanie projektowe polega na:

1. **Zaprojektowaniu hurtowni danych** dla gromadzenia danych o zatrudnieniu na rynku pracy.
2. **Implementacji procesów ETL** ładujących do hurtowni danych informacje pochodzące z pliku tekstowego CSV (https://moodle2.e-wsb.pl/mod/resource/view.php?id=2598241).
3. **Implementacji analitycznych raportów biznesowych** służących do wizualizacji zestawień danych pochodzących z hurtowni danych.

Opis kroków zadania:

1. Budowana hurtownia danych powinna gromadzić dane o liczbie osób pracujących w odniesieniu do: branży, lokalizacji, płci i przedziału wiekowego (zgodnie z plikiem https://moodle2.e-wsb.pl/mod/resource/view.php?id=2598241). Należy rozpoznać, które z wymienionych atrybutów maja charakter **wymiarów**, a które reprezentują **miarę** (miary). Następnie należy opracować **schemat gwiazdy**, obejmujący niezbędne **tabele wymiarów** i **tabelę faktów**. Wszystkie zaprojektowane tabele proszę umieścić na **diagramie graficznym** Nie jest konieczne samodzielne przygotowanie skryptu poleceń SQL tworzących te tabele, ponieważ można będzie do tego celu wykorzystać Pentaho Data Integration.
2. Za pomocą narzędzia **Pentaho Data Integration** należy zaimplementować procesy ETL, które umożliwią załadowanie danych z dostarczonego pliku CSV do tabel hurtowni danych o układzie gwiazdy. Należy zwrócić uwagę na następujące utrudnienia:
    a. dane są **zanieczyszczone** i powinny być skorygowane przez proces ETL, np. wielkość liter w nazwach branż (np. "Edukacja", "EdukacjA"), nazwa płci skrócona lub pełna (np. "K", "kobiety"), nadmiarowe spacje (np. "Transport, gospodarka magazynowa___i łączność") - dokumentacja: https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps
    b. wymiar wieku jest zapisany w kolumnach zamiast w wierszach; konieczne jest **"obrócenie" kolumn** z przedziałami wiekowymi do układu pionowego - w tym celu proszę zapoznać się z działaniem operatora Row Normalizer (https://wiki.pentaho.com/display/EAI/Row+Normaliser)
    c. plik CSV stosuje kodowanie znaków narodowych zgodne z **Windows1250**
3. Za pomocą narzędzia **Pentaho Report Designer** (lub podobnego) należy zaimplementować **trzy** analityczne raporty biznesowe według własnego projektu. W raportach tych muszą wystąpić następujące elementy (wystarczy każdy element zastosować jeden raz):
    a. prezentacja danych w formie **tabelarycznej**
    b. prezentacja danych w formie **wykresu graficznego**
    c. prezentacja danych w formie **tabeli przestawnej** (układ macierzowy)
    d. **parametr wywołania** raportu, wypełniany przez użytkownika
    e. element **graficzny**, np. logo

Przesyłanie rozwiązań:
Proszę o przesyłanie rozwiązań w formie plików załączonych do e-maila na mój adres: *maciej.zakrzewicz@wsb.poznan.pl*. W przypadku kompletnego rozwiązania będą to:

1. Diagram tabel hurtowni danych: 1 plik
2. Procesy ETL (Transformation): 5 plików (4 tabele wymiarów + tabela faktów)
3. Raporty: 3 pliki

Zasady oceniania:

Do zdobycia jest maksymalnie **15 punktów**, po 5 punktów za każdy z trzech kroków zadania:

1. Diagram hurtowni danych: 0-5 punktów, w zależności od poprawności.
2. Procesy ETL: po 1 punkcie za każdy proces (razem 5).
3. Raporty: po 1 punkcie za każdy z wymaganych elementów (razem 5).