Zdanie projektowe z hurtowni danych - grupa 2020

Wykorzystując dane o rozprzestrzenianiu się w koronowirusa zbudować hurtownię danych umożliwiającą analizę tempa rozprzestrzeniania się wirusa:

Wymiary analizy:

- Geografia – kraj (atrybuty: populacja, GDP) – kontynent

- Czas – dzień – miesiąc – rok

- Czas od pierwszej detekcji (numer kolejny dnia)

- Pacjent (wobec braku danych należy je wygenerować losowo lub próbki danych z Kaggle)

    - wiek

    - płeć)

Miary:

- liczba zakażeń (w okresie)

- liczba zgonów (w okresie)

- liczba pacjentów wyleczonych (w okresie)

- liczba nowych przypadków zakażeń (granulacja dzienna)

- liczba pacjentów zakażonych (stan na dzień)

- dynamika zakażeń –liczba nowych przypadków / liczba pacjentów zakażonych w dniu poprzedzającym.

Źródła danych:

https://github.com/CSSEGISandData/COVID-19

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/data

Opis realizacji zadania:

Hurtownia powinna umożliwiać doładowanie przyrostowe danych w zakresie zagregowanej liczby przypadków (pierwsze źródło danych) . Uruchomiony proces ETL powinien pobierać najnowsze opublikowane dane i zasilać nimi hurtownie.

W ramach realizacji zadania projektowego należy:

1. Zbudować proces ETL wykorzystując komponenty: SSIS (Microsoft SQL Server Integration Services).
2. Zbudować model hurtowni danych:
    a. Schemat przejściowy (stage)
    b. Schemat hurtowni danych

3. Model kostek wielowymiarowych w Microsoft SQL Sesrver Analysis Services
4. Raporty w Power BI sięgające do struktur wielowymiarowych w Analysis Service

Dokumentacja / sprawozdanie:

1. Opis założeń biznesowych (jakie założenia przyjęto)
2. Dokumentacja procesu ETL: diagramy SSIS, etapy transformacji danych
3. Model wymiarowy hurtowni (projekt)
4. Model wymiarowy – dokumentacja powykonawcza w Analysis Services (model kostki)
5. Zrzut ekranu z przykładowych raportów w PowerBI/Excel. Przykładowe analizy i sposób prezentacji wyników:
   https://www.worldometers.info/coronavirus/
   https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Kody źródłowe:

1. Skrypt do tworzenia struktur bazy danych
2. Projekt MSSIS
3. Projekt MSSAS
4. Plik Excel/PowerBI

Sprawozdania i dokumentacja techniczna umieszczana w repozytorium kodów.

Kryteria oceny

Dst

- proces ETL pobiera dane z plików pobranych bezpośrednio z zasobów sieciowych (pliki źródłowe nie mogą być korygowane ręcznie)
- poprawność transformacji danych i wyników końcowych
- dostęp do danych z Excela, biznesowy model danych zrozumiały dla użytkownika końcowego, możliwość budowy raportów ad-hoc
- zaimplementowane miary:
  - liczba zakażeń (w okresie)
  - liczba zgonów (w okresie)
  - liczba pacjentów zakażonych (na dzień)
- zaimplementowane wymiary
  - geografia
  - czas astronomiczny

Db

- zaimplementowane wszystkie miary i wymiary analizy (bez wymiaru danych pacjenta)
- możliwość raportowania ad-hoc zaimplementowana w PowerBI
- budowa przykładowych raportów w PowerBI
- Terminowość rozliczania kamieni milowych

Bdb

- proces ETL łączy się online do źródeł danych i pobiera najnowsze dane do aktualizacji
- czas od pierwszej infekcji mierzony indywidualnie dla kraju
- implementacja wymiaru opisującego cechy pacjenta demograficzne (wygenerowane losowo dane, przyjęcie poziom granulacji hurtowni – pacjent). Jako wzór można przyjąć dane z tabeli Kaggle.