Piotr Rybarczyk
Nr Indeksu 10117
Nr Grupy Z507

Projekt Zaliczeniowy
Hurtownie Danych

Hurtownia Danych dla Sieci Restauracji

Założenia Projektu

Ten projekt jest swojego rodzaju przedłużeniem projektu z przedmiotu Bazy Danych. Na przedmiocie Bazy Danych zajmowałem się tworzeniem bazy danych dla sieci restauracji, która będzie używana przez aplikacje klienckie takie jak: panel admina, panel zamówień czy raporty.

Ten projekt, jest skupiony na raportowej części tego projektu, jako że istnieje nieskończona ilość różnych raportów które możemy przygotować dla takiego biznesu, w swoim rozwiązaniu zamierzam skupić się głównie na dwóch obszarach czyli Zamówieniach i Restauracjach.

Jako techniczne założenia, zamierzam skorzystać z:

- Podejścia **ELT** (Extract Load Transform)
- Czterech różnych warstw:
    - **INGESTION** - bazowa kopia części danych z głównej bazy danych.
    - **CLEAN** - poziom zajmujący się wstępnym czyszczeniem danych.
    - **MODELED** - poziom w którym znajduje się główna część projektu czyli Fakty i Wymiary

Samo rozwiązanie zbudowane będzie przy wykorzystaniu następujących technologii:

- **Baza Danych**: Microsoft SQL Server
- **Tworzenie Hurtowni**: Microsoft Analysis Services

Wycena Projektu

Punkty

- 1 pkt za dokumentację
- 3 pkt za 6 faktów i 6 wymiarów
- 2 pkt sa Hurtownie
- 1 pkt za termin oddania przed egzaminem
- ? pkt za opis technologii jak ELT,
- => 7 + ?

Osoby

- Piotr Rybarczyk 10117 - punkty: wszystko.

Warstwy projektu

Warstwa Ingestion

Ten poziom, to prosta kopia głównej bazy danych. Trzymam go w raportowej bazie danych, ponieważ w dzisiejszych czasach pamięć jest tania w porównaniu do kosztów pracy ludzkiej. Posiadania poziomu ingestion umożliwia nam wielokrotne puszczanie transformacji w celu np wprowadzenia zmian, lub poprawy błędów.

Dodatkowo, dzięki takiemu zastosowaniu możemy pokusić się o wiele innych rozwiązań np.

- Zbudowanie historii zmian konkretnych danych przy wykorzystaniu Capture Data Change.
- Wykorzystanie wielu różnych mechanizmów do pobierania danych np. 3rd party services & Kafka

Jednak co najważniejsze, dzięki temu w bazie raportowej mamy dostęp do danych źródłowych.

Warstwa Cleaned

Drugi poziom, pierwszy istotnie zmieniający sposób prezentacji danych ale wciąż będący swoistą kopią danych ze źródła. Na tym poziomie, możemy wprowadzić istotne poprawki dla danych znajdujących się w warstwie Ingestion takie jak:

- Prezentacja tylko najnowszych danych tabeli
- Dodanie wersji dla danych wpisów w tabeli
- Zbudowanie wielu różnych na tych samych danych źródłowych.
- Ujednolicenie stref czasowych
- Usunięcie duplikatów
- Czy inne poprawki, które możemy wprowadzić bez ruszania warstwy Modeled

I wiele innych
Ta warstwa może opierać się tylko na nie zmaterializowanych widokach, dzięki czemu zawsze będzie reprezentować najnowszy stan danych.

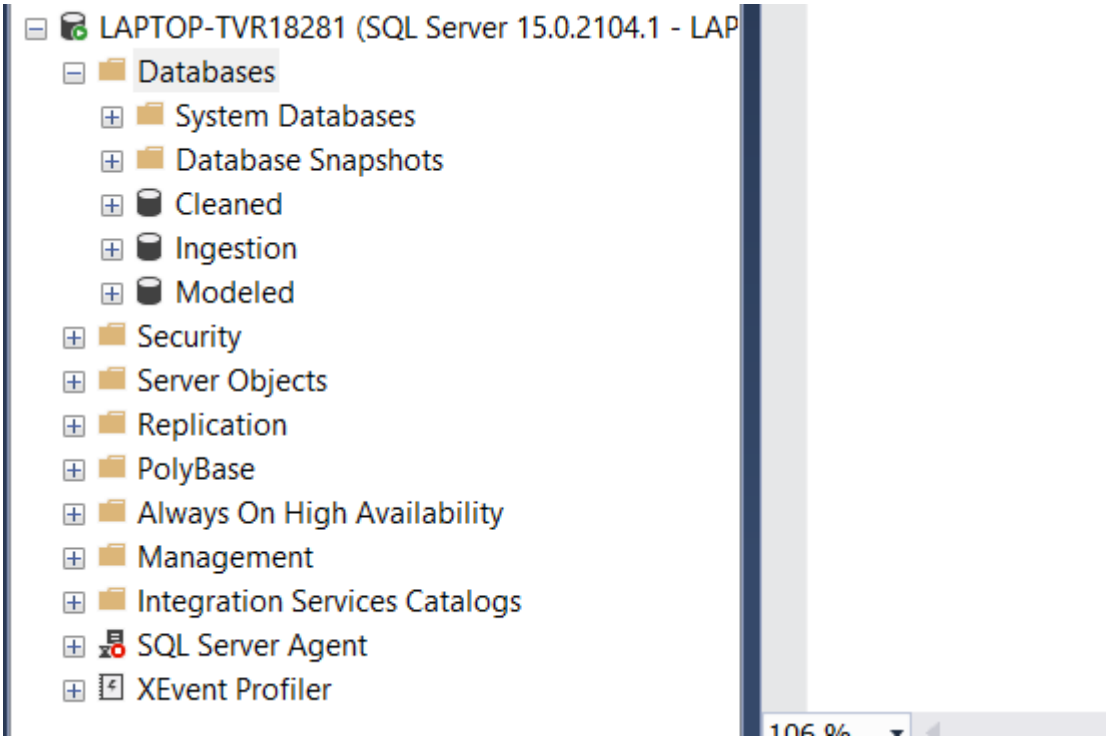
W przypadku problemu z wydajnością w dostępie do tych danych, możemy zmienić typ z View na Table lub Materialized View, żeby stworzyć inne indeksy, które przyspieszą odczyty. Jest to jednak zależne od tego na jakiej bazie tworzymy Data Warehouse.

Warstwa Modeled

Główna warstwa danych dla Hurtowni, to tutaj znajdują się już obrobione i gotowe do użycia w raportach dane w formie Faktów i Wymiary.
Ta warstwa może zawierać zmaterializowane widoki jak i niezmineralizowane widoki dzięki czemu możemy wybierać pomiędzy wydajnością odczytu dla zmaterializowanych widoków i świeżością danych.

Warstwa Ingestion

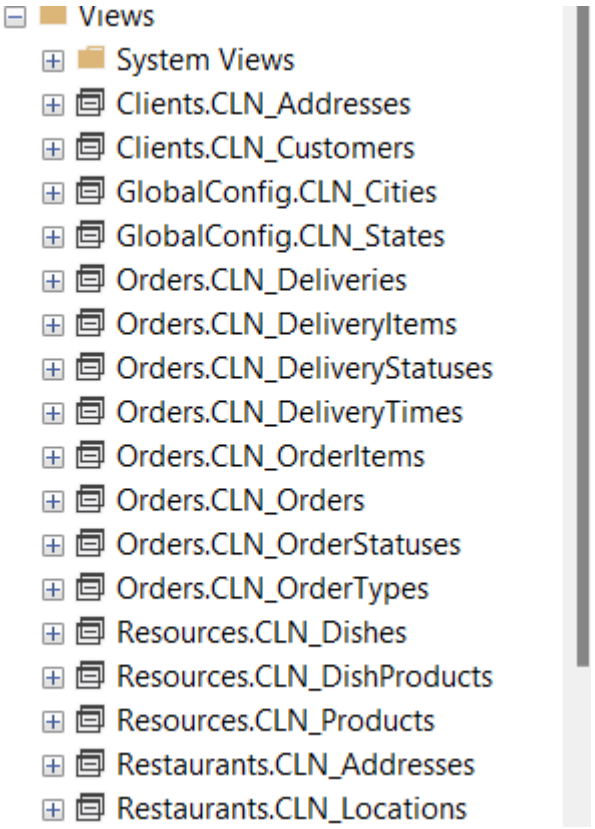
Warstwa Ingestion dla projektu została zbudowana z wykorzystaniem zwykłej kopii głównej bazy danych na tym samym serwerze SQL.
W produkcyjnym rozwiązaniu pokusiłbym się o inne rozwiązanie, np wykorzystanie Redshift, Snowflake lub innego Data Warehouse i ładowanie danych za pomocą CDC lub 3rd party service jak Fivetran.



Warstwa Cleaned

Warstwa Cleaned została utworzona, jako widoki na warstwę Ingestion, z pominięciem rekordów, które zostały usunięte. W założeniu biznes, nie powinien usuwać żadnych rekordów z bazy danych w myśl jeśli Order został złożony, należy go wycofać a nie usuwać jako, że są to dwie różne akcje biznesowe. Zdarzają się jednak przypadki, kiedy dane należy usunąć ponieważ zostały błędnie wprowadzone do systemu np. poprzez import.

Widoki mogą być tworzone za pomocą skryptów lub za pomocą oprogramowania typu DBT.



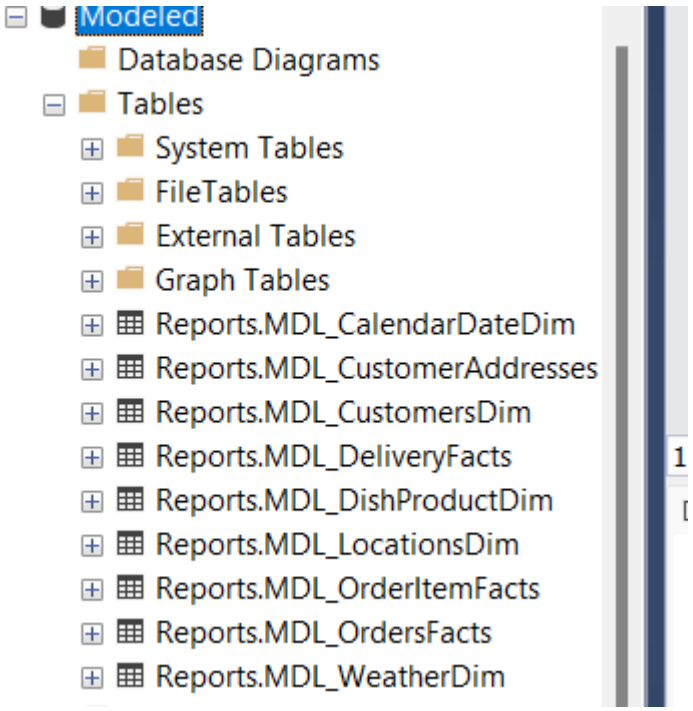
Warstwa Modeled

Główna warstwa naszego projektu, warstwa w której zajmujemy się właściwym modelowaniem danych. W założeniu, warstwa ta powinna korzystać tylko z warstwy Cleaned, aby uniknąć przeciekania nieobrobionych.
Moglibyśmy wymusić taką regułę poprzez wykorzystanie innego konta usera do pracy na tej warstwie, który nie miałby możliwości sięgania do warstwy Ingestion.

Sama warstwa Modeled zbudowana jest korzystając z podejścia **Constellation Star Schema,** tj. z wykorzystaniem wielu Tabel Faktów i Wymiarów, gdzie tabele Faktów nie mają pomiędzy sobą wzajemnej relacji innej niż poprzez tabele Wymiarów.

Każda tabela tutaj stworzona jest jako tabela budowana na bazie widoków z warstwy Cleaned.

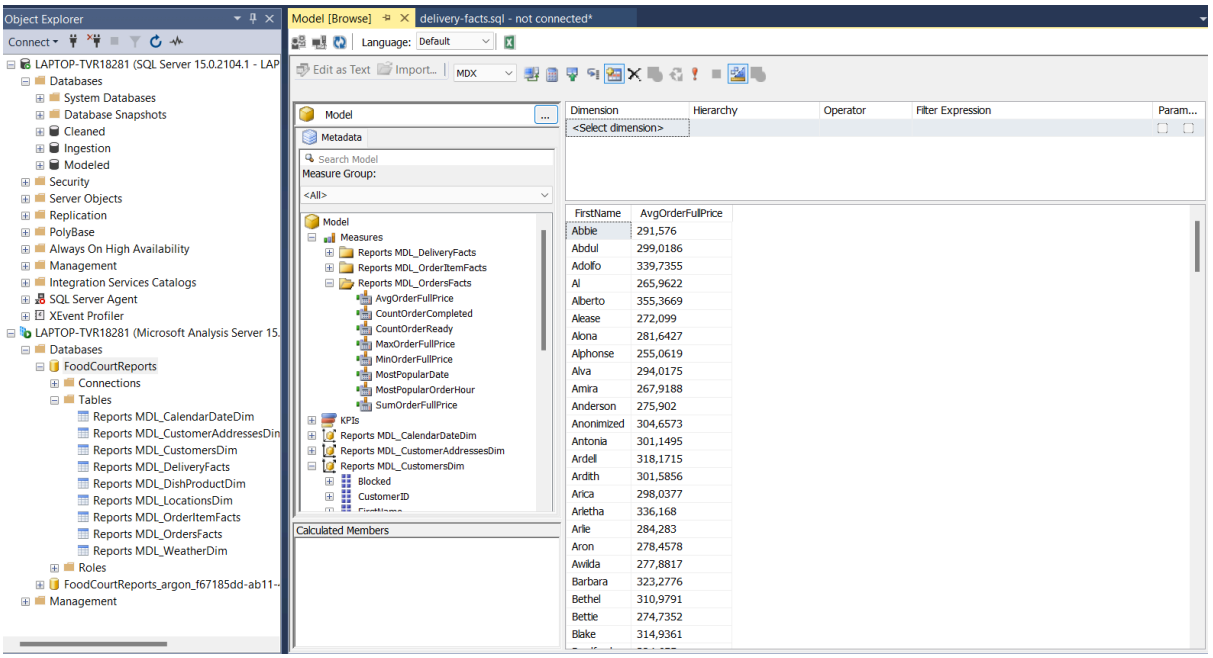
Tabele mogą być tworzone za pomocą skryptów lub przy pomocy technologii typu DBT.



Hurtownia Danych

Do utworzenia Hurtowni Danych korzystamy, z narzędzia Microsoft Analysis Services.
Po wprowadzeniu Miar, głównie na bazie cen, i hierarchii na bazie dat i geografii możemy wdrożyć hurtownie na serwer i udostępnić klientowi.
Dalej, można by pokusić się o lepsze przerobienie danych geograficznych. Np pozyskując dane o położeniu lokacji, i mając dane adresów klientów możemy analizować jak wpływa odległość dostawy na czas, koszt etc.

Do stworzenia hurtowni wybrałem model tabelaryczny korzystając z Microsoft Analysis Services, zastanawiałem się nad Kostką, ale skoro cały powyższy proces opiera się na fakcie, że dzisiaj zwykle tańszy jest dodatkowy RAM lub dysk twardy niż godziny pracy pracownika, takie rozwiązanie przeczyłoby powyższej idei.



Tabele Faktów

- **Order Facts -** tabela zawierające dane zamówień składanych w restauracjach.
- **Order Item Facts -** tabela zawierająca dane zamówionych dań w restauracjach
- **Delivery Facts -** tabela zawierająca dane dostaw wysłanych z restauracji.
- **Delivery Item Facts -** tabela zawierająca dane o dostawach każdego z produktu z osobna. Jeden produkt może być obecny w wielu dostawach bo np. ktoś zapomniał go spakować.
- **Location Menus Facts -** tabela zawiera dane o historii Menu per lokacja.
- **Reservations Facts** - tabela zawiera dane o historii rezerwacji w lokacji.

Tabele Wymiarów

- **Wymiary Daty i Czasu -** do tworzenia raportów np. o popularności dni, godzin czy kwartałów.
- **Wymiar Pogody** - data, miasto i pogoda, do tworzenia raportów pogodowych.
- **Wymiary Lokacji -** lokacja, adres lokacji etc. do tworzenia raportów popularności
- **Wymiar Pracownika -** dane pracownika

- **Wymiar Klienta** - dane klienta

Miary

- **Cenowe**: głównie min, max, avg cen poszczególnych zamówień czy produktów
- **Ilościowe**: głównie min, max, avg wielkości poszczególnych zamówień, ilości miejsc siedzących etc.

Hierarchie

- **Czasu -** rok, miesiąc, tydzień, dzień
- **Położenia -** geograficzne hierarchie stan -> miast. Można rozszerzyć dalej do dzielnic lub ulic.