Fundusze Europejskie Wiedza Edukacja Rozwój
Politechnika Wrocławska
Unia Europejska Europejski Fundusz Społeczny

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Hurtownie danych

Logiczna organizacja hurtowni danych

dr inż. Bernadetta Maleszka

Modelowanie danych biznesowych

1. Wybór danych
2. Czas w kluczu
3. Dane powiązane
4. Ziarnistość
5. Podsumowania
6. Złączenia tabel źródłowych
7. Tworzenie tabel wynikowych
8. Segregacja

Podsumowanie

Krok	Cel
Wybór danych	odpowiedni zakres, redukcja danych
Czas w kluczu	historia danych, „point-in-time -> over-time”
Dane powiązane	spójność, dostęp do niezbędnych danych
Ziarnistość	odpowiedni poziom szczegółowości (koszt przetwarzania danych)
Podsumowania	agregacje
Złączenia tabel źródłowych	krótszy czas oczekiwania na wynik
Tworzenie tabel wynikowych	poprawa wydajności dostarczania danych
Segregacja	zrównoważenie wydajności gromadzenia i dostarczania danych

Model konceptualny

- Warstwa pojęciowo - funkcjonalna
  - operuje na poziomie informacji, definiuje funkcje (biznesowe) HD
  - posługuje się językiem pojęć biznesowych
- Warstwa logiczna
  - operuje na poziomie informacji i danych
  - mapuje pojęcia biznesowe (informację) na język danych
- Warstwa fizyczna
  - operuje na poziomie danych
  - stanowi implementację warstwy logicznej

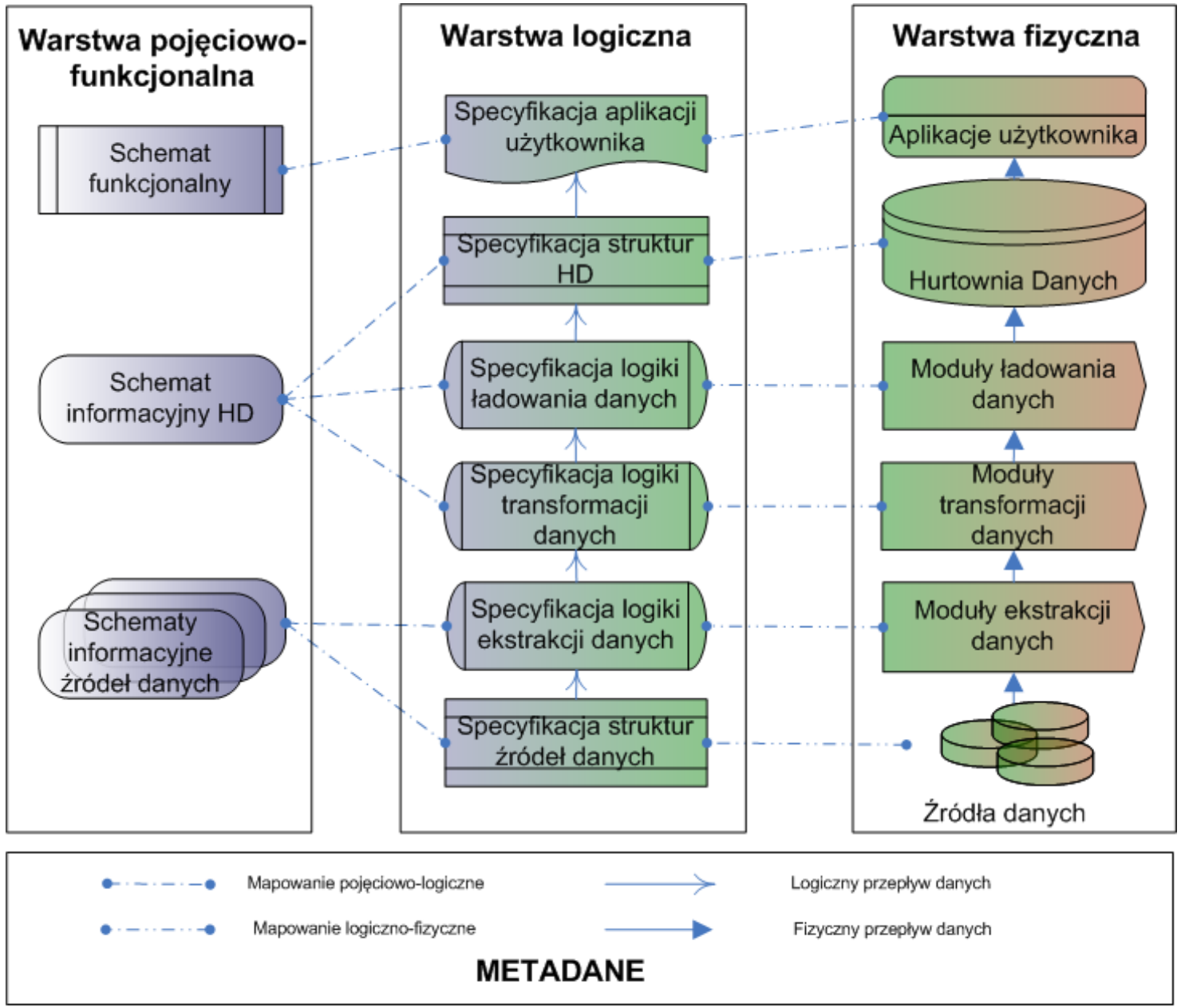
Modelowanie hurtowni danych

- Model biznesowy
  - Efekt analizy strategicznej
  - Identyfikacja miar i wymiarów dla poszczególnych procesów biznesowych

- Model logiczny (wymiarowy)
  - Model abstrakcyjny, konceptualny
  - Encje i atrybuty (reprezentowane w modelu relacyjnym jako tabele i powiązania między nimi)

Modelowanie hurtowni danych

- Model fizyczny
  - Wybór sposobu składowania danych
  - Formaty danych
  - Strategie partycjonowania
  - Wybór indeksów
  - Wybór materializowanych perspektyw



Opracowane na podstawie M.Pajer „Metody i technologie budowy…”

Warstwa logiczna

- Specyfikacja aplikacji użytkownika
- Specyfikacja struktur HD
- Specyfikacja struktur ETL
  - Specyfikacja logiki ekstrakcji danych
  - Specyfikacja logiki transformacji danych
  - Specyfikacja logiki ładowania danych
- Specyfikacja struktur źródeł danych

Model struktury danych

- Fakty (zdarzenia)
  - ujęcie ilościowe miary – np. cena, liczba sztuk, wartość, itp.
  - główna tabela HD
- Wymiary
  - nadają kontekst faktom – np. kto, gdzie, kiedy, jak, itp..
  - odrębne tabele
  - mogą być hierarchiczne

- Projekt logiczny HD: określenie faktów i wymiarów je opisujących

Rodzaje wymiarów

- Wymiary zgodne (ang. conformed dimensions)
  - wymiar, który pozostając w relacji z wieloma faktami ma takie samo znaczenie
  - integracja różnych źródeł
  - dwa wymiary są uzgodnione jeśli są identyczne lub jeden jest podzbiorem drugiego
  - Ten sam wymiar używany dla różnych tabel faktów

- Wymiar wielokrotnego stosowania (ang. role-playing dimension)
  - wymiar przechowywany w jednej tabeli, ale wykorzystywany wielokrotnie w tej samej kostce (np. data)

Rodzaje wymiarów

- Wymiar abstrakcyjny (ang. junk dimension)
  - Połączenie różnych wymiarów o małej liczebności atrybutów, np. płeć i grupa wiekowa
  - Iloczyn kartezjański „małych” wymiarów
  - Cel: poprawa wydajności zapytań

- Wymiar zdegenerowany (ang. degenerate dimension)
  - Liczność wymiaru jest porównywalna z liczbą faktów
  - Klucz biznesowy przechowywany w tabeli faktów

Wolno zmieniające się wymiary

- Przyczyny:
  - Powiązanie pozycji wymiaru z faktem jest zmieniane lub anulowane
  - Wartości atrybutów pozycji wymiaru ulegają zmianie (w kontekście czasu) w rozpatrywanym wycinku rzeczywistości

- Typy:
  - Zmiana traktowana jest jako błąd (Typ 0)
  - Pamiętana jest ostatnia wartość (nadpisanie -Typ 1)
  - Pamiętana jest cała historia zmian (Typ 2)
  - Pozostawia się historię zmian w ograniczonym zakresie np. trzy ostatnie zmiany (Typ 3)
  - Nieaktualne dane przenieś do tabeli historycznej (Typ 4)

Wymiary szybkozmienne

- Atrybut lub grupa atrybutów zmienia się szybko, ale w ograniczonym zakresie
  - szybkość definiowana jest przez rzeczywistość biznesową

Rozwiązania:

- Miniwymiar
  - tabela odpowiadająca wszystkim dopuszczalnym wartościom
  - jeśli zmienia się kilka atrybutów, to łączymy te miniwymiary w wymiar abstrakcyjny
- Dodatkowa tabela faktów typu „fakty bez faktów”
  - łączy wymiar (atrybuty nieszybkozmienne) i miniwymiar lub wymiar abstrakcyjny
  - może łączyć się też z innym wymiarem, np. czas wprowadzenia zmiany

Wymiar hierarchiczny

- Hierarchie występujące naturalnie
- Zbalansowanie hierarchii
- Hierarchia typu rodzic-dziecko
- Domyślny element wymiaru

Tabela faktów - rodzaje

- Fakty bez faktów (ang. factless facts)
  - Fakty bez miar, samo zdarzenie jest faktem
  - Przykłady: wizyta pacjenta u lekarza, strzelenie bramki przez zawodnika, udział poszkodowanego w kolizji/wypadku, itp. Możliwa jest sytuacja, że mamy informacje o miarach przy tych zdarzeniach.

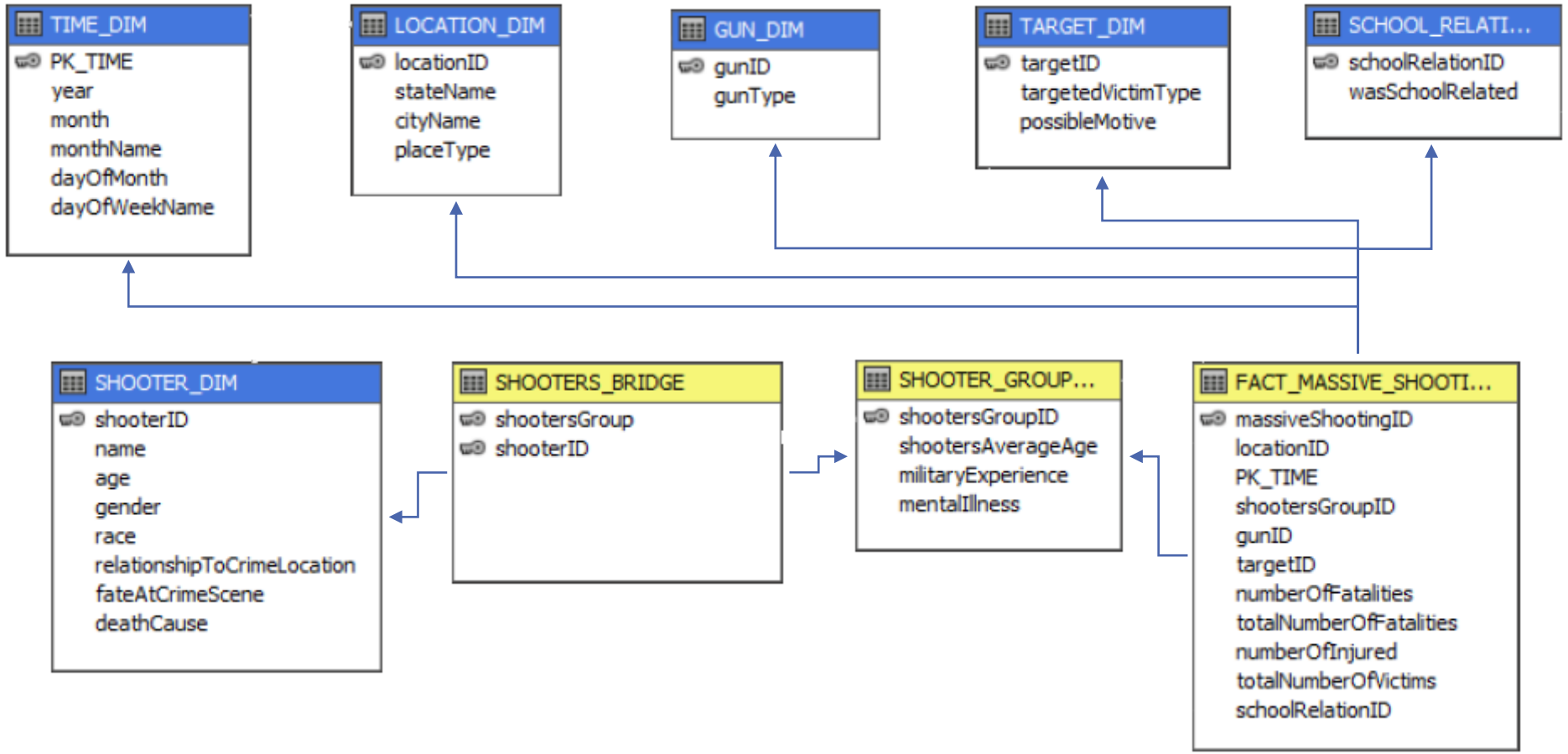
- Brakujące miary
  - NULL – niezbędne zabezpieczenie `IF NOT IsNull(atrybut)`
  - Wartość domyślna – problem ze znaczeniem
  - Wartość domyślna + dodatkowa flaga bool’owska ((true, 1), (false, 0))

Most (ang. bridge)

- W efekcie denormalizacji możemy uzyskać tabele połączone relacją wiele-do-wielu
  - wielowartościowe wymiary, np. wielu autorów utworu muzycznego
  - wielowartościowe atrybuty, np. wiele umiejętności/zainteresowań pracownika

- Pytania:
  - jak wyliczyć zysk ze sprzedaży konkretnego albumu / sumaryczny zysk pojedynczego autora?
  - jak przygotować zestawienie zysku firmy w zależności od elementarnych umiejętności pracownika?

Most


TIME_DIM
PK_TIME
year
month
monthName
dayOfMonth
dayOfWeekName
LOCATION_DIM
locationID
stateName
cityName
placeType
GUN_DIM
gunID
gunType
TARGET_DIM
targetID
targetedVictimType
possibleMotive
SCHOOL_RELATI...
schoolRelationID
wasSchoolRelated
SHOOTER_DIM
shooterID
name
age
gender
race
relationshipToCrimeLocation
fateAtCrimeScene
deathCause
SHOOTERS_BRIDGE
shootersGroup
shooterID
SHOOTER_GROUP...
shootersGroupID
shootersAverageAge
militaryExperience
mentalIllness
FACT_MASSIVE_SHOOTI...
massiveShootingID
locationID
PK_TIME
shootersGroupID
gunID
targetID
numberOfFatalities
totalNumberOfFatalities
numberOfInjured
totalNumberOfVictims
schoolRelationID

Sposoby uniknięcia mostów

- Zmiana ziarnistości faktów
  - przykład: wiersz to transakcja -> wiersz to konkretny produkt z transakcji

- Wskazanie wartości „pierwotnej”
  - wskazanie podstawowej wartości z jednym kluczem obcym
  - oznaczenie odpowiedniego atrybutu jako „pierwotny”/ „podstawowy”

- Dodanie wielu atrybutów do tabeli wymiarów
  - przykład: kolejne atrybuty to nazwy umiejętności -> wartości atrybutów `true/false`
  - skalowalność -> wystarczające dla stałej, ograniczonej liczby wartości

Sposoby uniknięcia mostów

- Dodanie kolumny zawierającej konkatenację wartości atrybutów do wymiaru
  - możliwe tylko dla ograniczonej liczby wartości
  - niezbędny ogranicznik pomiędzy kolejnymi wartościami
  - łatwa prezentacja
  - trudność z zapytaniami (wieloznaczne wyszukiwania, wolniejsze działanie)
  - problemy z wyliczaniem sum miar
  - problemy z grupowaniem/filtrowaniem według konkretnych wartości atrybutów

Projektowanie miar

- Jakie dane mogę zastosować jako miary?
- Jakich miar wyliczanych potrzebuję?
- Czy jest zapewniona addytywność miar?

Specyfikacja struktur ETL

- ETL – ogół działań mających na celu:
  - pobranie danych ze źródeł,
  - przekształcenie ich do postaci pożądanej w modelu HD
  - sprawdzenie ich poprawności
  - wykonanie niezbędnych operacji, np. agregacji
  - załadowanie danych do struktur HD

Specyfikacja logiczna ETL

- Extract
  - określenie typu i zakresu danych źródłowych
  - określenie sposobu pozyskania danych
  - określenie warunków dostępności i form transmisji danych
  - zapewnienie systemu kontroli poprawności ekstrakcji

- Transform:
  - przekształcenie typów i wartości
  - przekształcenie do pożądanej struktury
  - zapewnienie poprawności i spójności ekstraktów
  - zapewnienie wydajności procesu przetwarzania, np. uwzględnienie właściwej kolejności operacji

Specyfikacja logiczna ETL

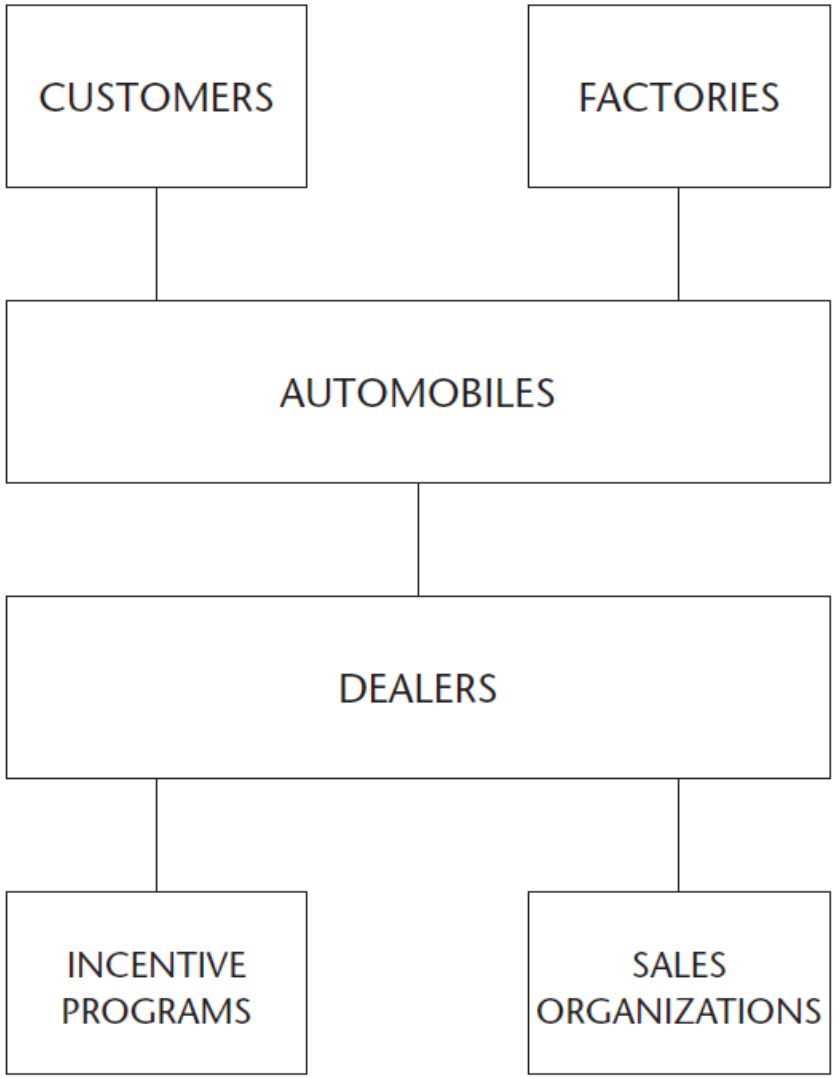
- Load:
  - wczytanie poprawnych danych w odpowiedniej kolejności
  - mapowanie elementów do odpowiednich struktur HD
  - zasilanie tematycznych HD
  - wyliczanie predefiniowanych raportów i analiz

Podsumowanie

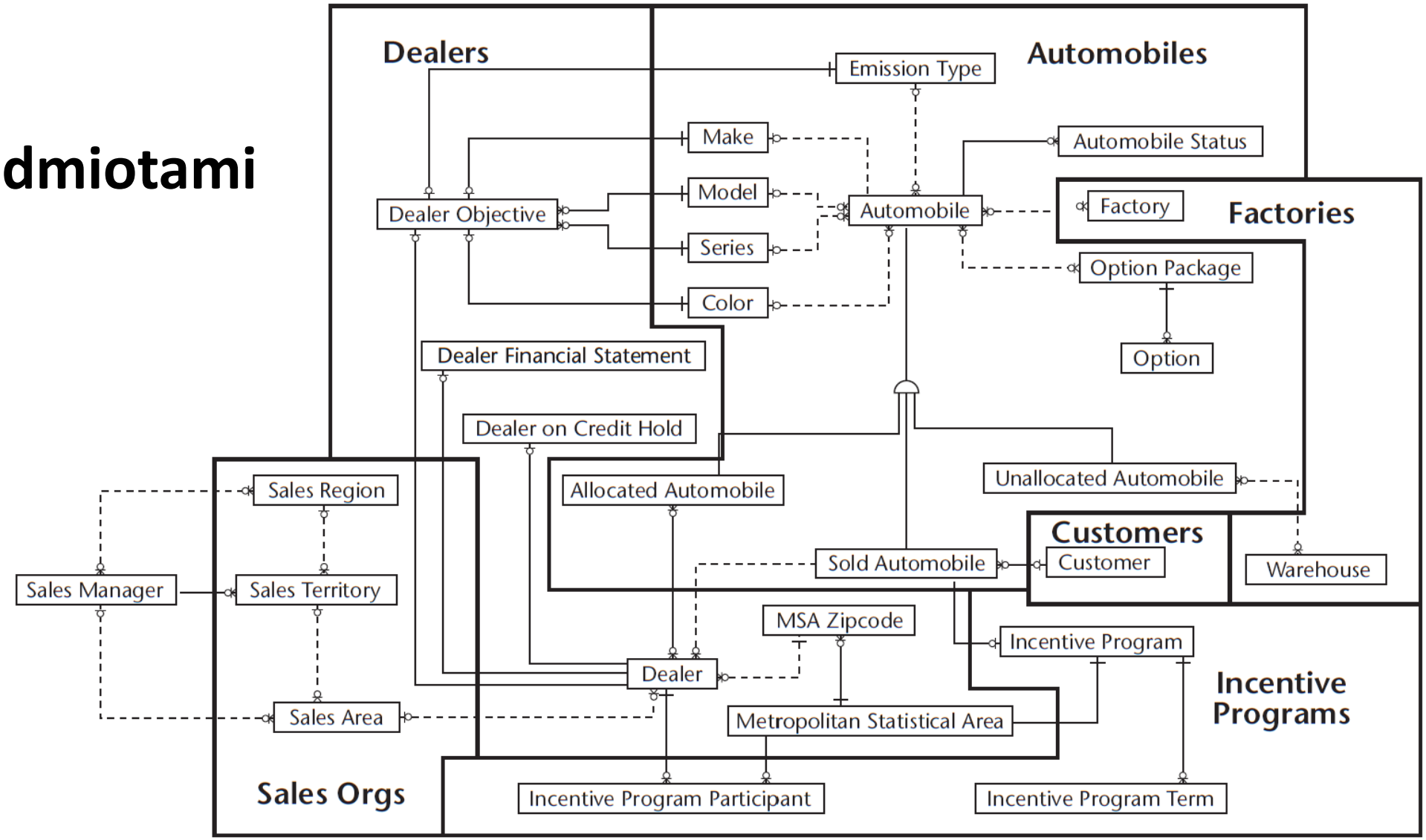
- Warstwa logiczna odpowiada projektowi technicznemu w inżynierii oprogramowania
- Specyfikacja powinna zawierać opis wszystkich algorytmów
- Warstwa logiczna jest niezbędna do rozpoczęcia implementacji

Przykład

1. Analiza sprzedaży aut
2. Obszary analizy:
   - Klient
   - Fabryka
   - Auto
   - Dealer
   - Promocja
   - Organizacja sprzedaży



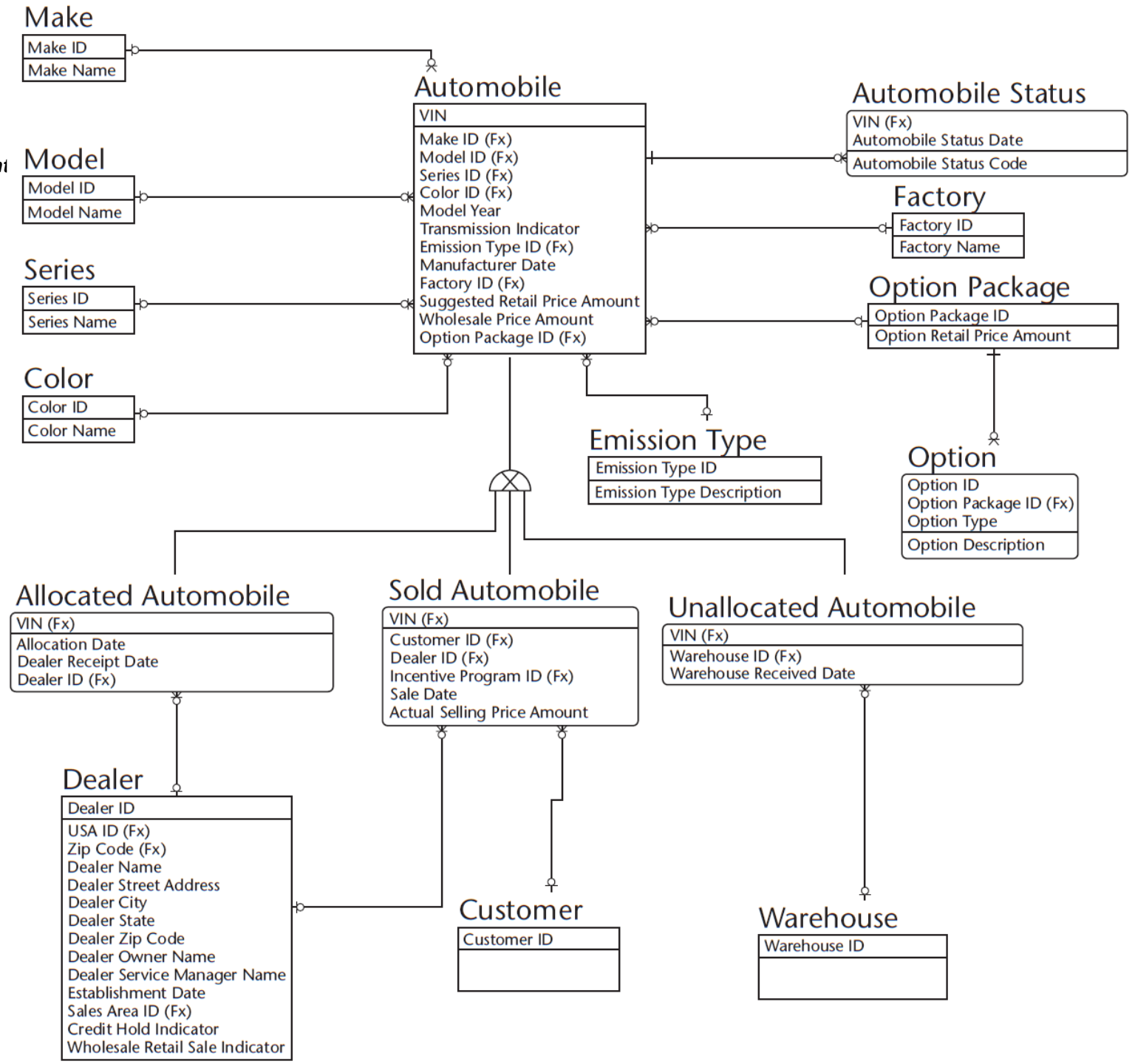
Relacje między podmiotami



Zadanie:

Zaproponuj hurtownię danych służącą do analizy sprzedaży aut:

1. Co będzie faktem?
2. Jakie będą miary?
3. Jakie będą wymiary?
4. Przygotuj 10 przykładowych zapytań analitycznych, na które hurtownia pozwoli uzyskać odpowiedź.



Hurtownie danych

Dziękuję za uwagę

dr inż. Bernadetta Maleszka