Fundusze Europejskie
Wiedza Edukacja Rozwój

Politechnika Wrocławska

Unia Europejska
Europejski Fundusz Społeczny

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Hurtownie danych

Podstawy hurtowni danych

dr inż. Bernadetta Maleszka

Przypomnienie

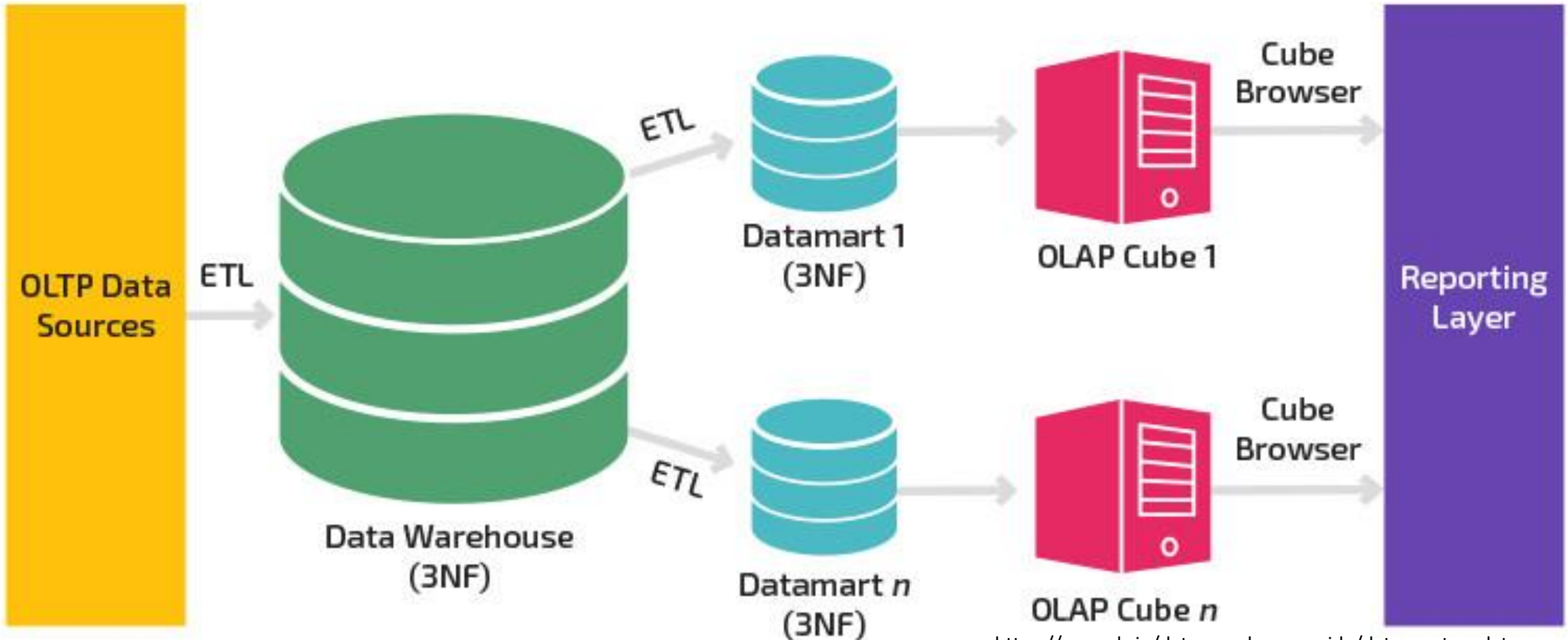
- Czym jest hurtownia danych?
- Różnice pomiędzy systemami transakcyjnymi a analitycznymi
- Podstawowe pojęcia:
  - Fakt
  - Wymiar
  - Miara
  - Kostka

Hurtownia danych wg Inmon’a

- Top-down:
  - Modelowanie -> projektowanie -> integracja -> rozprzestrzenianie danych -> raportowanie

- Dane źródłowe -> centralna hurtownia danych
  - „single version of the true” (3PN)
  - dane atomowe
  - największa możliwa szczegółowość danych
- Centralna hurtownia danych -> targowiska danych (data mart)
- Data mart -> wielowymiarowa kostka (OLAP) -> raporty

Inmon Model



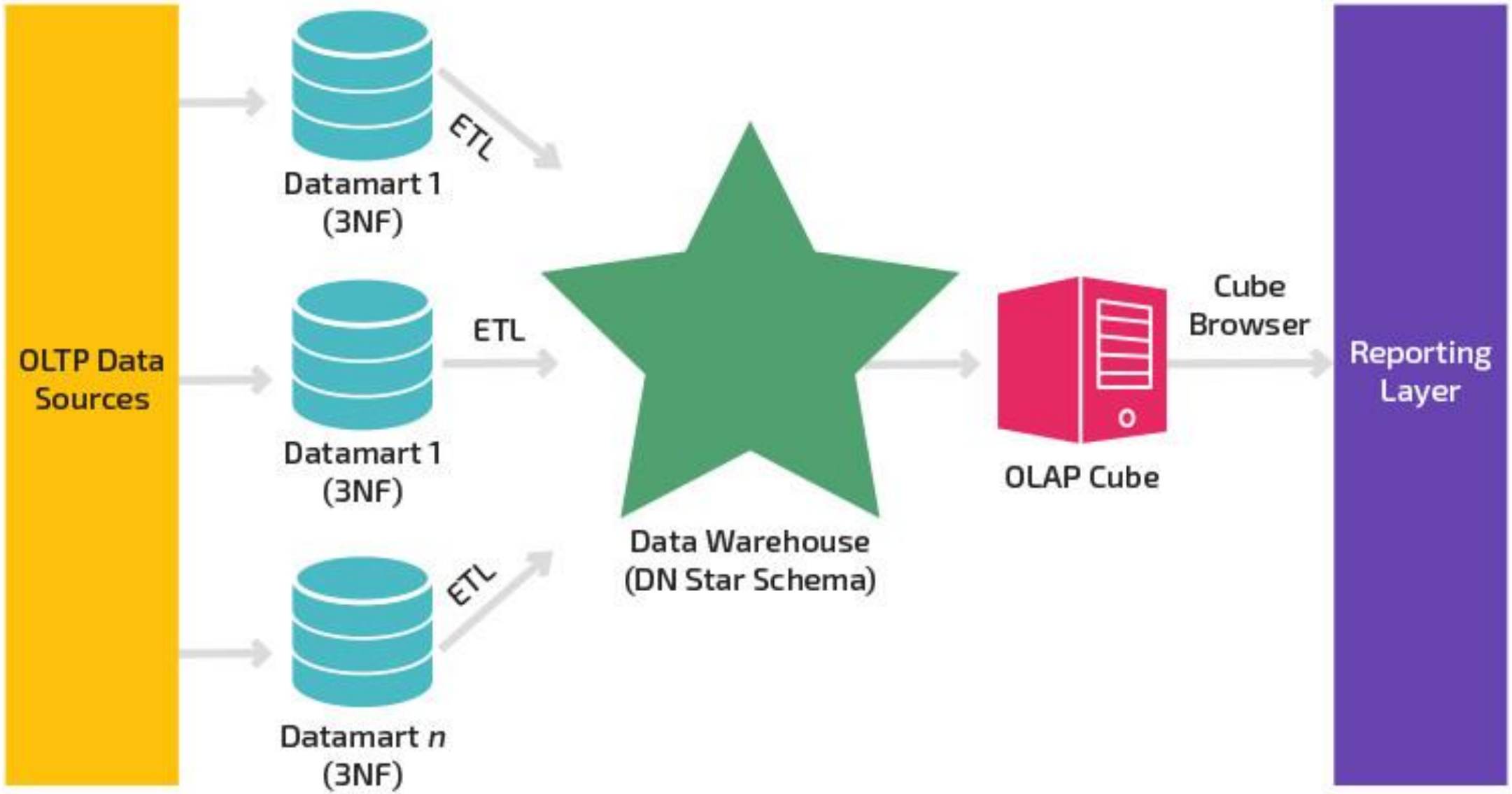
https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Hurtownia danych wg Kimball’a

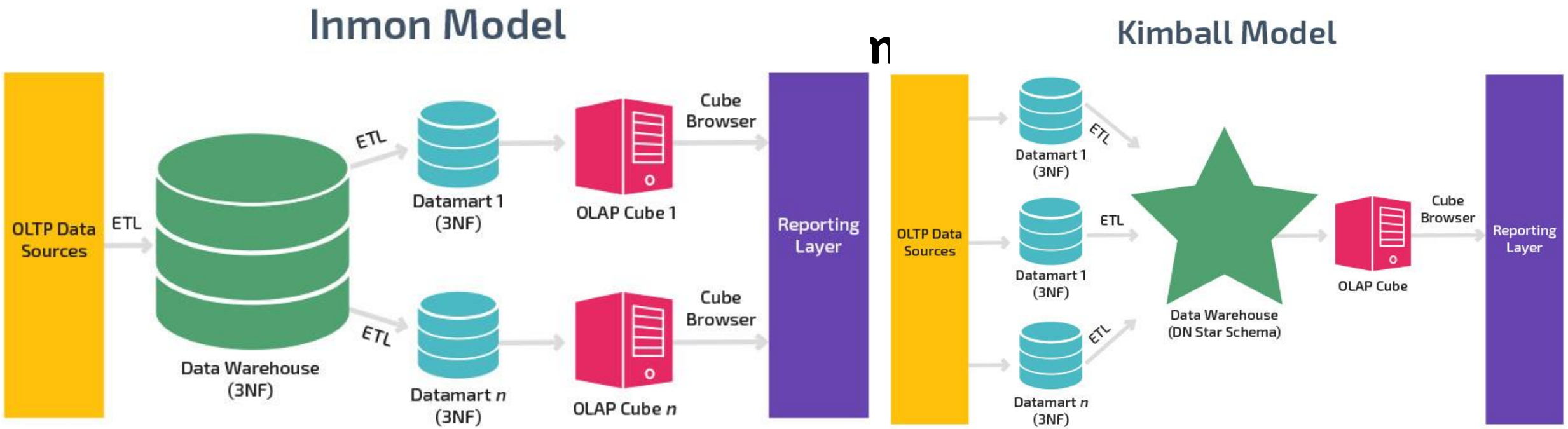
- Bottom-up
  - Integracja (zachowanie jakości danych) -> modelowanie -> projektowanie -> dostęp -> raporty

- Dane źródłowe -> targowiska danych
  - niezależne, tematyczne
  - dane atomowe lub zagregowane
  - Lokalizacja fizyczna: centralna lub rozproszona
- Data mart -> wielowymiarowa kostka (OLAP) -> raporty

Kimball Model



https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/



- Kimball: *The data warehouse is nothing more than the union of all data marts*
- Inmon: *You can catch all the minnows in the ocean and stack them together — they still do not make a whale*

https://medium.com/cloudzone/inmon-vs-kimball-the-great-data-warehousing-debate-78c57f0b5e0e

Etapy projektowania hurtowni danych

- Zrozumienie „potrzeby biznesu”
- Zrozumienie dziedziny problemowej
- Problemy w określonym wycinku rzeczywistości
- Identyfikacja potrzeb, celu i możliwości analiz biznesowych
- Wspieranie procesów decyzyjnych

Modelowanie wielowymiarowe

- Analiza dziedziny problemowej
  - Identyfikacją i zrozumieniem procesów biznesowych
- Identyfikacja problemów i potrzeb w ramach rozpatrywanej dziedziny (biznesu)
- Ocena dostępności i jakości źródeł danych
- Określenie wymagań w kontekście ustalonych procesów i celów biznesowych

Kluczowe etapy

1. Selekcja procesów biznesowych
2. Ustalenie poziomu szczegółowości (ziarnistości) rejestracji faktów
3. Identyfikacja faktów (zdarzeń biznesowych oraz wielkości pomiarowych istotnych w kontekście zarządzania i podejmowania decyzji)
4. Identyfikacja kontekstu (wymiarów) analizy faktów procesów biznesowych

Modelowanie konceptualne

1. Wyznaczenie procesów biznesowych
2. Określenie celu i zakresu analiz biznesowych w wybranych obszarach
3. Ocena dostępności i jakości źródeł danych
4. Zdefiniowanie modeli konceptualnych w kontekście uzgodnionych, wymaganych analiz faktów

Modelowanie konceptualne

- Czy i jakie zbiory danych źródłowych są dostępne
  - gdzie wiadomo gdzie się znajdują dane
  - czy mamy dostęp do danych źródłowych?
- Kto w organizacji potrzebuje informacji udostępnianych w formie raportów?
- W jaki sposób można poprawić proces decyzyjny w zakresie krótko i długo terminowym?
  - Więcej informacji
  - Udostępnić informację większej liczbie osób
  - Zmienić sposób dostępu do informacji

Modelowanie konceptualne

- Jaki rodzaj informacji uznawany jest za potrzebny w procesie decyzyjnym?
- Czy są grupy osób, które nie mają dostępu do informacji lub dostęp jest ograniczony, a ma to wpływ na podejmowane decyzje?
- Czy mamy możliwość uzyskać odpowiedź na pytania w rodzaju:
  - Co jeśli?
  - Dlaczego tak/nie?
  - Czy można?

Przykład – ogólna charakterystyka obszaru analizy

- Fabryka samochodów i podwykonawcy części aut.
- Produkcja dwóch marek samochodów – kilka modeli z każdej marki.
- Auta można nabyć jedynie za pośrednictwem dealerów.
- Dealerzy są rozliczani ze swojej sprzedaży miesięcznie/kwartalnie/itp.
- Funkcjonują wspólne oferty promocyjne dla całego obszaru sprzedaży.
- Każda fabryka podzespołów i każdy dealer operuje własnym sprzętem i oprogramowaniem.
- …

Przykład – obszar analizy

- Jaki jest miesięczny trend sprzedaży pod względem liczby i kwot sprzedawanych w dolarach każdej marki, modelu, serii i koloru (MMSC) dla konkretnego dealera, według każdego obszaru sprzedaży, regionu sprzedaży i stanu?
- Jaki jest wzorzec miesięcznej ilości zapasów według MMSC dla każdego dystrybutora, według każdego obszaru, regionu sprzedaży i stanu?
- Jak zmienia się miesięczna liczba sprzedanych samochodów ze względu na MMSC o określonym typie emisji - według dealera, fabryki, obszaru i regionu sprzedaży - w porównaniu z tymi samymi przedziałami czasowymi w poprzednim roku / poprzednich latach?
- Jaki jest trend w rzeczywistej miesięcznej sprzedaży (w dolarach i liczbach) MMSC dla każdego dystrybutora, obszaru i regionu sprzedaży w porównaniu do ich celów? Użytkownicy wymagają tych informacji zarówno według sum miesięcznych, jak i narastająco z roku na rok (YTD).
- Jaka jest historia (dwuletnie porównania) miesięcznej liczby jednostek sprzedawanych przez MMSC i powiązanych kwot w dolarach przez detalistów w porównaniu do hurtowników?

Przykład – obszar analizy

- Jaka jest miesięczna sprzedaż według MMSC w tym roku w porównaniu do tego samego czasu w ubiegłym roku dla każdego dystrybutora?
- Jaki jest miesięczny trend według MMSC dla poszczególnych rodzajów promocji, według dealera, obszaru i regionu sprzedaży?
- Jaki jest miesięczny trend w średnim czasie, jaki zajmuje dealerowi sprzedaż określonej MMSC (zwanej prędkością i równą liczbie dni od otrzymania przez dealera samochodu do daty sprzedaży) według obszaru i regionu sprzedaży?
- Jaka była średnia miesięczna cena sprzedaży MMSC dla każdego dealera, obszaru i regionu sprzedaży?
- Jaki jest trend sprzedaży gotówkowych i kredytowych dla każdego dealera i rodzajów promocji na przestrzeni miesięcy i lat (porównać odpowiadające okresy sprzedaży)?
- Porównać miesięczne ceny sprzedaży i ilości od ostatniego modelu do bieżącego modelu nadwozia dla każdego regionu sprzedaży? Modele nadwozia zmieniają się co cztery lata.

Przykład – obszar analizy

- Jaki jest miesięczny trend sprzedaży pod względem liczby i kwot sprzedawanych w dolarach każdej **marki, modelu, serii i koloru** (MMSC) dla konkretnego **dealera**, według każdego **obszaru sprzedaży, regionu sprzedaży i stanu**?
- Jaki jest wzorzec miesięcznej ilości zapasów według **MMSC** dla każdego **dystrybutora**, według każdego **obszaru, regionu sprzedaży i stanu**?
- Jak zmienia się miesięczna liczba sprzedanych samochodów ze względu na **MMSC** o określonym **typie emisji** - według **dealera, fabryki, obszaru i regionu sprzedaży** - w porównaniu z tymi samymi **przedziałami czasowymi** w poprzednim roku / poprzednich latach?
- Jaki jest trend w rzeczywistej miesięcznej sprzedaży (w dolarach i liczbach) **MMSC** dla każdego **dystrybutora, obszaru i regionu** sprzedaży w porównaniu do ich celów? Użytkownicy wymagają tych informacji zarówno według sum miesięcznych, jak i narastająco z roku na rok (YTD).
- Jaka jest historia (dwuletnie porównania) miesięcznej liczby jednostek sprzedawanych przez **MMSC** i powiązanych kwot w dolarach przez **detalistów** w porównaniu do **hurtowników**?

Przykład – obszar analizy

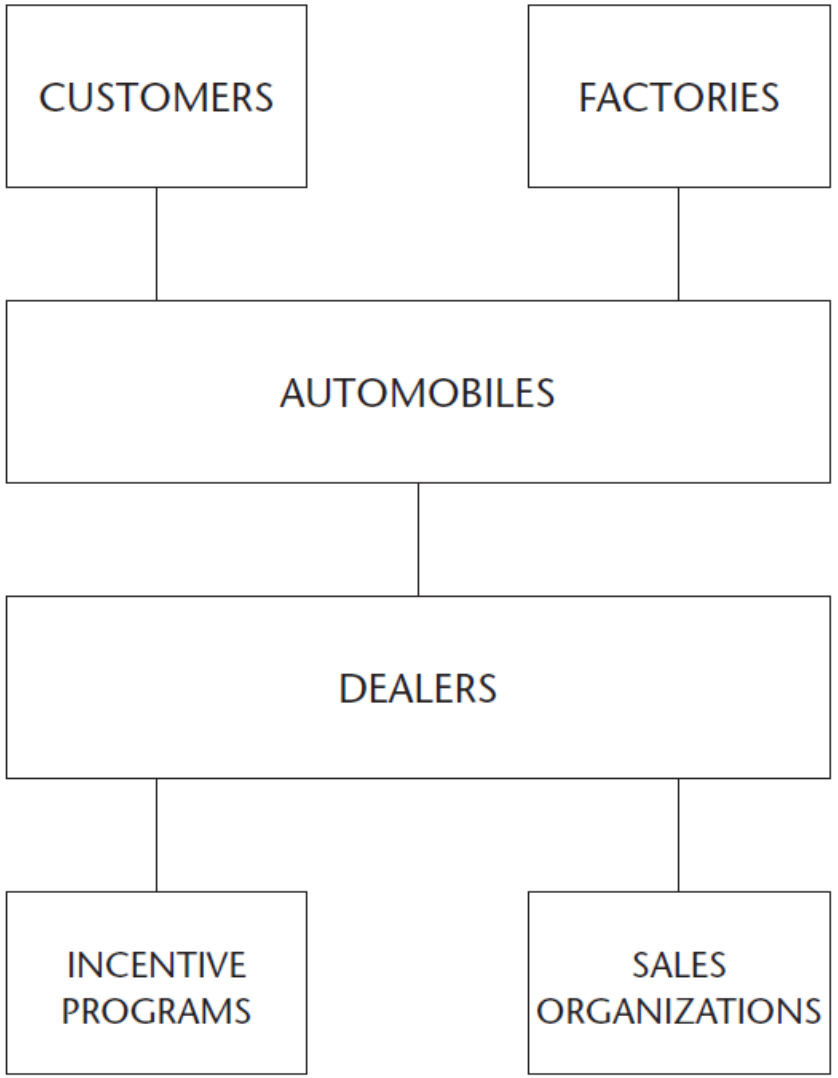
- Jaka jest miesięczna sprzedaż według **MMSC** w tym roku w porównaniu do tego samego **czasu** w ubiegłym roku dla każdego **dystrybutora**?
- Jaki jest miesięczny trend według **MMSC** dla poszczególnych **rodzajów promocji, według dealera, obszaru i regionu sprzedaży**?
- Jaki jest miesięczny trend w średnim czasie, jaki zajmuje **dealerowi** sprzedaż określonej **MMSC** (zwanej prędkością i równą liczbie dni od otrzymania przez dealera samochodu do daty sprzedaży) według **obszaru i regionu sprzedaży**?
- Jaka była średnia miesięczna cena sprzedaży **MMSC** dla każdego **dealera, obszaru i regionu sprzedaży**?
- Jaki jest trend sprzedaży gotówkowych i kredytowych dla każdego **dealera i rodzajów promocji** na przestrzeni **miesięcy i lat** (porównać odpowiadające okresy sprzedaży)?
- Porównać miesięczne ceny sprzedaży i ilości od ostatniego modelu do bieżącego **modelu nadwozia** dla każdego **regionu sprzedaży**? Modele nadwozia zmieniają się co cztery lata.

Zrozumienie „potrzeby biznesu”

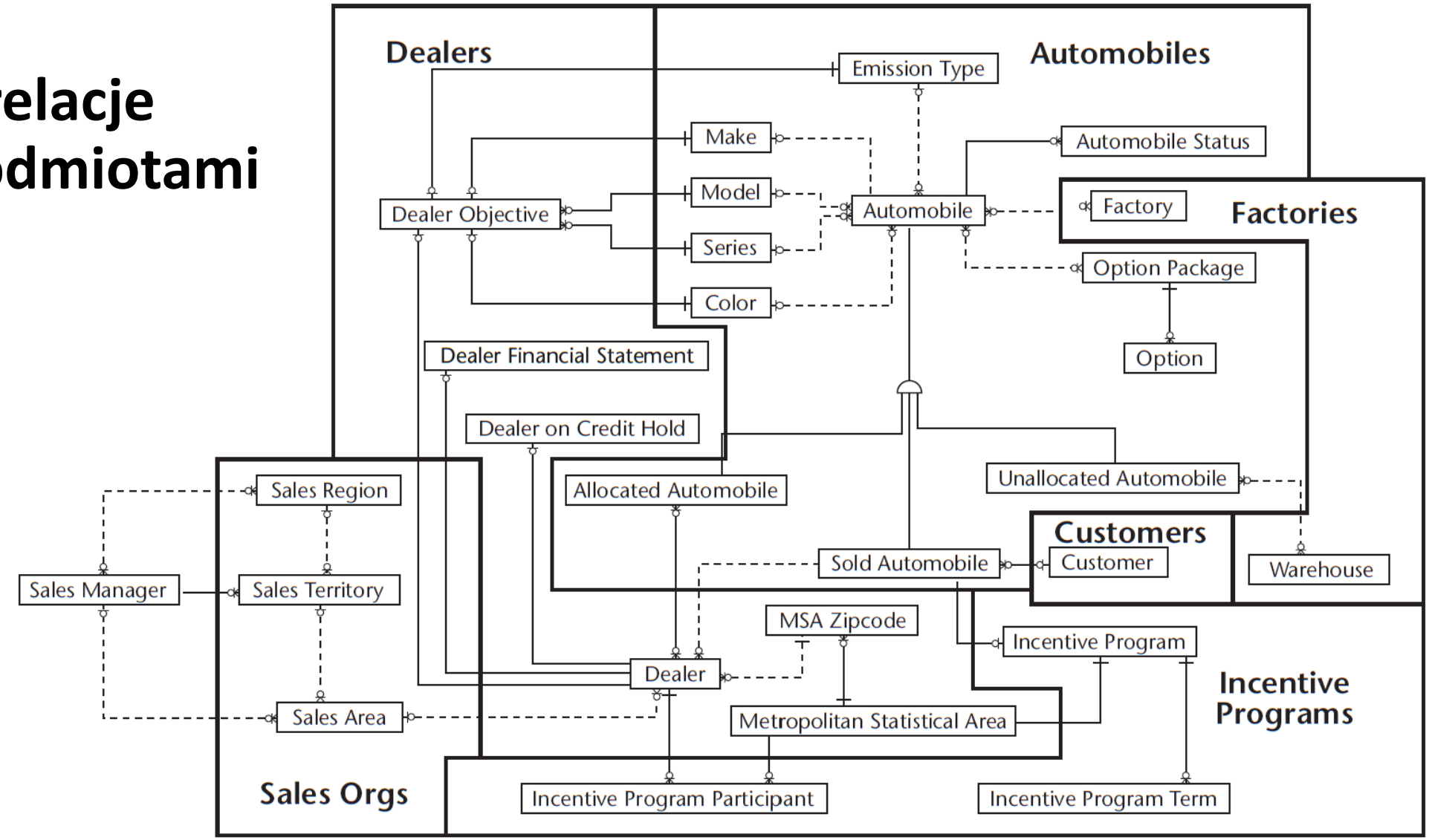
1. Zidentyfikuj obszar, z którego będą pobierane dane.
2. Zidentyfikuj interesujące podmioty w obszarze analizy i ustal ich identyfikatory.
3. Określ relacje pomiędzy tymi podmiotami.
4. Dodaj atrybuty.
5. Potwierdź strukturę modelu.
6. Potwierdź zawartość modelu.

Przykład

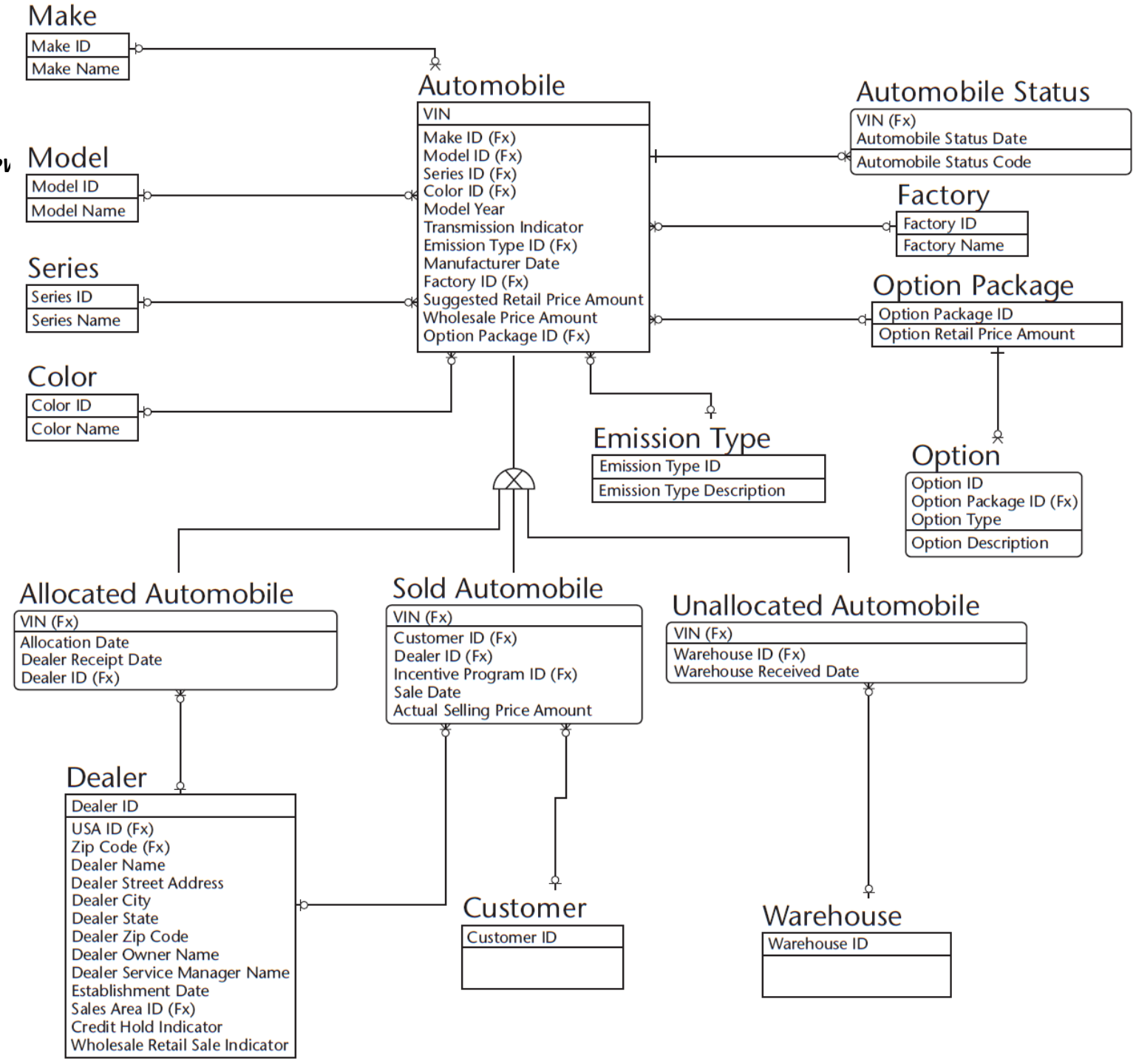
1. Zidentyfikuj obszar – analiza sprzedaży aut
2. Zidentyfikuj interesujące podmioty w obszarze analizy i ustal ich identyfikatory:
   - Klient
   - Fabryka
   - Auto
   - Dealer
   - Promocja
   - Organizacja sprzedaży



3. Określ relacje między podmiotami



4. Dodaj atrybuty



Przykład cd.

5. Potwierdź strukturę modelu
   - 3PN
   - elastyczność, stabilność, spójność
   - rzadko spotykana w wersji zaimplementowanej

6. Potwierdź zawartość modelu
   - uzgodnienie poprawności z przedstawicielami biznesu
   - sprawdzenie zgodności z regułami biznesowymi

Zrozumienie dziedziny problemowej

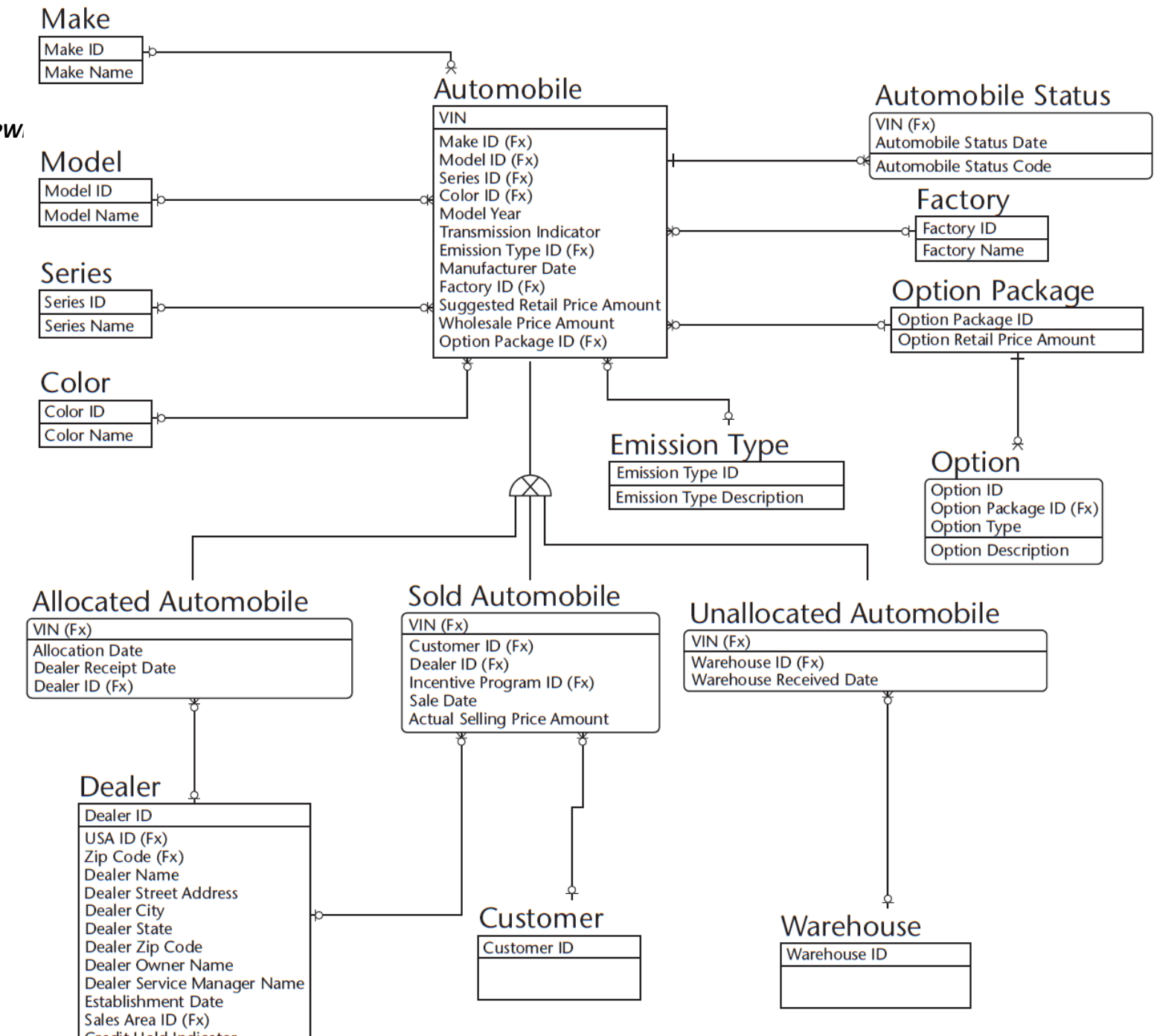
- Biznesowy model danych -> model hurtowni danych
- Najważniejsze aspekty:
  - identyfikacja wymagań
  - wybór atrybutów
  - zapewnienie spójności danych
  - tworzenie widoków i targowisk danych
  - optymalizacja

Metodologia tworzenia HD

1. Aspekty biznesowe:
    1. Wybierz interesujące dane
    2. Dodaj czas do klucza – perspektywa czasowa
    3. Dodaj dane pochodne – zapewnienie spójności
    4. Określ poziom ziarnistości
2. Aspekty wydajnościowe:
    1. Dodaj podsumowania – w zależności od ustalonego poziomu szczegółowości
    2. Dokonaj niezbędnych złączeń tabel
    3. Utwórz tabele wymiarów i faktów
    4. Segreguj dane – optymalizacja zapytań

Dyskusja

- Zgodność modelu z wymaganiami biznesowymi?
- Problemy?
- Kwestie do wyjaśnienia?



Etapy projektowania hurtowni danych

1. Charakterystyka dziedziny problemowej
2. Krótki opis obszaru analizy
3. Problemy i potrzeby
4. Cel przedsięwzięcia
    1. Oczekiwania
    2. Zakres analizy
5. Źródła danych (lokalizacja, format, dostępność)
6. Wstępna analiza źródeł danych

Wstępna analiza źródeł danych

1. Profilowanie danych
    1. Analiza danych
    2. Ocena przydatności danych w pliku do tworzenia hurtowni danych
2. Definicja typów encji/klas (wraz z własnościami) oraz związków pomiędzy nimi
3. Propozycja wymiarów, hierarchii, miar (w tym nieaddytywnych)
4. Model konceptualny
5. Model logiczny
6. Implementacja bazy danych

Tworzenie kluczy do tabel

- Potencjalne problemy:
  - niespójność – przykłady?
  - unikalność wartości

- Atrybuty będące potencjalnymi kandydatami na klucz:
  - istniejące w systemie
  - uznane standardowe klucze
  - klucze sztuczne

Wymiar zdegenerowany

- Jeżeli tabela faktów zawiera, oprócz kluczy obcych i miar, dodatkowe kolumny, to oznacza, że te kolumny pełnią funkcję zdegenerowanego wymiaru.
- Mogą to być naturalne klucze obiektów stanowiących kontekst analizy faktów.

- W przypadku wymiaru zdegenerowanego nie wykorzystuje się tabeli wymiarów.

Wymiar czasowy

- umożliwia analizę biznesową w kontekście historii zdarzeń (faktów)
- często ma strukturę hierarchiczną:
  - rok – kwartał – miesiąc – dzień

- rejestracja czasu:
  - czas wykonania transakcji
  - dane historyczne
  - logi DBMS
  - porównywanie plików
  - ingerencja w system

Wymiar czasowy

- Różne rodzaje kalendarza
- Różne długości miesięcy
- Obsługa dni wolnych
- Często nadmiarowo przechowywane atrybuty:
  - Nr dnia w roku (w roku fiskalnym)
  - Nr dnia w miesiącu
  - Nr miesiąca
  - Nazwa miesiąca (różne wersje językowe)
  - Nazwa dnia tygodnia (różne wersje językowe)
  - Data początku i końca tygodnia
  - Nr kwartału

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)

Który to typ?

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	**Nowak**	Sprzedawca

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	**Kowalska**	Sprzedawca

Który to typ?

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	Nowak	Sprzedawca

Praconik_ID	PESEL	Imię	Nazwisko	Stanowisko	Od	Do	Status
004352	90120923877	Katarzyna	Nowak	**Sprzedawca**	**2014-07-01**	**2019-03-26**	**nieaktualne**
0049872	90120923877	Katarzyna	Nowak	**Kierownik**	**2019-03-27**	**9999-12-31**	**aktualne**

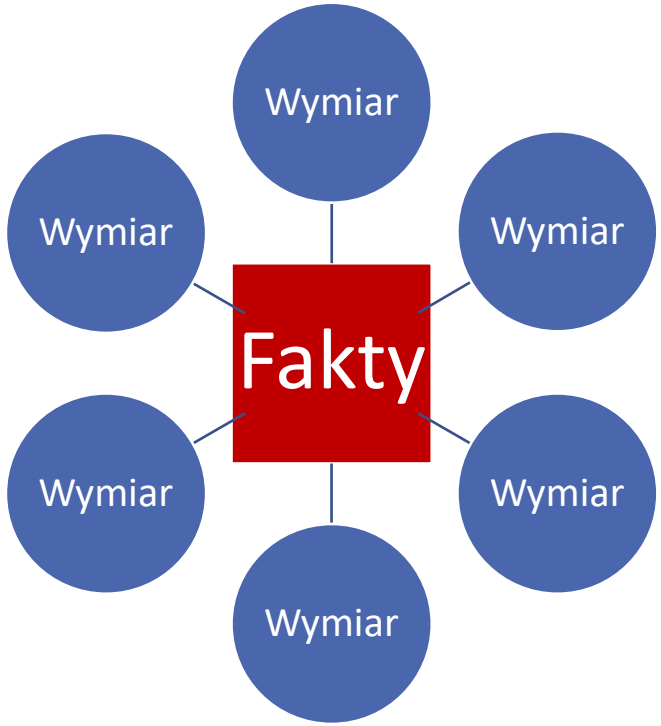
Który to typ?

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	Nowak	Sprzedawca

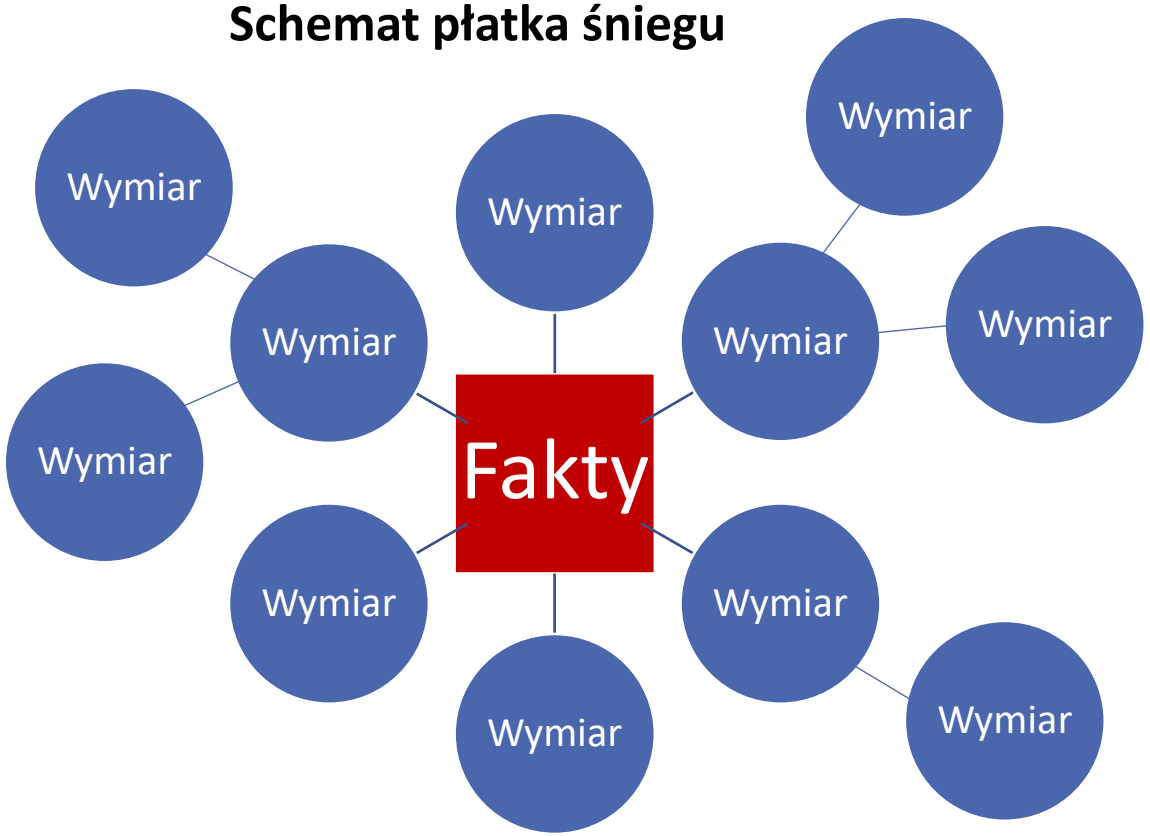
Praconik_ID	PESEL	Imię	Nazwisko	Aktualne stanowisko	Poprzednie stanowisko
004352	90120923877	Katarzyna	Nowak	Kierownik	Sprzedawca

Schematy modeli analitycznych

Schemat gwiazdy

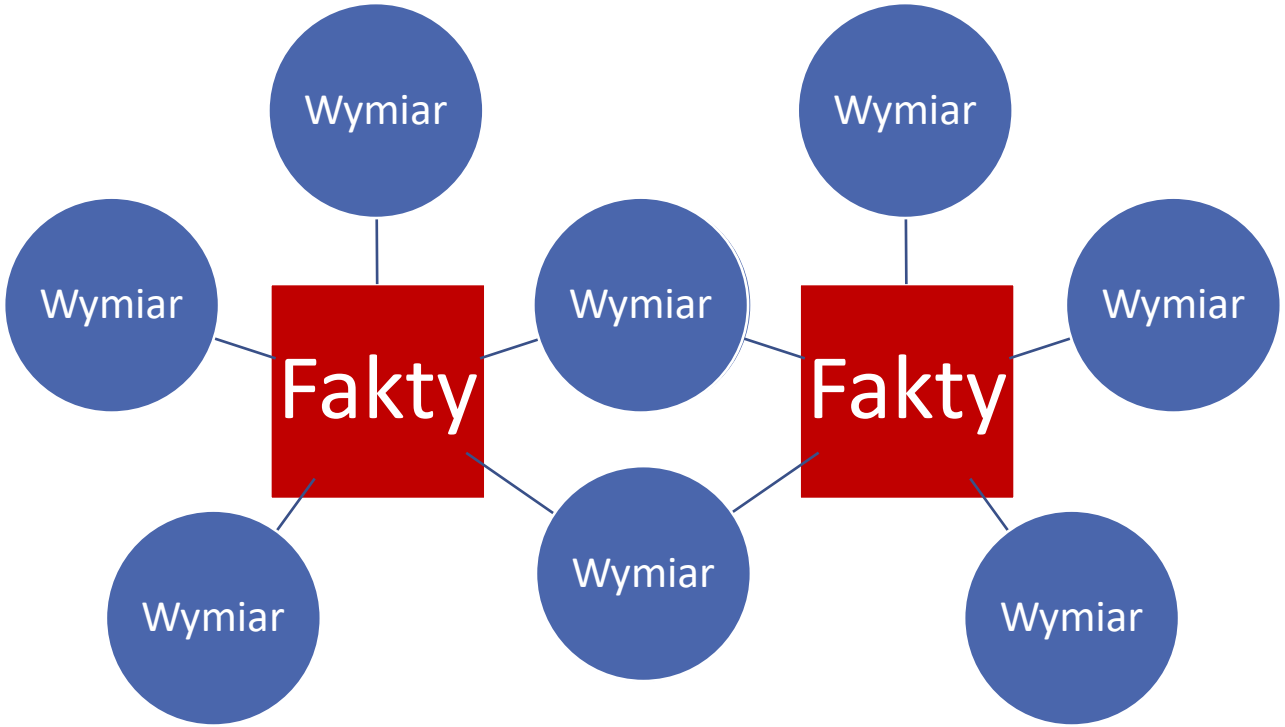


Schemat płatka śniegu



Schematy modeli analitycznych

Konstelacja faktów



Hurtownie danych

Dziękuję za uwagę

dr inż. Bernadetta Maleszka