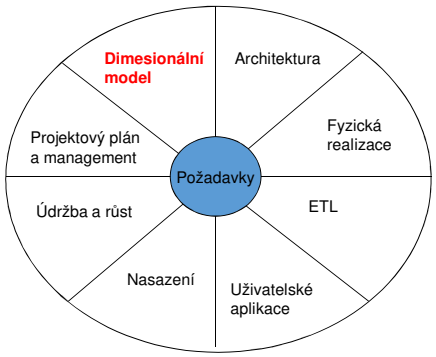
Databázové systémy a metody zpracování dat

4.přednáška



Dimenzionální modelování

- Dimenzionální modelování je rozdílné od klasického ERD datového modelování
- ERD
  - Normalizace
  - Odstranění redundance
  - Málo srozumitelné člověku
  - Optimalizován na vkládání dat a update
- Dimenzionální modelování
  - Důraz na srozumitelnost pro uživatele
  - Dosaženo standardní strukturou – fakta a dimenze
  - Základní přístup – denormalizace, redundance
  - Optimalizován na vyhledávání dat a složité analýzy
  - Základy položeny v 60. tých letech (General Mills and Dartmount University, Nielson Marketing Research)

Dimenzionální modelování

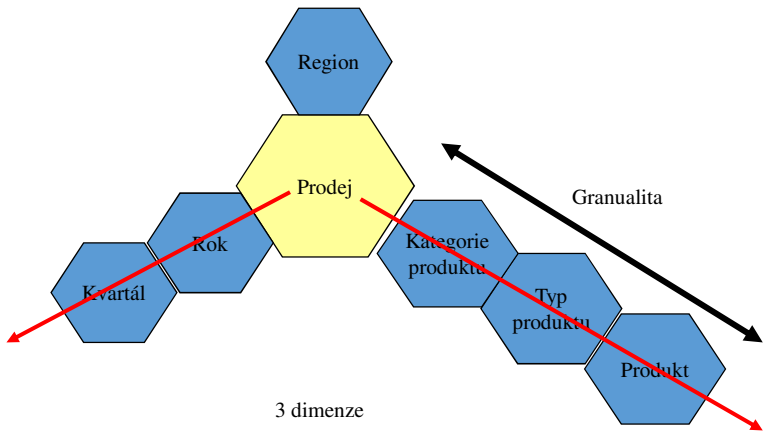
- Výhody dimenzionálního modelu
  - Standardní – navazují na to OLAP aplikace, tvůrci reportingových aplikací, …
  - Dobře rozšiřitelný (bez dopadu na aplikace)
    - Přidání nových fakt se stejnou granualitou
    - Přidání dimenze
    - Přidání atributu dimenze
  - Standardní způsoby modelování reálných problémů
    - SCD
    - Sledování událostí (Factless fact table)
    - Heterogenní produkty
    - …

Dimenzionální modelování

- Základní přístup k modelování dat v datovém skladě
- Oproti klasickému relačnímu modelu dochází k denormalizaci
- Proč dimenzionální modelování:
  - Přehledné, uživatelsky pochopitelné
  - Důraz na obchodní logiku
  - Denormalizace
    - Menší počet tabulek, spojení
    - Rychlejší odezva
    - Většina údajů v jedné tabulce
    - Indexy

Dimenzionální modelování

- Základní myšlenka multidimenzionálního modelování



Základní typy tabulek

- V datovém skladě se vyskytují dva hlavní typy tabulek:
  - Faktové tabulky
  - Dimenzionální tabulky

Dimenzionální modelování

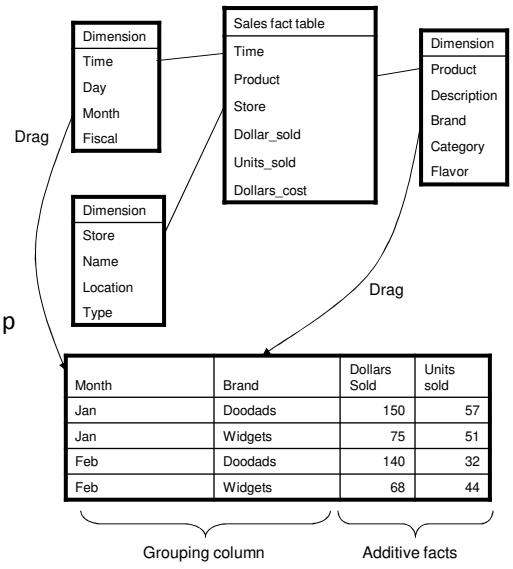
- Nejužitečnější fakta – jsou často numerická a aditivní
- Dimenze tvoří vstupní bod do DW
  - Omezení v dotazech
  - Hlavičky řádků v reportech

Dimenzionální tabulky

- Dimenizonální tabulky zachycují úhel pohledu na sledované ukazatele
- Představují vlastně „číselníky“
- Typické dimenze jsou:
  - Čas
  - Zákazník
  - Produkt
  - Prodejna
  - Smlouva

Dimenzionální tabulky

- Atributy v dimenzi:
  - Textové hodnoty (nebo se chovají jako textové)
  - Diskrétní
  - Slouží pro definici omezení a agregace v výstupech
- K výstupům lze využít všechny atributy dimenzí
  - Jednoduché SQL (Select …group by…order by)



Dimenzionální tabulky

- Každá dimenzionální tabulka obsahuje jeden jednoznačný identifikátor a popisné atributy
- Dimenzionální tabulka je denormalizovaná

D_CAS
ID_Cas (*)
Datum
Rok
Kvartal
Mesic
Den
Víkend
…

D_Produkt
ID_Produkt (*)
Nazev
Kategorie
Subkategorie
…

D_Zakazník
ID_Zakaznik (*)
Jmeno
Pohlavi
Rok narozeni
Země
Kraj
Okres
Obec
…

Dimensionální tabulky

- Atributy v dimenzi:
  - Hierarchické (kategorie – subkategorie – produkt)
  - Nehierarchické (barva_produktu)
- Výstupy většinou kombinují oba typy atributů
- Muže existovat i více hierarchií

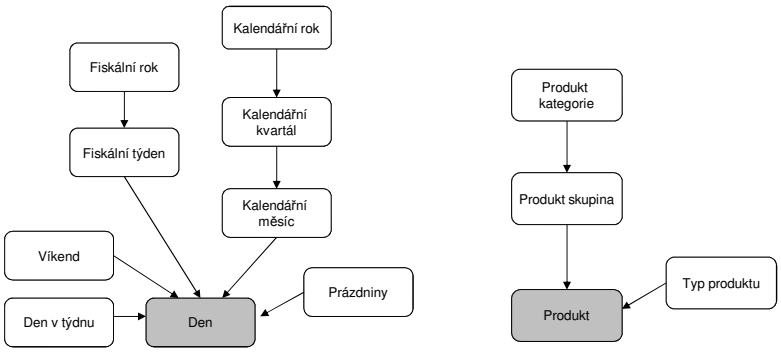
Product dimension	
Product key	
Description	
Marketing brand	Hierarchie 1
Marketing subcategory	
Marketing category	
Financial brand	Hierarchie 2
Financial subcategory	
Financial category	
Flavour	Atributy v žádné hierarchii
Package type	

Hierarchie v dimenzi

- Dimenzionální tabulka v sobě nese různé hierarchie – vztahy 1:M mezi atributy dimenze
- Hierarchie mají vždy stromovou strukturu
- Příklady hierarchie:
  - Rok -> Kvartál -> Měsíc -> Den
  - Rok -> Týden -> Den
  - Země -> Kraj -> Okres -> Obec
  - Kategorie produktu -> Sub kategorie -> Produkt

Hierarchie – zápis

- Způsob zápisu hierarchie v dimenzi:



Dimensionální modelování

- Dimensionální atributy jsou důležité (vstupní brána do DW)
- Mají být
  - Popisné
  - Nepoužívat kódy
  - Bez chybějících (null) hodnot – nahradit např. Neuvedeno
  - Zajištění kvality (bez překlepů, nemožných hodnot, zastaralé, sirotci…)
  - Standardizace (např. standardní zápis adresy)

Výběr dimenzí

- Počet dimenzí by se měl pohybovat okolo 5 až 15
- Méně dimenzí – něco chybí, zda nelze dodat:
  - Kauzální dimenzi (promoce, kontrakt, počasí, podmínky obchodu, …)
  - Další časové dimenze (hlavně u faktových tabulek zachycující položky celku (např. objednávka -> objednané produkty)
  - Dimenzi ve více rolích (např. místo odkud se volalo, kam se volalo)
  - Status dimenzi – označující současný status dimenze nebo snímku (např. nový zákazník)
  - Auditní dimenzi
  - Degenerovanou dimenzi
  - Junk dimenzi
- Přidání dimenzí často nezmění granualitu původní faktové tabulky
  - Lze přidat i do provozujícího datového skladu

Výběr dimenzí

- Více jak 20 až 30 dimenzí ukazuje na možnost jejich spojení
  - Nejsou nezávislé – spojit do jedné
    - Obchodně patří k sobě (např. značka, kategorie, oddělení)
    - Korelace (viz –Junk dimenze)

Faktové tabulky

- Faktové tabulky slouží k uchování informací o sledovaných ukazatelích
- Mezi typické ukazatele lze zařadit:
  - Počet prodaných kusů
  - Hodnotu prodaného zboží (v Kč)
  - Počet zákazníků
  - Počet smluv
  - Výše škody
  - Délka hovoru
- Z hlediska datového se většinou jedná o číselné údaje, které lze agregovat

Faktové tabulky

- Granualita faktové tabulky: míra podrobnosti sledovaných ukazatelů, jejich přesný význam
  - Např.: počet prodaných kusů za den daného zboží v dané prodejně
- Přiřazený dimenzí:
  - Popisy, které nabývají jedné hodnoty pro jeden záznam ve faktové tabulce (s danou granualitou)
  - Při více hodnotách (M:N vztah) řešit přes pomocné tabulky

Typy ukazatelů

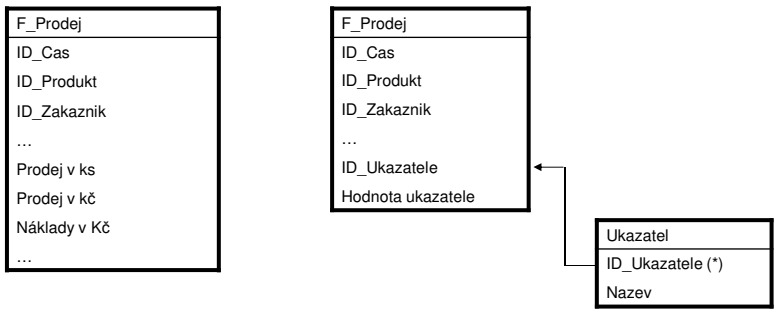
- Ukazatele máme tří typů:
  - Aditivní – agregovatelné (sčítáním) přes všechny dimenze (např. prodej v ks)
  - Semiaditivní – agregovatelné (sčítáním) jen přes některé dimenze (např. stav zásob v ks)
    - Nutné agregovat jinými agregačními funkcemi např. průměr
    - POZOR nelze vždy použít funkci AVG v SQL – problém prázdných období (kdy nebyla transakce)
  - Neaditivní – neagregovatelné (např. některá textová fakta – počasí při nehodě)

Faktová tabulka

- Z datového hlediska:
  - Faktová tabulka obsahuje cizí klíče dimenzionálních tabulek
  - Tyto cizí klíče tvoří primární klíč faktové tabulky
  - Navíc obsahuje faktová tabulka jednotlivé ukazatele
- Kimball's law: Každý vztah M:N je faktová tabulka, z definice

Ukazatele v faktové tabulce

- Existují dva způsoby zápisu ukazatelů do faktové tabulky:



Typy faktových tabulek

- Transakční
  - Zachycuje jednotlivé transakce, jednotlivé akce v daný časový okamžik
  - Zachycuje jen transakce jež se udály (jestliže zákazník nic nekoupil nebude v faktové tabulce)
  - Obvyklý fakt: Množství (v dané transakci)
  - Obvykle se po naplnění dále neprovádí update
  - Ukazuje chování, vývoj v čase
- Snímková
  - Zachycuje stav k určitému časovému okamžiku (periodicky)
  - Většinou měsíční
  - Obvykle existuje jeden záznam pro všechny kombinace významných dimenzí
  - Sledovaná fakta – často složité výpočty, někdy vhodné přebírat z OLTP systémů (již ověřená čísla)
  - Umožňuje efektivně generovat výstupní reporty s často složitě vypočitatelnými ukazateli
    - Není efektivní je generovat přímo z transakcí

Typy faktových tabulek

- Akumulovaná
  - Zachycuje stav v daný okamžik
  - Většinou obsahuje několik časových dimenzí (kdy byl záznam naposledy updatován, datum jednotlivých sledovaných fází)
    - Řada obsahuje „null“ hodnoty, které jsou postupně vyplňovány
    - Potřeba umělých klíčů v časové dimenzi na hodnotu „Dosud neznámo“
  - Dochází k update v faktové tabulce při změně stavu
    - Pro sledovanou událost jeden záznam ve faktové tabulce, který je postupně updatován
  - Vhodná tam kde sledovaná událost má daný čas trvání

Typy faktových tabulek

- Vhodné začít s snímkovou nebo akumulovanou fakt tabulkou
- Postupně lze přidat i transakční
  - Pro pokročilé analýzy chování

Dimensionální modelování

- Čtyři kroky tvorby faktové tabulky
  - Výběr datového tržiště
    - Jeden datový zdroj vs. Více
    - Začít s řešením kde jeden datový zdroj
  - Určení granuality dat
    - Měla by být co nejdetailnější
    - Potřeba přesně vydefinovat, určuje co bude v fakt tabulce uloženo
    - Určení typy faktové tabulky
  - Výběr dimenzí
    - Vychází z určení granuality plus další dimenze vyhovující navržené granualitě
    - Granualita dimenze nemůže být nižší než granualita faktové tabulky
  - Určení faktů (ukazatelů)
    - Vychází z granuality a typu faktové tabulky
    - Ukazatele s rozdílnou granualitou vytvořené např. z důvodu urychlení výpočtu je třeba uložit do zvláštní faktové tabulky (např. součty pro výpočet procent z detailních ukazatelů, …)

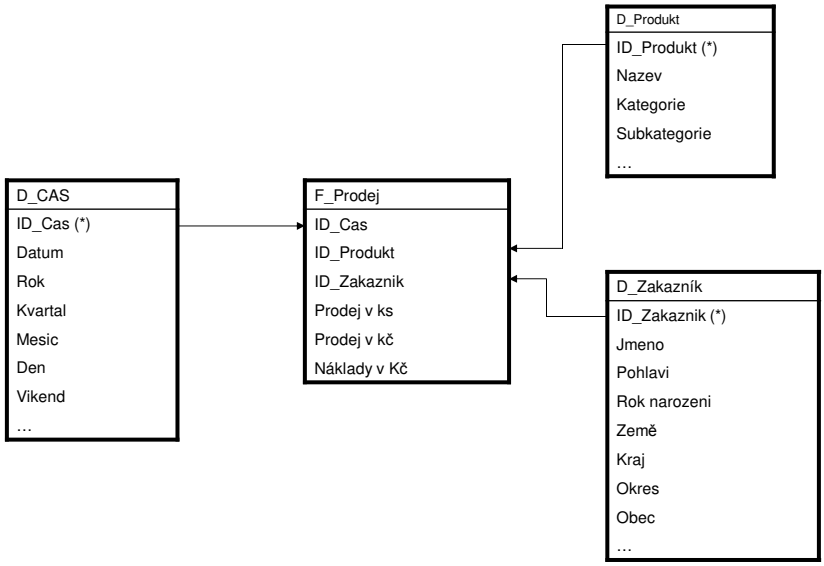
Různá granualita fakt

- Fakta s různou granualitou musí být uloženy v různých faktových tabulkách
- Často je potřeba alokovat fakta s vyšší granualitou na nižší pro možnost srovnaní
  - Např. alokovat náklady objednávky na jednotlivé produkty pro porovnání s výnosy
  - Viz dále Parent-child modelování

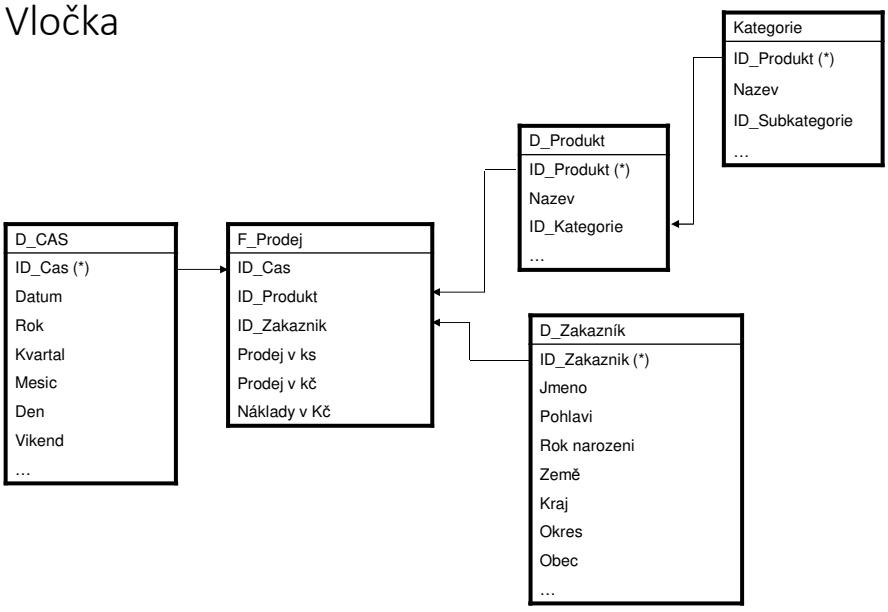
Uspořádání tabulek

- Existují dvě základní schémata uspořádání faktových a dimenzionálních tabulek:
  - Hvězda
  - Sněhová vločka

Hvězda



Vločka



Výhody/Nevýhody uspořádání

- Hvězda
  - Preferované uspořádání
  - Přehlednější uspořádání z hlediska uživatele
  - Snadněji udržovatelé
  - Méně spojení než u vločky
- Vločka
  - Méně přehledné
  - Náročnější na údržbu
  - Nutné pro napojitelnost dalších faktových tabulek (BUS architektura)

Dimensionální modelování

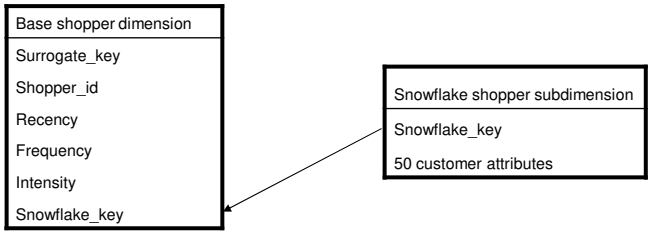
- Doporučuje se používat star schéma
  - Snowflake – není tak srozumitelné uživateli
  - Na druhou stranu je někdy vhodné při napojení jiných faktových tabulek s rozdílnou granualitou
  - Použijeme-li fyzicky snowflake – je vhodné odstínit uživatele pomocí views
  - Snowflake ušetří trochu místa ale většinou zanedbatelné (hlavní velikost je ve faktech)
    - Existují výjimky – extra velké dimenze (např. zakaznici)

Kdy Snowflakes

- Před uživatelem je možné Snowflakes skrýt za pomoci views
  - Fyzický design – řízen rychlostí, efektivností
  - Logický design – řízen uživatelskou přívětivostí, jednoduchostí
- Např. dimenze zákazník
  - Klasický zákazník (20%)
  - Návštěvník www (80%)
- Společné atributy:
  - Shopper surrogate key
  - Shopper ID (fixed ID for each physical shopper)
  - Recency
  - Frequency
- Klasický zákazník
  - Five name attributes
  - 10 location attributes
  - 10 behavior attributes
  - 25 demographic attributes.

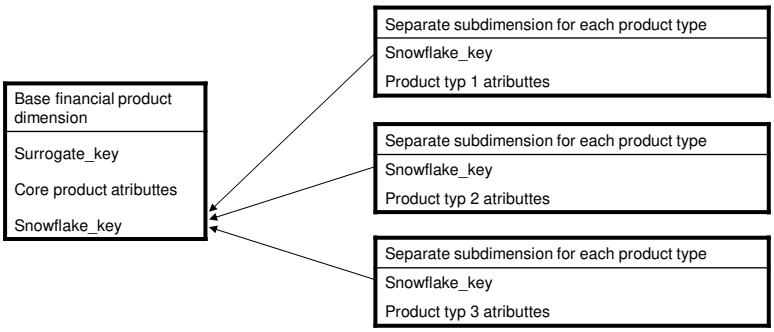
Kdy Snowflakes

- Příklad:



Kdy Snowflakes

- Finanční produkty
  - Mnoho produktů, některé atributy společné, jiné rozdílné
  - Jedna dimenze – mnoho null hodnot v nevyplněných atributech
- Snowflake key lze nahradit umělým klíčem společným přes všechny tabulky (viz unity dimension)
- Uživatele lze odstínit pomocí views



Kdy Snowflakes

- Časová dimenze – různé kalendáře v různých organizacích
- Pozor:
  - Subdimenze má větší kardinality než dimenze
  - Primární klíč subdimenze je snowflake_key a Organization
  - Potřeba specifikovat organizaci při spojení tabulek (pak 1:1)