

Zdroje

- <https://is.cuni.cz/studium/predmety/index.php?tid=&do=download&did=35279&kod=NDBI027>
- <http://forum.matfyz.info/viewforum.php?f=472>

1. Modelování

1.1. Příklad 1 z přednášky:

**Sledujte ukazatele:**  
počet zapsaných studentů, počet pokusů, známka

**Podle:**  
Ročník  
Semestr  
Předmět  
Učitel  
Zápočet  
Zkouška

Jak budou vypadat dimense a hierarchie?

**Řešení:**

**1. Metriky:** Co budeme měřit

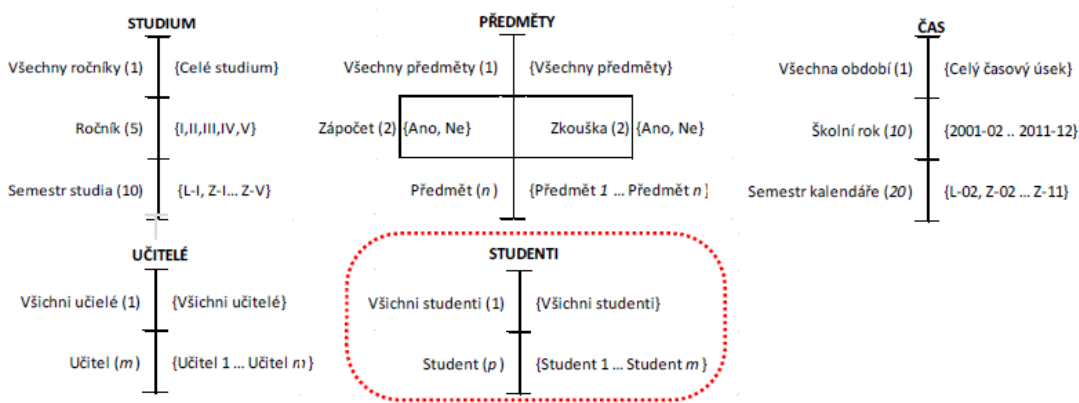
Metrika	Typ metrik
Počet zapsaných studentů	Semiaditivní
Počet pokusů	Aditivní
Známka	Neaditivní

Pozn.:

**Počet zapsaných pokusů:** Dle definice může mít různý význam:  
a) Počet studentů zapsaných na předmět  
b) Počet studentů zapsaných na předmět v jednom semestru  
• V tomto případě by to mohla být plně aditivní metrika  
• Záleží ovšem ještě na hierarchiích (viz dále)

**Známka:** Uložená hodnota vs. presentovaná metrika  
• Uložená hodnota: atomická (co to znamená ... viz dále)  
• Presentovaná metrika: agregace (průměr, medián,...)  
• Počítat ze všech pokusů nebo jen z úspěšných?  
• Co když zapsaný student na zkoušku vůbec nešel?

**2. Dimense:** – podle čeho budeme sledovat



Pozn.: Dimense STUDENTI přidána, aby bylo možno sledovat průměrné známky

- Obsahuje dostatečně atomické elementy (Student)
- Pravděpodobně existují odpovídající data (známky)
- Pokud bychom nepoužili dimensi STUDENTI, museli bychom ukládat průměry známek pro kombinaci:

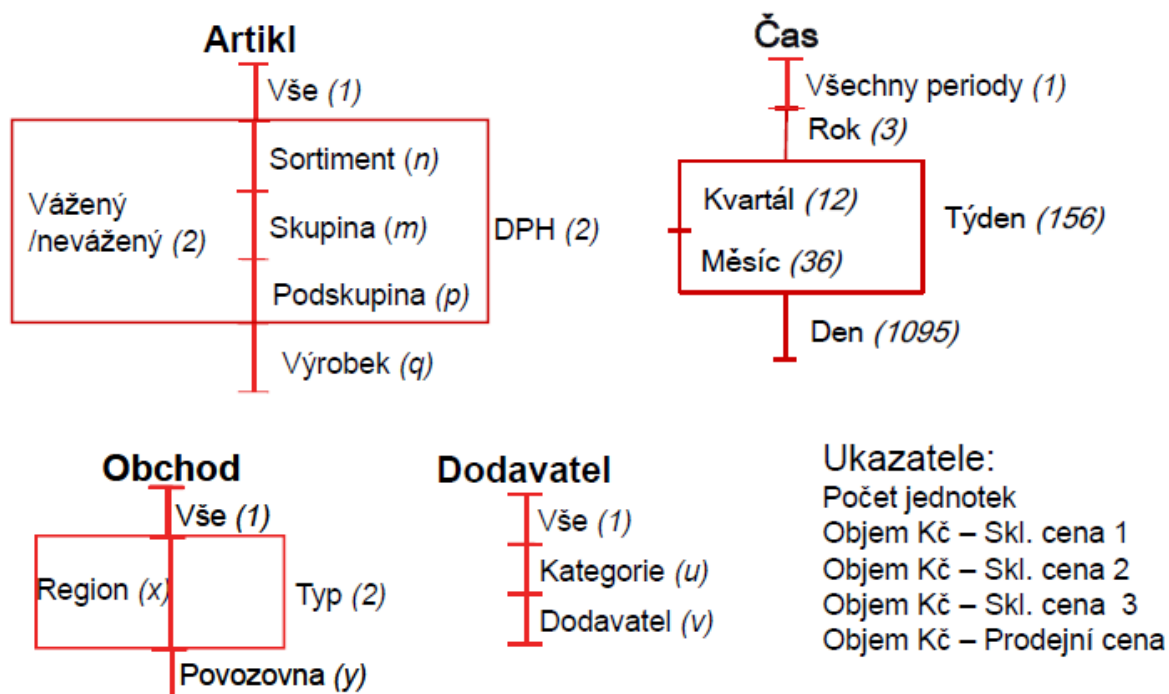
Semestr\_studia x Předmět x Semestr\_kalendáře x Učitel + počet pokusů, ze kterých se průměr počítá

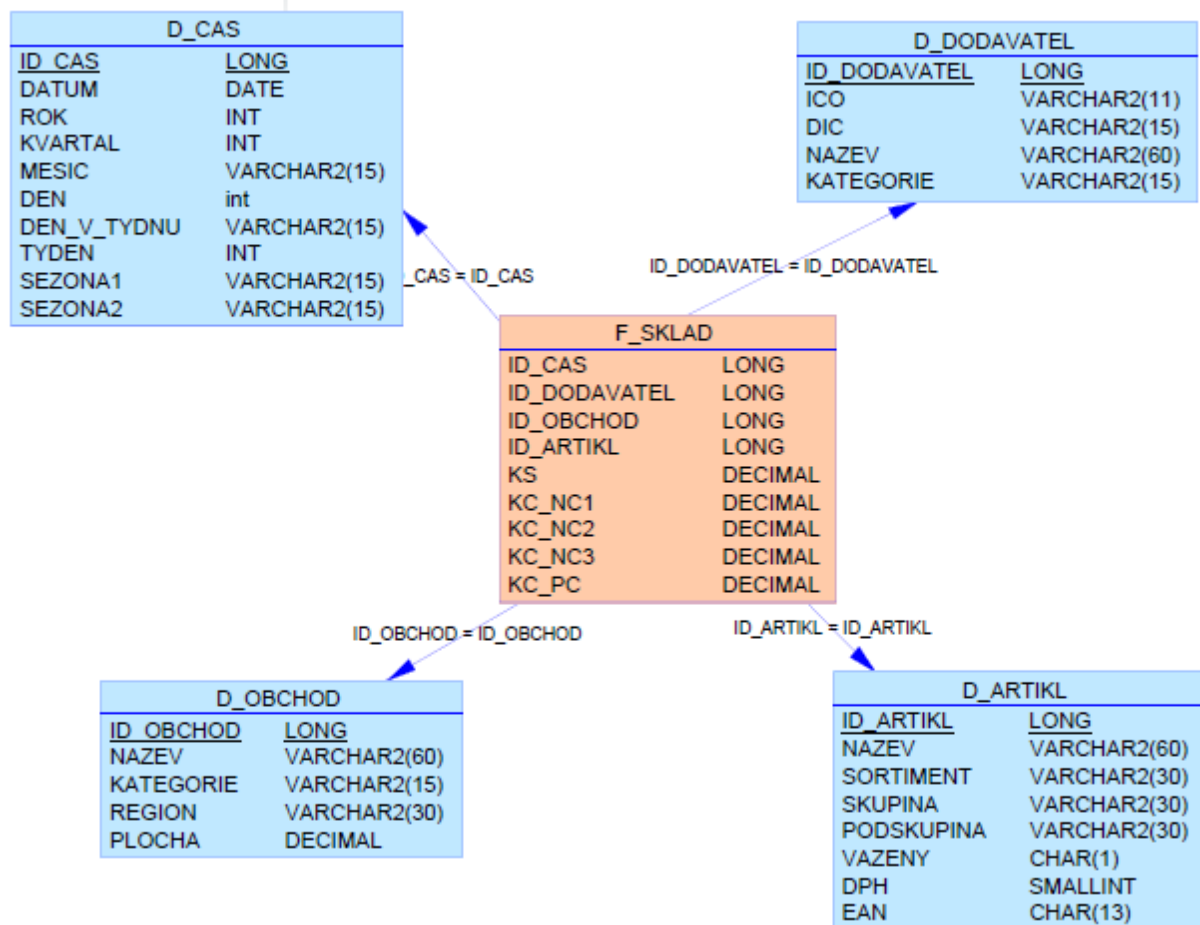
- Výpočet by byl v rámci ETL

1.2. Příklad 3 z přednášky:

- **Proces:** Sledování stavu skladu
- **Granularita:** Denní snímky
- **Dimenze:**
  - čas (den, ...)
  - region->obchod, sklad
  - kategorie->dodavatel
  - sortiment->skupina->podskupina->artikl
  - sazba DPH
  - vážený/nevážený
  - trvanlivost (datum)
- **Ukazatele:**
  - ks (měrných jednotek)
  - Kč (skladová cena 1, skladová cena 2,...)
  - Kč (prodejní cena)

**Řešení:**





### 1.3. Cvičení 1 z přednášky:

- Při analytických interview v obchodním řetězci jste se dověděli tyto informace:
  - řetězec má 120 prodejen
  - řetězec prodává 20 000 artiklů členěných na úseky, rodiny, podrodiny, artikl
  - nákupčí nakupuje více artiklů, pro jeden artikl existuje vždy jeden hlavní nákupčí
  - dodavatel dodává více artiklů, pro jeden artikl existuje vždy jeden hlavní dodavatel
- v rámci první etapy byly identifikovány tyto informační procesy:
  - sledování prodeje, marže, zisku
  - porovnání skutečného prodeje s plánem prodeje
  - sledování výkonnostních ukazatelů: Prodej na zaměstnance, zisk na zaměstnance
- Definujte jednotlivé procesy, jejich granularitu, dimenzionalitu a ukazatele
- Nakreslete MDS diagram multidimenzionálního prostoru

### Řešení:

- **Proces 1:** Sledování prodeje, marže, zisku
  - **Ukazatele:**
    - Prodej v jednotkách
    - Prodej v realizované prodejní ceně (PR\_RPC)
    - Prodej v teoretické prodejní ceně (PR\_TPC)
    - Prodej v nákladové ceně (PR\_NC)
    - Zisk = PR\_RPC – PR\_NC
    - Marže = Zisk / P\_RPC
  - **Časová granularita:** Denní snímky
  - **Dimense:**
    - ČAS: Rok → Den
    - GEOGRAFIE: Region → Obchod
    - DODAVATEL: Hlavní dodavatel → Dodavatel/Subdodavatel
    - NÁKUPČÍ: Hlavní nákupčí → Nákupčí

- PRODUKT: Úsek → Rodina → Podrodina → Artikel
- PROMOCIE: Typ promoce → Promoce
- **Proces 2:** Porovnání prodeje s plánem prodeje
  - **Ukazatele:**
    - Prodej v jednotkách
    - Prodej v realizované prodejní ceně (PR\_RPC)
    - Prodej v teoretické prodejní ceně (PR\_TPC)
    - Plán v realizované prodejní ceně (PL\_RPC)
    - Plnění plánu (v %) =  $PR\_RPC / PL\_PLC$
  - **Granularita:** Týdenní/Kvartální snímky obchod, rodina produktů
  - **Dimenze:**
    - ČAS: ROK → týden
    - GEOGRAFIE: Region → Obchod
    - PRODUKT: Úsek → Rodina
    - Verze plánu
- **Proces 3:** Sledování výkonnostních ukazatelů
  - **Ukazatele:**
    - Prodej v jednotkách
    - Prodej v realizované prodejní ceně (PR\_RPC)
    - Zisk (viz Analýza (1))
    - Počet zaměstnanců (PZ)
    - Prodej na zaměstnance =  $PR\_RPC / PZ$
    - Zisk na zaměstnance =  $Zisk / PZ$ .
  - **Granularita:** Týdenní snímky
  - **Dimenze:**
    - ČAS: Rok → týden
    - GEOGRAFIE: Region → Obchod
- MD model (MDS zápis)

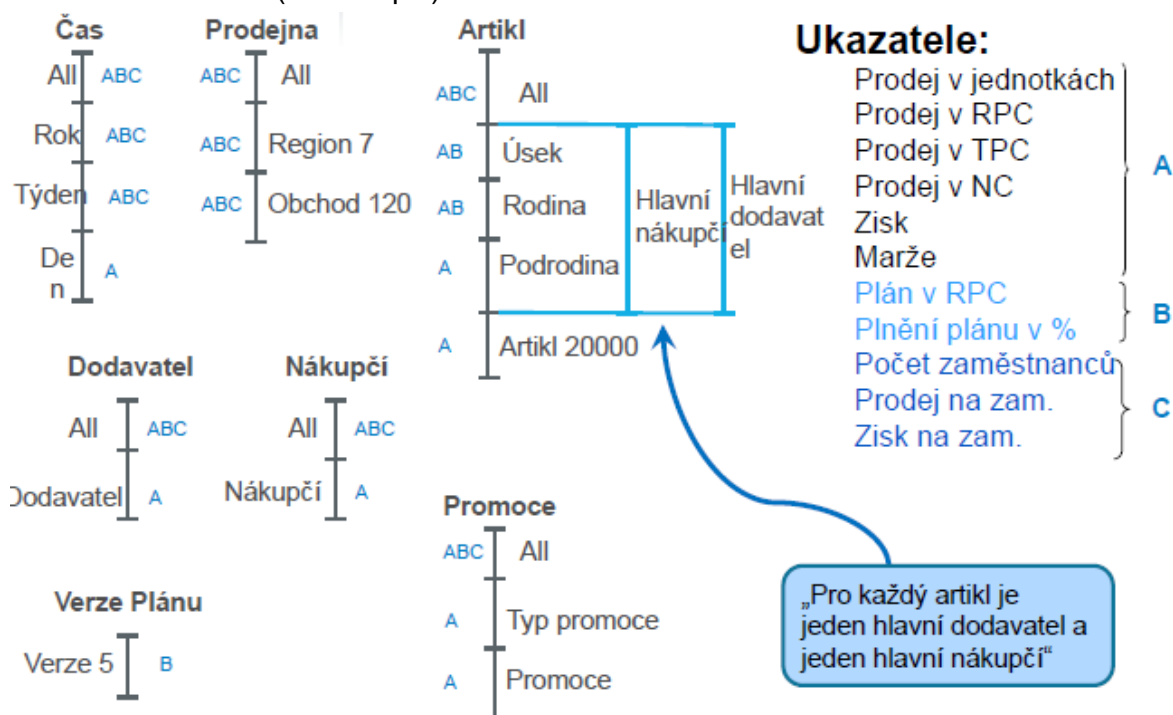
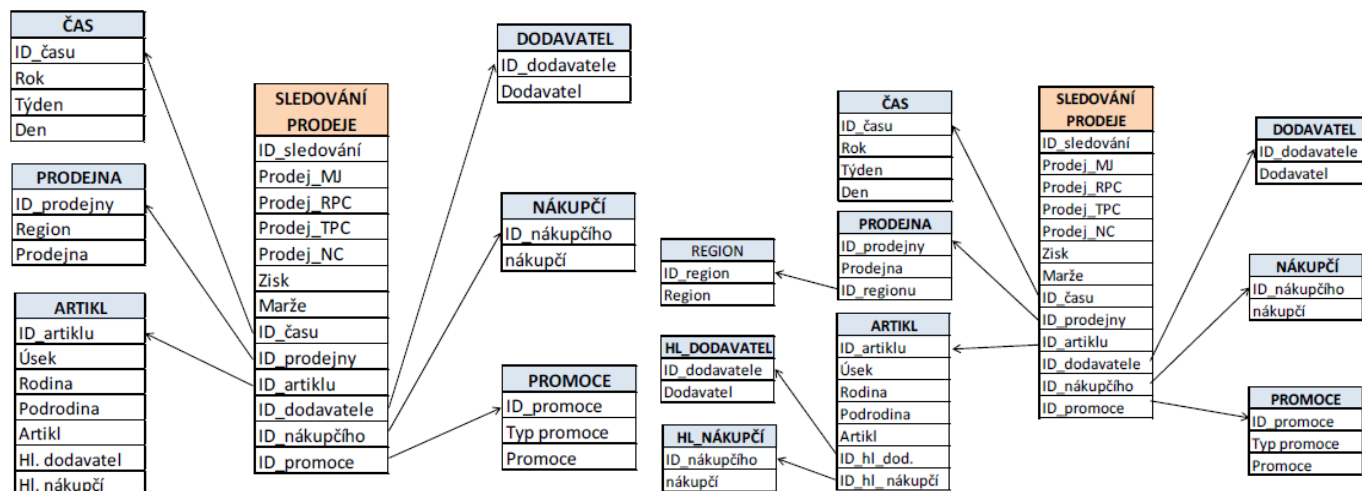


Schéma (hvězda a vložka):



1.4. Navrhněte multidimensionální model, který umožní sledovat výsledky zkoušek na vaší fakultě podle následujícího zadání:

- V čase po semestrech a rocích
- Podle ročníků
- Podle předmětů
- Podle vyučujících
- Podle místa bydliště studentů (země, region)

Zajímají nás následující ukazatele:

- Průměrná známka
- Počet pokusů na studenta
- Počet pokusů na předmět
- Počet pokusů na učitele

**Řešení** (pracovní verze):

**Postup**

- Stanovení metrik
- Stanovení dimensí
- Stanovení hierarchií dimensí
- Mapování metrik na dimense

-Hierarchie vs, agregace (slidy 4, str 16)

Stanovení metrik -> je ze zadání:

**Ukazatele/metriky** (bude potřeba upřesnit, co se tím myslí):

- Průměrná známka                      neaditivní
- Počet pokusů na studenta            aditivní
- Počet pokusů na předmět            záleží co to znamená, pokud za semestr, tak lze vnímat jako aditivní
- Počet pokusů na učitele              totéž

->budou nás zajímat třeba průměry za nějaký čas, to by dávalo smysl

surovými daty tedy budou:

student-předmět-učitel-známka-kolikátý pokus

->chce to dimenzi student

Stanovení dimensí jsme opět dostali v zadání (první seznam).

**Hierarchy**

záleží, jestli můžeme předpokládat nějakou vazbu předmětů a vyučujících, pokud by jeden předmět učil jen jeden vyučující, je to hierarchie.

semestry a roky jsou hierarchické

ročníky bych do toho spíš nemotala - leda každý semestr ještě rozdělit na současní prváci, současní

druháci, současní.... a to je otázka, jestli je k něčemu.  
student může být hierarchicky pod svou geografickou lokací

### Dimense:

- Čas
  - Rok, semestr
- Ročník
- Předmět
  - Vyučující, předmět
- Student
  - Země, region, student

písmena jsou která metrika je sledována v dané dimenzi - příklad ze slidů  
číslo je, kolik takových věcí rozlišujeme.

### ▪ MDS - Multidimensional Domain Structure (Erik Thomsen)

#### Metriky:

- (a) Prodej v ks
- (b) Zásoba v ks
- (c) Počet zaměstnanců
- (d) Průměrný plat

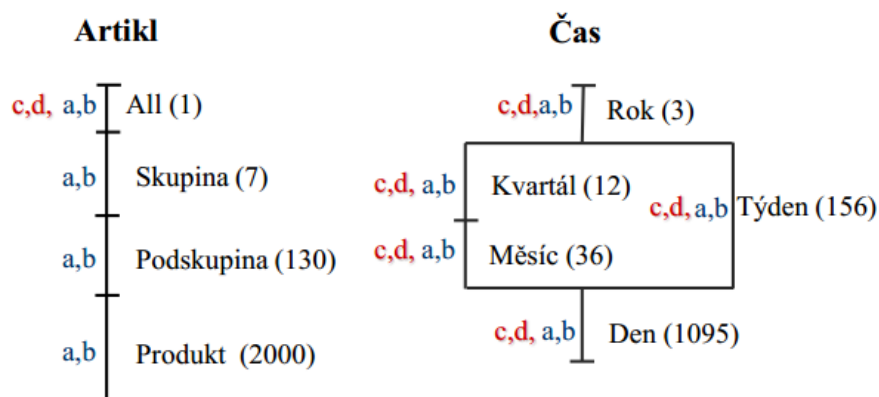
#### Dimense:

Artikl

- Produkt, podskupina, skupina

Čas

- Rok, kvartál, měsíc, týden, den



**Otázka:** Mohou být metriky sledovány i pro dimenzi Artikl?  
Za jakých okolností?  
Co pak metriky vyjadřují?

jeden hezký příklad slidy 4 od strany 26 je hotový rozbor....

1.5. Navrhněte databázové schema pro model z předchozího příkladu

1.6. Navrhněte hvězdicové schema pro model z předchozího příkladu

1.7. Navrhněte multidimensionální model, který analysovat výsledky fungování ZOO dle následujícího zadání

Sledovat ukazatele

- a) Spotřebované krmivo v kg
- b) Počet chovaných živočichů
- c) počet ošetřovatelů
- d) Náklady
- e) Tržby ze vstupného
- f) Průměrný věk chovaných živočichů

Podle následujících kategorií

- a) Živočišný druh, rod, čeleď, řád
- b) Čas - roky, měsíce, dny v týdnu, sezóny
- c) Stáří chovaných živočichů
- d) Údaje o živočichu (pohlaví, váha, barva, nebezpečnost, typ stravy - maso/zelenina/vše, spotřeba v kg)

**Řešení:**

### Ukazatele/metriky:

- |                                    |              |
|------------------------------------|--------------|
| • Spotřebované krmivo v kg         | aditivní     |
| • Počet chovaných živočichů        | semiaditivní |
| • počet ošetřovatelů               | semiaditivní |
| • Náklady                          | aditivní     |
| • Tržby ze vstupného               | aditivní     |
| • Průměrný věk chovaných živočichů | neaditivní   |

**Dimense:**

- Živočich
  - Živočišný druh, rod, čeleď, řád, živočich(pohlaví, váha, barva, nebezpečnost, typ stravy - maso/zelenina/vše, spotřeba v kg, stáří)
- Čas
  - roky, sezóny, měsíce, dny v týdnu

1.8. Navrhněte databázové schema pro model z předchozího příkladu

1.9. Navrhnete Multidimensionální model dle zadání níže. Znázorněte např. pomocí Multidimensional Domain Structure (Dle Thomsena). Zkuste se krátce zamyslet nad tím, jak to udělat se sledováním objemu prodeje v různých zemích a naznačit řešení

## Proces: Sledování prodejů

### Granularita: Denní snímky

### Dimensione:

- čas (rok, kvartál, měsíc, den, den v týdnu)
- region->země, region, pobočka
- sortiment->skupina->podskupina->artikl
- sazba DPH

**Ukazatele:**

- Počet prodaných výrobků (ks)
- objem prodeje (Kč, Eu)

**Řešení:**

### Ukazatele/metriky:

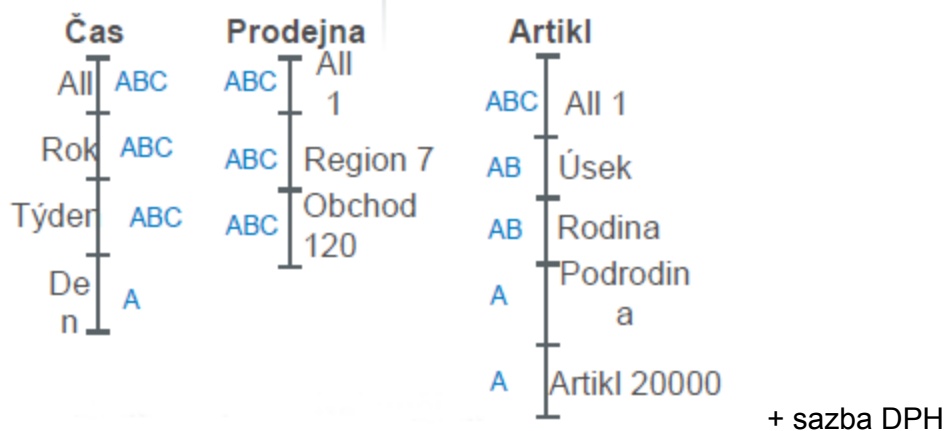
- Počet prodaných výrobků (ks) aditivní
- objem prodeje (Kč, Eu) aditivní

**Dimense:**

- Čas
  - rok, kvartál, měsíc, den, den v týdnu
- Pobočka
  - země, region, pobočka
- Artikl
  - skupina->podskupina->artikl
  - sazba DPH

**MDS:**

namalujte něco ve stylu



1.10. Navrhněte DB model pro předchozí příklad - Hvězdicové nebo vločkové schema - tabulky, sloupce, vazby.

1.11. Řešíte úlohu čištění adres. Na vstupu máte sloupce: psč, obec, ulice, číslo orientační a číslo popisné. Navrhněte metriky datové kvality (alespoň 10)

- Metriky kvality atributu PSČ
  - Počet záznamů se správně vyplněným PSČ
  - Počet záznamů s nevyplněným PSČ
  - Počet záznamů s nesprávným formátem PSČ
  - Počet záznamů s PSČ, které nelze dohledat v externím zdroji
  - Počet záznamů, kde PSČ v externím zdroji neodpovídá názvu obce
  - Počet záznamů, pro něž lze dohledat PSČ podle názvu obce
  - Počet případů, kdy evidentně různé obce mají stejné PSČ
- Metriky kvality atributu Obec
  - Počet záznamů s nevyplněným Obec
  - Počet záznamů s Obec, které nelze dohledat v externím zdroji
- Metriky kvality atributu Číslo popisné
  - Počet záznamů s nevyplněným Číslo popisné



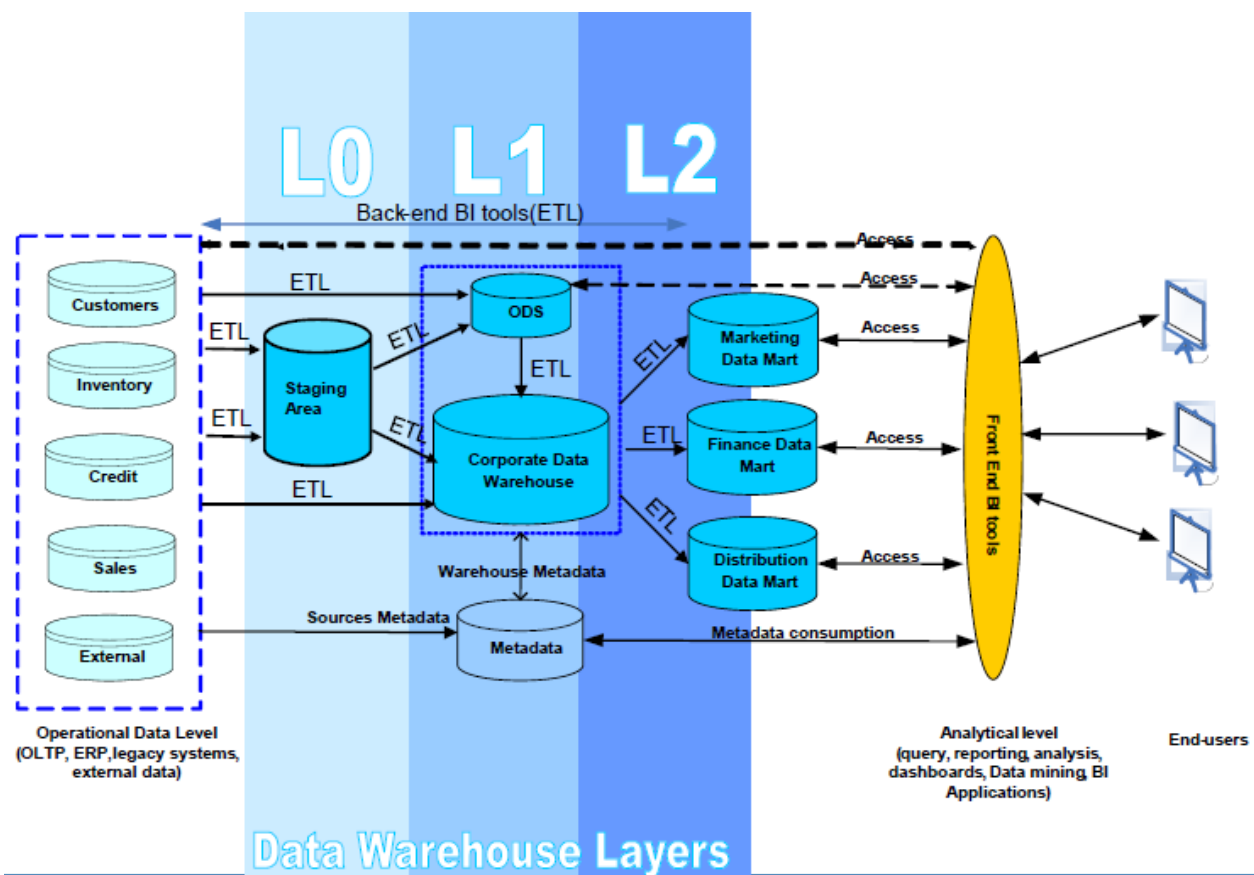
## 2. Architektura a modelování

### 2.1. Jaké jsou rysy, v nichž se typicky liší [DW](#) a [OLTP](#)

- DW (Data Warehouse) - je zvláštní typ relační databáze, která umožňuje řešit úlohy zaměřené převážně na analytické dotazování nad rozsáhlými soubory dat.
- OLTP (Online transaction processing) - je technologie uložení dat v databázi, která umožňuje jejich co nejsnadnější a nejbezpečnější modifikaci v mnohauživatelském prostředí. Jedná se o přístup používaný v současné době v převážné většině databázových aplikací. Jako protiklad k OLTP je především k analytickým účelům nad rozsáhlými databázemi používána technologie [OLAP](#).
- Aplikační vs. Subjektová orientace
- Detail vs. Agregace
- Časové diskrétní hodnoty vs. Snímky za období
- Uživatelé: Řadoví úředníci vs. Manageři
- Přístup: Update vs. Read
- Opakující se vs. Heuristické zpracování
- Požadavky: OLTP - předem známy, DW - většina předem neznáma
- Standardní životní cyklus vs. Specifický životní cyklus DW
- Přístup k informacím v jednom okamžiku: Jednotka informace vs. Sada informací
- Transakční orientace vs. Orientace na analýzu
- Požadavky na výkon: pro OLTP životně důležitá, pro DW ne tolik
- Přístupová práva pro update: Pro OLTP rozhodující, pro DW nezajímavé
- High availability: pro OLTP důležitá, pro DW ne
- Správa a používání: OLTP jako celek, DW - subzety
- Redundance: OLTP - nežádoucí, DW - typická
- Struktura: OLTP - statická, DW - flexibilní
- Zpracovávané objemy dat: malé vs. velké
- Operace: OLTP - rutinní úlohy, DW - manažerské potřeby
- Pravděpodobnost přístupu ke konkrétním datům: OLTP - vysoká, DW - nízká až střední

### 2.2. Jaké jsou základní logické vrstvy datového skladu - stručně popište. (●)

- **L0 (Data Staging Area)**
  - „Nárazník“ mezi operativním a analytickým světem
  - autonomní tabulky, ne kontroly, ne referenční integrita
- **L1 (Integrated Layer)**
  - „RELAČNÍ VRSTVA“
  - konsolidovaný datový sklad
  - bez faktových a dimenzních tabulek
  - 2.-3 NF, referenční integrita, integrovaná, konsolidovaná data, historie, subjektově orientovaná
- **L2 (Presentation Layer)**
  - presentační vrstva
  - multidimensionální model, určeno pro analýzu



- 2.3. Jaký je rozdíl mezi Staging areou a L0? (●)
- vrstva L0 obsahuje Staging areu
- 2.4. Jaký je rozdíl mezi Staging areou a ODS? (●)
- 2.5. Co je ODS a k čemu slouží?
- **Staging Area** (ve vrstvě L0)
    - Nárazník mezi operativními systémy a DW nebo i ODS
    - Konsolidace dat, Vyrovnávání časového nesouladu
    - Persistence (ukládání) extraktů
    - Transformační operace
  - **ODS** se používá až v L1 (Integrated Layer)
    - **Operational Data Store** - db pro integraci dat z různých zdrojů pro další operace na datech, často orientace na subjekt (např. zákaznická ODS)
    - Data se pak předávají (příp. transformují) do DW pro reporting.
    - Aktuální data, menší objem
- 2.6. K čemu je dobrá ve **Staging area** referenční integrita? (●)
- RI zajišťuje pořádek v databázi a v její celé historii
  - Povinnost vazeb významně zjednodušuje čtení při analýsách a reportingu
    - (můžeme se vyhnout ošetřování NULL (IS NULL, IS NOT NULL))
- 2.7. Co jsou **Datamarty**, k čemu slouží a proč vznikají? (●)
- “rozhraní” pro DW - DM obsahují požadovanou podmnožinu dat určitého DWH, ale jsou rychlejší a flexibilnější (co do změny údajů - navržen spíše pro rychlost než flexibilitu)
  - Můžou být nezávislé (data DWH se duplikují na Data marte), alebo závislé (neduplikují data, používají data z DWH)
  - pro subjekty co nemají zájem o celý DWH a chtějí rychlejší a flexibilnější systém, do kterého může naraz přistupovat víc uživatelů
  - Menší množství dat, časté sumarizace
- 2.8. Co jsou to **kreativní indexy**? Stručně popište a příklad (●)
- Optimalizace modelu (Denormalizace)
  - Exitující systémy → Extrakce:
    - mírně sumarizovaná data
    - vysoce sumarizovaná data
    - kreativní indexy, profily

- Příklady kreativních indexů:

- největší zakázky
- nejméně aktivní účty
- nejpozdější dodávky

2.9. Co znamená v DW **partitioning** a k čemu je to dobré? Stručně popište (●)

- Rozdělení dat na oddělené fyzické jednotky
- K čemu je to dobré
  - s daty se může zacházet odděleně -> vyšší výkon
- Data ve velkém bloku se nedají:
  - snadno restruktuálnízovat
  - libovolně indexovat a v případě potřeby sekvenčně prohledávat
  - jednoduše reorganizácia
  - snadno obnovovat a aktualizovat
  - jednoduše monitorovat

2.10. Typy faktů

- **Aditivní**
  - Agregace přes všechny dimenze, např.: objem prodeje
- **Neaditivní**
  - Nelze agregovat, např: popis, jakost
- **Semiaditivní**
  - Typicky aditivní pro všechny dimenze mimo času
  - např. stav účtu, objem zásob
  - pro vazbu na čas se používá jiná agregace (průměr,min,max)

2.11. Uveďte příklad semiaditivního ukazatele.

- Stav skladu (ks n. Kč) je ukazatel, který nelze sčítat v čase - tzv. semiaditivní

2.12. Jaké jsou základní **typy pomalu se měnících dimenzí** a čím se liší? (●)

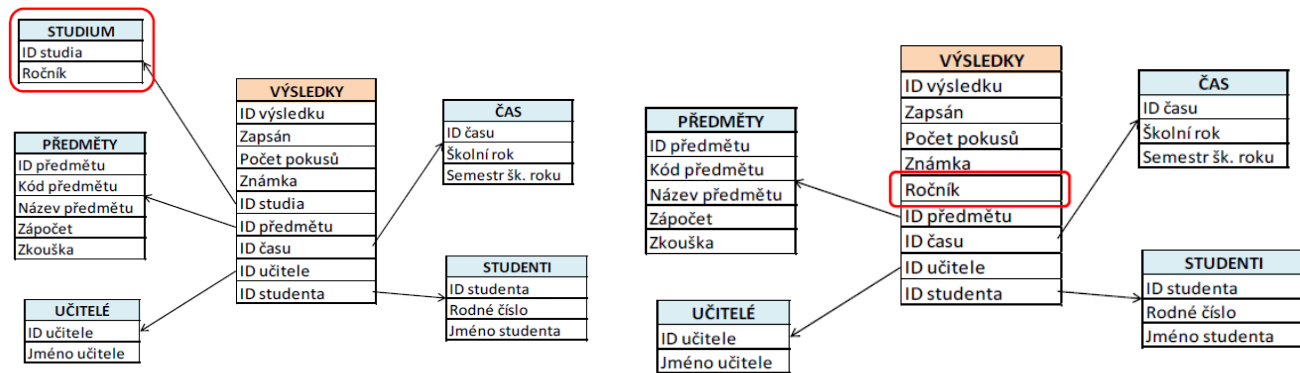
- typopické pro dimenze Produkty, Zákazníci
- **Type 1** - Přepsání záznamu novou hodnotou
  - Ošetření změn: Přidání nových záznamů a update existujících záznamů v případě změn
  - Historie: Žádná
- **Type 2** - Vytvoření nového záznamu s novým umělým klíčem
  - Ošetření změn: Přidání nových záznamů a verzování změn
  - Historie: Plná (verzováním, platností záznamu od-do, atribut pro akt. záznam)
- **Type 3** - Vytvoření sloupce „old“ pro atributy co chceme sledovat předchozí stav
  - Ošetření změn: Přidání nových záznamů a uchování současné a předchozí hodnoty v případě změny
  - Historie: Částečná

2.13. Jaký je rozdíl mezi minidimensí a subdimensí?

- Minidimenze: skupina atributů je oddělena do samostatné tabulky, kde každý řádek představuje unikátní kombinaci hodnot (ppt04/slide 51)
- Subdimenze: vypadají jako snowflake, ale mají odlišnou charakteristiku (ppt04/slide 49)
- Minidimenze je obdoba subdimenze
- Minidimenze má vazbu na tabulku faktů na rozdíl od subdimenze, která se váže na dimenzionální tabulku

2.14. Co je **degenerovaná dimenze**? (●)

- Obvykle představují pouze záznam v tabulce faktů
- Většinou bez vazby na další tabulky
- Nepoužívají se umělé klíče, ale produkční (do faktové tabulky je přímo vloženo např. číslo objednávky)
- Použití:
  - pro seskupování položek patřících do jednoho kontejneru (např. objednávky)
  - pro vazbu do produkčních systémů
- **Příklad** (dimenze studium zdegenerovala):



### 3. Data-Mining (Dobývání znalostí/Dolování dat)

#### 3.1. Co je to data-mining (DM)?

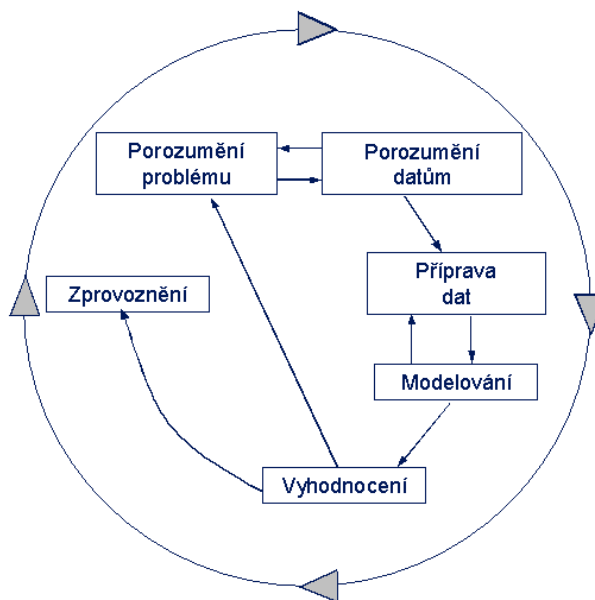
- Je to proces extrahování skrytých vzorů z dat. So zvyšujícím sa obsahom dát (zdvojnásobenie raz za 3 roky) slúži ako nástroj, ktorý tieto data transformuje na informácie

#### 3.2. Jaké jsou hlavní kroky Data Miningu a co se v nich děje

#### 3.3. Jaké jsou základní fáze typického DM projektu? (●)

#### 3.4. Vyjmenujte fáze **CRISP-DM** (●)

- Porozumění problému - Co řešit (Business understanding)
- Porozumění datům - Kde vzít data (Data understanding)
- Příprava dat - Jak data připravit (Data preparation)
- Modelování - Jak data analyzovat (Data modelling)
- Vyhodnocení - Co jsme zjistili (Evaluation)
- Zprovoznění - Jak výsledky využít (Deployment)



- **Zajímavosti:** dnes prakticky standard
  - používá např. systém SPSS Modeller (dříve Clementine)
  - vyvinulo konsorzium: SPSS, NCR(Teradata), Daimler AG, OHRA(pojistovna)

#### 3.5. Popište jednu vybranou fázi v detailu, vyjmenujte její hlavní rizika.

- **Deployment** - Upravit získané znalosti do podoby použitelné pro zákazníka

- Implementace klasifikačního algoritmu v user-friendly podobě
- Příprava uživatelského manuálu
- Instalace programu na pobočkách banky a zaškolení uživatelů
- Změna metodiky poskytování úvěrů a příslušná změna vnitřních předpisů banky
- Rizika:
  - Zákazník musí pochopit, co je třeba učinit pro efektivní využití dosažených výsledků!
  - špatná implementace

#### 3.6. Vyberte jeden DM algoritmus a popište jeho princip (ne detaily, ale princip).

3.7. Popište stručně základní princip **rozhodovacího stromu** (●)

- Kořen – všechny záznamy
- Uzel se dělí dle podmínky na hodnoty atributů na své syny
- Ideál – listy jsou „čisté“, tj. obsahují jen záznamy jedné třídy
- Cesta kořen -> list odpovídá rozhodovacímu pravidlu
- Blíže ke kořeni se štěpí podle významnějších atributů
- Učení s učitelem
- Výhody
  - Akceptují chybějící hodnoty
  - Akceptují spojité i diskrétní hodnoty
  - Snadná transformace na rozhodovací pravidla
  - Dá se použít jako výborný prostředek na zjištění nejdůležitějších proměnných
  - Mohou být interaktivní nebo se generovat celé na základě určitých stop kritérií
- Nevýhody
  - Potíže s jinými regiony než obdélníkovitými
  - Mohou být příliš velké pro rozumné využití

3.8. Vyberte si algoritmus (jiný než v předchozím bodě) a popište jeho výhody a nevýhody.

3.9. Jaké jsou výhody a nevýhody **neuronových sítí** (●)

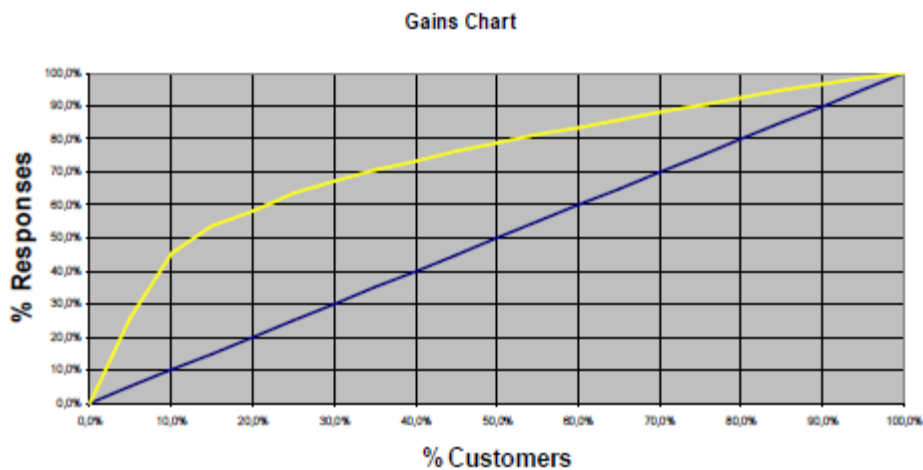
- Výhody
  - Velmi dobré zejména při predikci spojitých atributů
  - Možnost použít pro mnoho typů problémů
  - Dobré výsledky i ve složitých doménách
  - Schopnost přirozeně řešit nelineární vazby mezi vstupy a výstupy
- Nevýhody
  - Nejasná interpretace modelu
  - Možnost konvergovat k lokálnímu minimu
  - Výpočetně náročnější
  - Vstupní proměnné jsou spojité
  - Nutná příprava dat – škálovatelnost
  - Není jednoznačná vazba na významnost proměnných
  - Obecně neumí pracovat s chybějícími hodnotami

3.10. Jaký je rozdíl v učení s učitelem a bez učitele? K čemu je co dobré? (●)

- Model se učí na základě historických dat
- Učení s učitelem
  - V historických datech je obsažen i výstup modelu (cílová proměnná)
- Učení „bez učitele“
  - V historických datech není obsažena cílová proměnná (definuje ji až model)
  - Typickým příkladem je segmentace

3.11. Co je **lift** modelu? (●)

- měří se jím kvalita modelu
- v předmětu Dobývání znalostí se to překládá jako „zlepšení“
  - $\text{zlepšení}(\text{pravidlo}) = \frac{p(i \& j)}{p(i)p(j)} = \frac{\#i \& j}{\#i \#j}$
  - Pravidla: IF Položka\_i THEN Položka\_j
  - #i znamená #transakcí\_obsahujících\_i
- srovnává se výsledek modelu (pravidel) proti např. náhodnému výběru nebo průměru
- Lift je třeba brát na adekvátním percentilu, např. 10%
  - Typicky lift dosahuje hodnot mezi 2,5 a 3,5 na zmíněném 10% kvantilu
- Lift chart (kategorická proměnná)



3.12. Co je overfitting (**přeučení**)? Čím je způsobeno a jak mu zabránit? (●)

- Přeučení modelu u data mining-u
- Naučený model je příliš svázan s trénovacími daty
- Přesnost modelu je vysoká na trénovacích datech, ale nízká na nových datech
- Jak mu zabránit
  - Rozdělení trénovacích dat (učení – test)
  - Rozhodovací stromy – prořezávání, menší hloubka stromu
    - Některé algoritmy ukončí včas generování stromu (prepruning)
    - Většina nejdříve vygeneruje strom a pak ho ořeže (postpruning)
    - Prořezávání zvyšuje chybu na učicích množině, ale doufáme, že na reálných datech chybu zmenší

3.13. Jaké uplatnění nachází DM v bankách? Z business pohledu.

- retence
- cílený marketing
- detekce podvodu
- credit risk
- money laundering
- segmentace klientu

3.14. Jaké uplatnění nachází DM v telekomunikacích? Z business pohledu.

- segmentace klientu
- cílený marketing
- detekce podvodu
- credit risk
- analýza obchodu

3.15. Jaké znáte úlohy DM? Z technického pohledu.

- Klasifikace (Jaké je riziko dané žádosti o úvěr?)
- Odhadování (Jaká je dlouhodobá hodnota zákazníka?)
- Predikce (O které zákazníky můžeme v nejbližších 6 měsících přijít?)
- Analýza nákupního košíku (Které produkty se nakupují společně?)
- Shlukování (Jaká skupina klientů se chová podobně - segmentace?)
- Deskripce (Jaké jsou závislosti údajů v komplikovaných datech?)

3.16. Jaký formát dat je typicky vyžadován do SW pro data mining (pro algoritmy)?

- Typicky prvky n-dimenzionálního vektorového prostoru nad N nebo R.

## 4. Technologie

4.1. Co je **OLAP**, k čemu je to dobré a čím se vyznačuje? (●)

- **On-line Analytical Processing** - je technologie uložení dat v databázi, která umožňuje uspořádat velké objemy dat tak, aby byla data přístupná a srozumitelná uživatelům zabývajícím se analýzou obchodních trendů a výsledků (BI). Způsob uložení dat se svým zaměřením liší od běžněji užívaného **OLTP**, kde je důraz kladen především na snadné a bezpečné ukládání změn v datech v konkurenčním (víceuživatelském) prostředí.
- K čemu dobré:
  - Uživatel má možnost formulovat hypotézy

- Systém poskytuje nástroje pro jejich ověření
- Čím se vyznačuje
  - Základem je zobrazování multidimenzionální matice (kostky)
  - Technické řešení OLAP
    - ROLAP, MOLAP, HOLAP, DOLAP (vid' otázka č. 22)

4.2. Co je **FASMI** a co to znamená?

- FASMI je alternativní termín pro OLAP
- Charakteristika OLAP:

- **F**ast - nesmí záviset na množství dat (do 30sec)
- **A**nalysis - druhotné zpracování (musí být dost. flexibilní)
- **S**hared - konzistence, bezpečnost dat
- **M**ultidimensional - multidimenzionální model
- **I**nformation - zaměřený na informace (pro business users)

4.3. Jaké jsou technické typy řešení OLAP - stručně charakterizujte, schematicky znázorněte, uveďte výhody a nevýhody

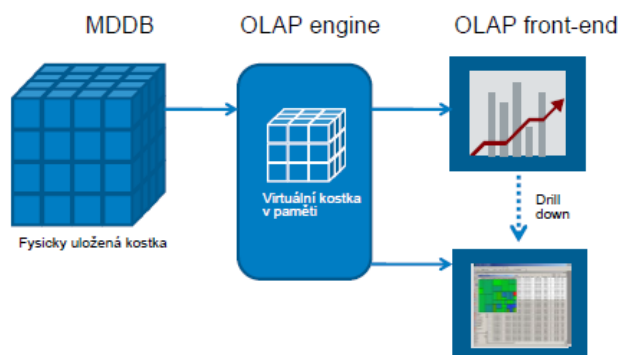
4.4. Co je ROLAP, k čemu je to dobré a čím se vyznačuje?

4.5. Jaké jsou výhody a nevýhody MOLAP?

4.6. Co je HOLAP, jaké komponenty využívá a jaká data kde uchovává? (●)

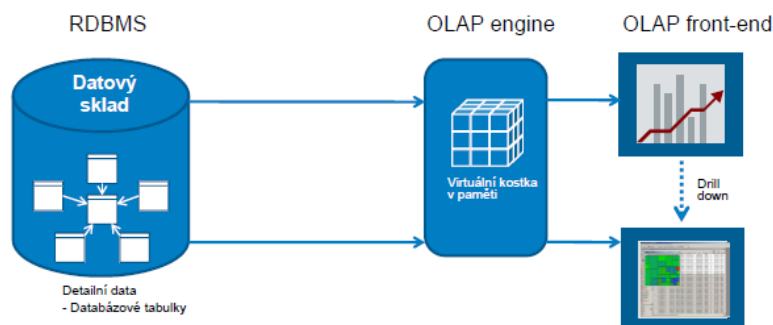
#### ■ **MOLAP**: Multidimenzionální databáze

- využití MDDb a MD zobrazování
  - všechna data (agregace i detaily) jsou v multidim. db (MDDb)
  - fyzicky (na disku) uložená „kostka“
- vhodný pro malé a středně velké db
  - Výhody: Rychlost (výkon), optimalizováno pro multidim. data
  - Nevýhody: Malá flexibilita, nároky na prostor, omezený počet dimenzí



#### ■ **ROLAP** : Relační OLAP

- detailní data zůstávají v originálních relačních tabulkách, speciální relační tabulky jsou použity pro uložení agregací
- zobrazována jsou multidimensionálně
- vhodný pro velké nebo distr. db nebo pro málo dotazovaná data
  - Výhody: flexibilita, škálability
  - Nevýhody: nároky na výkon, správu



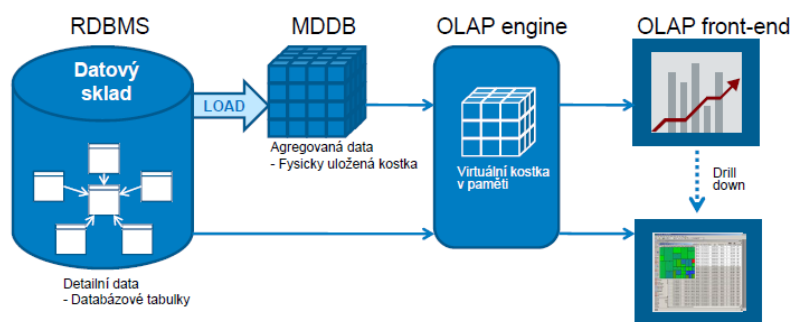
#### ■ **HOLAP**: Hybridní OLAP - kombinace MOLAP a ROLAP

- Kombinuje prvky MOLAP a ROLAP
  - Originální data ponechává v relačních tabulkách (RDBMS)
  - Agregace ukládá v multidimenzionálním formátu (MDDb)
- Poskytuje přístup k velkým databázím při zachování velké rychlosti při přístupu k



agregovaným datům

- Výhody: přístup k velkým datům, současně rychlé agregace
- Nevýhody: Údržba dat na dvou místech a z toho vyplývající problémy



- DOLAP: Dynamický OLAP: Virtuální MD matice postupně budována v paměti, zdrojová data v RDBMS
  - Výhody: Neomezená flexibilita
  - Nevýhody: Nároky (omezení) na RAM, závisí na výkonosti DB, kostka se vždy znova buduje
- DOLAP: Desktop OLAP - část MD kostky downloadována "na desktop"
  - Výhody: Autonomnost analýzy
  - Nevýhody: složitá aktualizace

4.7. Uvedte příklady technologií pro [OLAP](#).

- Multidimensionální databáze
- Agregace
- Writeback
- Drill-through

4.8. Vysvětlíte pojmy [ETL](#), [ELT](#), [EAI](#), [EII](#).

4.9. Jaký je rozdíl mezi ETL a ELT? Kdy je výhodnější použít ETL a kdy ELT? (●)

- přesun,transf. dat ze zdrojových systému do cílových (DW, Datamarty, soubory, ...)
- **ETL** (Extract Transform Load)
  - Cíl: extrahovat a přetransformovat data ze zdroje do datového skladu, většinou prováděno dávkově a po skupinách
  - Extract – extrahování dat z 1 nebo více zdrojů
  - Transform – transformace dat (čištění, přeformátování, standardizace, agregace nebo aplikace business pravidel)
  - Load – nahrání dat do cíl. systémů (datového skladu, ODS, DM, souborů, ...)
  - +/-
    - Výhody: může provádět více komplexní operace, lépe se škáluje, může pracovat in-stream
- **ELT** (Extract Load Transform)
  - Extract - výběr dat ze zdroje
  - Load - nahrání do vrstvy datového skladu
  - Transform – transformace dat v rámci databáze ale **jejím jazykem**
  - Příklady: Oracle OWB, MS SSIS
  - +/-
    - Výhody: často lepší nástroje v DB, využívá škálování DB, menší zátěž na síť
- **EAI** (Enterprise application integration)
  - Podobně jako ETL ale ne dávkové spracování ale **real-time**
  - Zatím nedostačující výkon, takže pouze u jednoduchých transformací
- **EII** (Enterprise information integration)
  - is a process of information integration, using data abstraction to provide a single interface (known as uniform data access) for viewing all the data within an organization, and a single set of structures and naming conventions (known as uniform information representation) to represent this data; the goal of EII is to get a large set of heterogeneous data sources to appear to a user or system as a single, homogeneous data source.

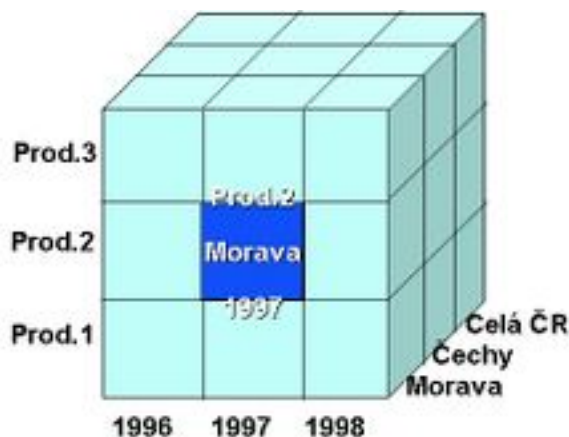
4.10. Uvedte příklady technologií [ETL](#).



- Informatica PowerCenter
- IBM DataStage
- Ab Initio
- SAS Integration Studio
- MS DTS
- Oracle Warehouse Builder

4.11. Co je **multidimenzionální kostka** a co obsahuje? (●)

- 1 tabulka faktů + n dimenzních tabulek



- Obsahuje

- Tabulku faktů (metrik)
  - Buňky MD kostky → fakty ... sloupce tabulky s hodnotami ukazatelů
  - Souřadnice buněk MD kostky → dimenzní klíče ... cizí klíče do tabulek dimensí
- Tabulky dimensí (hierarchií)
  - „Legenda“ k souřadnicím MD kostky → sloupce s hodnotami identifikačních atributů (elementů)
  - Souřadnice na osách MD prostoru → primární klíče tabulek dimensí

4.12. Co jsou operace **drill-down** a **drill-up** (uvedte příklad) (●)

- Navigace v hierarchii dimenzí - směrem k většímu detailu (drill-down) a k menšímu detailu (drill-up)
- Příklad: Zobrazování hodnot prodejů za rok s rozpadem na měsíce, dny (drill-down)

4.13. Co je **materializované view**, jaké má výhody a nevýhody? (●)

- Výsledek příkazu select se uloží do speciální („neviditelné“ systémové) tabulky
- aktualizace
  - Průběžně („trigger based“, vícenásobný insert/update/delete)
  - Jednorázově – počítá se s dávkovým loadem
- zhodnocení
  - výhody:
    - rychlejší než normální view
    - pokud je dobře navrženo tak rychle předcachuje dlouhotrvající dotazy
    - data lépe čitelná
  - nevýhody:
    - musí se navrhout dobře jinak se pořád precachovává
    - aktualizovatelná views nesmí obsahovat agregace

## 5. Datová kvalita a MDM

5.1. Co je parsing, standardisace, deduplikace? (●)

5.2. Co je parsing, k čemu je dobrý a jak se provádí (●)

5.3. Co je unifikace zákaznických dat, k čemu je to dobré, jaké hlavní problémy jsou s ní spojeny?

5.4. Co je deduplikace a jak se provádí?

- **Parsing** - rozpoznávání obsahu datových položek

- používá se při automatickém čištění dat pro rozeznání dat, které je třeba opravit
- způsob realizace:
  - definice vzoru hledaných dat: `<pattern definition='{FIRST_NAME} {LAST_NAME} MLADSI' name='F L mladsi' />`

- vyhledání tokenov

## ■ Standardizace

- Zlepšení stavu dat, náprava defektů
- Převod na **jednotný formát**

## ■ Unifikace - určení záznamu předst. 1 subjekt (osoba, adresa, produkt, vozidlo ...)

- Dobré: z dat dostanem konkrétní, smysluplnou informaci a přidělíme jí nový jednoznačný identifikátor jako unifikovanému subjektu
- Problémy: např. s parsováním - v sloupci pro jméno najdeme "Pan" nebo "Jméno Příjmení", nebo můžou být problémy se správností unifikace - záznam se neunifikuje, i když unifikovaný měl být

## ■ Deduplikace (Surviving record identification) - Stanovení nejlepšího reprezentanta

- Deduplikované databáze obsahují (právě) jeden záznam pro každého konkrétního jedince – reprezentant (Mělo by jich být méně)
- Reprezentant nemusí (ale může) být master, Záleží na metodě jeho tvorby:
  - např. Nejlepší z nejlepších (BoB)
  - nebo ten Master z unifikace
  - nebo některý ze závazného číselníku

## ■ Identifikace - Pro nové záznamy – nalezení (unifikovaného) subjektu, kterému záznam patří

### 5.5. Co je metoda **BoB**? (●)

- při deduplikaci
- Konstrukce Master záznamu: vytvoří nejlepší záznam složením z částí záznamů (každý kus může vzít z jiného zdroje)

### 5.6. Co je to **System of Record**? (●)

- systém pro úložiště a správu Master dat
- Súčasti:
  - Databáza (buď modelovanie nového modelu, alebo použitie existujúceho riešenia)
  - Aplikácia pre správu Master dát
  - Interfaces, API, ...

### 5.7. Naznačte stručně, co je MDM Hub a jaké má funkce.

- Master Data Management = Služby přístupu k Master datům (jeden z nich je System of Record)
- Data a služby (funkce):
  - Často řešení Centrální DB (něco jako DW nebo ODS) - tvoří hub, skrz nějž jsou synchronizována master data, metadata a fyzická data
  - Mohou to být master tabulky nebo master soubory, v nichž se shromažďují a uspořádávají záznamy
  - Někdy využití existujících aplikací ([CRM](#), [ERP](#)), pokud už obsahují potřebné definice

### 5.8. Co je householding, jaké jsou typy HH a k čemu je to dobré

- Seskupení klientů, kteří mají něco společného
- Při householdingu se hledají vztahy mezi klienty (skutečnými lidmi)
- Typy
  - Riskově orientovaný HH
    - „Liberální“ přístup (volnější pravidla, potenciální vazby ...)
    - Každý klient tvoří jádro 1 HH
    - Klient může být ve více HH
  - Marketingově orientovaný HH
    - „Konservativní“ přístup (co nejpřesnější identifikace HH, opatrnost)
    - Každý klient je právě v 1 HH
- K čemu dobré
  - Obchod a marketing
    - Nabídka produktů pro celý household
    - Profitabilita klientů
    - Hodnota, potenciál a riziko na úrovni householdu
  - Ošetřování rizik, fraud management ...

## 6. Ostatní

- 6.1. Co byste se chtěli ještě dozvědět?
  - Modelovanie DWH v konkrétnych modelovacích nástrojoch ako Power Designer.
- 6.2. Co znamenají zkratky [TQM](#), EDQM?
  - Různé metodiky pro řízení DQ
  - Total Quality Management (TQM) is a business management strategy aimed at embedding awareness of quality in all organizational processes.
  - Enterprise Data Quality Management (EDQM) - Data Quality Insurance for the Enterprise
- 6.3. Jak je to s mazáním dat z DW?
  - Single Snapshot - celá tabulka je smazána a naplněna znovu
  - Sequential Snapshot - každá aktualizace přidá nový snashot k předešlým datům
  - Incremental - každá aktualizace přidá pouze nové záznamy
  - Incremental with Update - každá aktualizace přidá nové záznamy a aktualizuje existující záznamy
- 6.4. Jaké jsou základní technologické komponenty DW (stručně charakterisujte)
  - Datová úložiště - databáze pro ukládání dat DW, datamartů, ODS etc.
  - Integrace - extrakce, transformace a load - přesuny mezi zdrojovými systémy a jednotlivými částmi a vrstvami DW
  - Visualisace: Analytické nástroje, reportovací nástroje, monitorování etc.
  - Datová kvalita
  - Modelování
  - Správa metadat
- 6.5. Která část kursu vás nejvíce zaujala?
  - :-)
- 6.6. Uveďte 3 nejpodstatnější vlastnosti řešení [DWH](#), kterými se liší od jiných aplikací a systémů
  - Orientace na subjekt - uživatele
  - Integrovanost - spojení několika zdrojů
  - Nízká proměnlivost
  - Historizace