# 1. Data x informace, sémantika, ontologie, sémantický web, RDF, nové možnosti zpracování dat

# 1. Informace a data – hodnota dat

- Hodnota dat
  - Data nemají cenu sami o sobě
  - Cenu má jejich porozumění a informace v nich obsažené
- Business Inteligence (BI)
  - Techniky a nástroje umožňující transformovat hrubá data do užitečných a srozumitelných informací pro potřeby uživatelů

# 1. Informace a data – hodnota dat

- Analýzy používané v BI
  - Deskriptivní analýza
    - Analýza historických dat – strategický reporting
    - Analýza aktuálních dat – operativní reporting
  - Prediktivní analýza
    - Prediktivní modely založené na historických datech
    - Předpovídají budoucnost
  - Preskriptivní analýza
    - Používá modely pro určení optimálního chování a akcí
    - Data driven decision making

# 1. Informace a data – data

- Zobrazují vlastnosti, stavy objektů a probíhajících procesů v reálném prostředí kolem nás
- **Strukturovaná**
  - Zachycují explicitně fakta, atributy a objekty
  - Pevně daná struktura umožňuje jednoduše vybírat relevantní hodnoty
  - Př. obsah relační databáze
- **Nestrukturovaná**
  - Vyjádřená určitým tokem bitů bez další specifikace
  - Př. videozáznam, obrázek, textový dokument
- Př.: obsah uložišť dat v organizaci - obsah databází, soubory a dokumenty, email, obsahy disků, znalosti zaměstnanců

# 1. Informace a data – informace

- Data obohacená o další údaje – metadata
  - Datový formát
  - Definice
  - Relevance
  - Platnost
  - Správce
- Data s přiřazenou sémantikou, kontextem
- Informace je časově pomíjivá, data trvalá
- Odpovědi na otázky Kdo? Co? Kde? Kdy?
- Př.: prezentace dat v aplikaci

# 1. Informace a data – znalosti

- Zpracované a analyzované informace
- Zobecněné poznání reality
  - zobecnění procesů a stavů
- Na jejich základě je možné se rozhodovat
  - Vazby; Trendy; Vzory chování
- Př.: analýzy, reporty, výzkumné zprávy
- Porozumět jak? Pochopení vzorů
- Životní cyklus znalosti
  - Tvorba, získání, zjištění znalosti – učení
  - Uchování, uspořádání – porozumění
  - Přenos, šíření, sdílení
  - Používání, aplikace

# 1. Informace a data – moudrost

- Vysoce analyzované informace
- Soubor znalostí vycházející z pochopení podstaty problematiky v daných souvislostech

- Porozumět Proč? Pochopení principů
  - Nástroje BI
  - Pokročilé analýzy

# 1. Hierarchie dat v organizaci

- Metadata
  - Data popisující struktury, význam a použití ostatních dat
- Reference data
  - Číselníky, které nevlastní a nespravuje organizace
- Enterprise structure data
  - Číselníky organizace, organizační struktura, zaměstnanci, obchodní procesy a funkce, aplikace, bezpečnostní přístupy, …
- Transaction structure data
  - Struktura základních datových entit v organizaci
  - Hlavní kniha, katalog zboží, struktura faktury, …

# 1. Hierarchie dat v organizaci

- Transaction Activity Data
  - Obchodní data organizace včetně všech odvozených dat
- Transaction audit data
  - Loggy, audity
- Zvyšuje se


  - sémantický obsah, důležitost kvality
  - objem dat, frekvence aktualizací, kratší životnost

# 1. Data management

- Data management zahrnuje všechny disciplíny související se správou dat
- Komplexní množina postupů, konceptů, procesů a široké škály doprovodných systémů, které umožňují organizaci získat kontrolu nad datovými zdroji
- Souvisí s celým životním cyklem dat od jejich vzniku až po zastarání

# 1. Kompetence data managementu

- Co je kompetence
  - Cíle (mission, vision)
  - Procesy (plánovací, kontrolní, vývojové, operativní)
  - Pravidla (politiky, doporučení, best practices, knowledge base)
  - Metriky
  - Organizace – vlastníci dat (data owner), data stewardship, data stewardship committee, BI oddělení, oddělení bezpečnosti, oddělení (datové) kvality, databázoví administrátoři
  - Nástroje – systémy pro správu dat, zálohovací systémy

# 1. Kompetence data managementu

- Datová architektura
- Datová kvalita
- Metadata
- Bezpečnost
- Dokument a kontent management
- Data warehousing a BI
- bigData
- master data management a správa číselníků
- databázový vývoj
- provozování datových systémů

# 1. data management a enterprise architecture

- enterprise architecture
  - kompetence jednotného pohledu na organizaci zahrnující jako obchodní procesy, tak technické řešení
- př. EA fram.
  - Zachman Framework, TOGAF, Archimate
- Základní entity EA
  - Business funkce
  - Business procesy
  - Organizace
  - Aplikace
  - Datové typy

# 1. data management a enterprise architecture

- EA řeší vazby mezi entitami, aby byla schopna odpovědět na otázky:
  - Které obchodní funkce podporuje tato aplikace
  - Co jsou technicky úzká místa tohoto obchodního procesu?
  - Pokud změníme organizaci, která aplikace a které obchodní procesy je třeba upravit?

# 1. Informační tok dat v organizaci

- Agendové aplikace
- Datové služby
  - Datová kvalita
  - Integrace
  - Master data management
- Datový sklad
  - Jednotný model agendově a aplikačně nezávislý
  - Jednotně spravovaný na úrovni celé organizace
  - Kompletní historie
- Operativní datový sklad
  - Real-time, near-to-real-time řešení
- Specializované datamarty

# 1. XML

- XML
  - poskytuje základní syntaxi pro strukturované dokumenty
- XML schema
  - je jazykem pro omezení struktury XML dok, rozšiřuje XML o datové typy

# 1. Sémantický web

- Termín zaveden kolem r. 2000 pro oblast výzkumu vzniklou spojením
  - Nástrojů a standardů sítě www
  - Technologie reprezentace a zpracování znalostí, zejména
    - Modelování znalostí (ontologické inženýrství)
    - Formální logiky
- Později se zapojily i další komunity
  - Zpracování přirozeného jazyka, text/web mining databáze, (mezi-)podnikové procesy, filosofie, zpracování neurčitosti, sociální sítě, …
- Dialog komunit je přínosem už sám o sobě

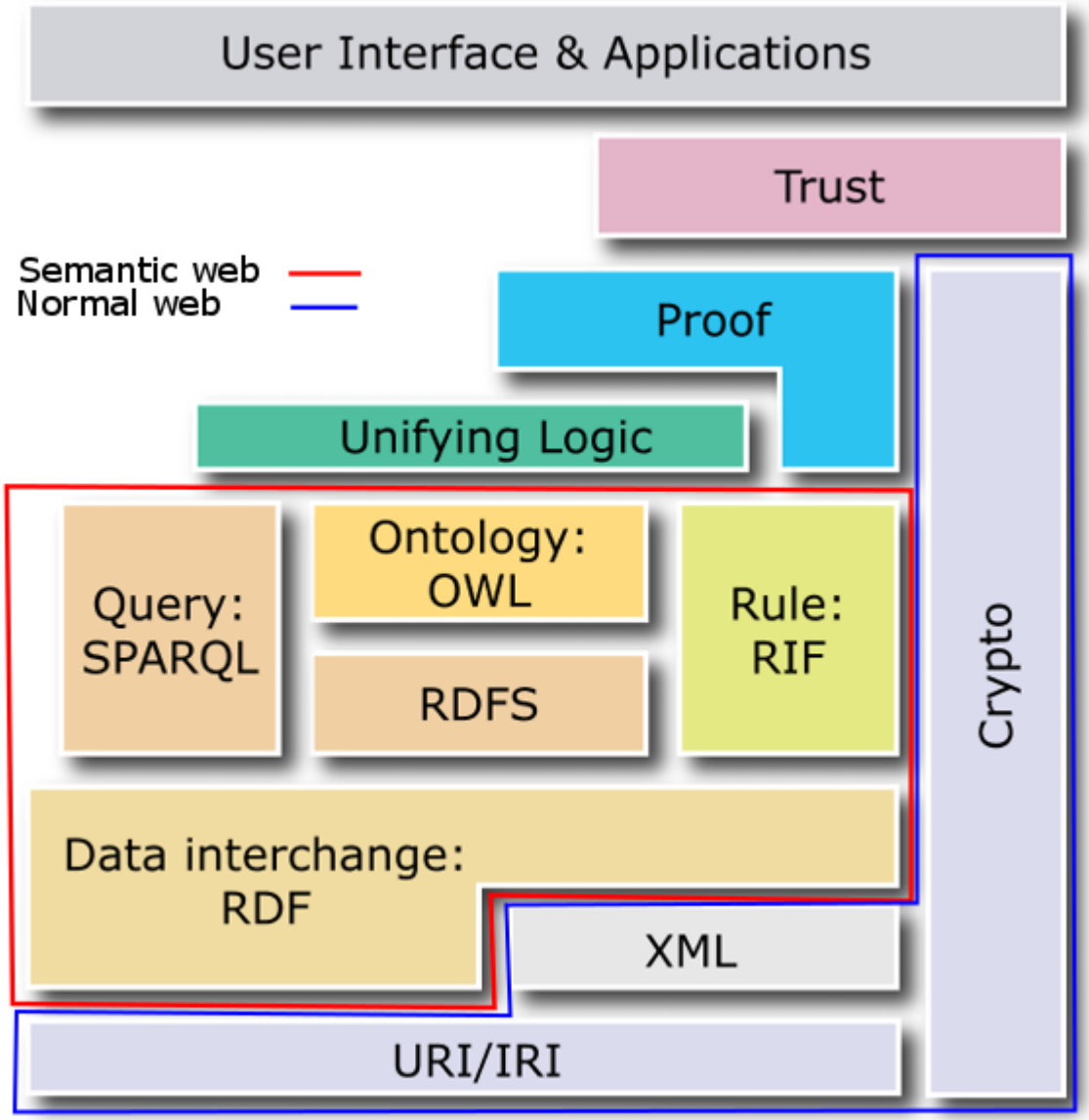
# 1. Sémantický web

- Cíl (Tim Berners-Lee)
  - Web dostupný nejen pro lidi, ale i pro počítače musí být schopen formálně reprezentovat informace a definovat jejich význam
- SW je reprezentací dat na webu, rozšíření současného webu o metadata, které specifikují přesný význam dat
- Cílem je zpřístupnit (popsat) informace na webu pro strojové automatické zpracování, ne jenom pro zobrazení člověkem
- Tak aby se při přenosu dat z různých zdrojů neztrácel jejich význam

# 1. Sémantický web

- Př.: automatické propojení několika datových množin do aplikace třetí strany a zpřístupnění tak, rozsáhlejších informací pro uživatele z jednoho UI
  - Např. s možností agregací dat určitého tématu
- Jádrem současné koncepce sémantického webu jsou data reprezentovaná v jazyce RDF, s významem definovaným pomocí ontologií, s odvozováním nových informací zejména pomocí pravidel

# 1. Sémantický web - vrstvy



# 1. XML a sémantika

- Sémantika
  - Význam sdělení pro příjemce
- Stromová struktura XML pouze předepisuje způsob zaznamenávání dat, nic nevypovídá o jejich významu
- Sémantiku musí přidat „lidský“ uživatel

# 1. RDF - Resource Description Framework

- Framework (množina specifikací, model)
- Poskytuje jednoduchý jazyk pro vyjádření informací o zdrojích na webu
- Zdrojem může být obecně cokoliv, dokumenty, lidé, fyzické objekty i abstraktní koncepty, čísla, řetězce, co je jednoznačně identifikovatelné
- Základní prvek – trojice
  - <subjekt><predikát><objekt>
  - Každá trojice představuje jedno platné tvrzení
  - Př.: <pavel><bydlet><praha>
  - Zdroj X nabývá pro vlastnost Y hodnoty Z
  - Vyjadřuje orientovaný vztah mezi dvěma zdroji

# 1. RDF

- Množina trojic lze reprezentovat orientovaným grafem
- Modulární  - trojice jsou na sobě nezávislé
- Zdroje jednoznačně identifikovatelné – URI/IRI
- RDF a ontologie
  - Nová tvrzení můžeme odvodit tehdy, když konkrétní zdroje přiřadíme k obecným *třídám* jakožto jejich *instance* pomocí konstrukce *rdf:type*
  - Vlastnosti definované u tříd se pak promítají do jejich instancí
  - Struktura tříd a jejich vlastnosti mohou být definovány v *ontologiích*

# 1. RDF - slovníky

- RDF datový model neposkytuje žádné informace o sémantice zdrojů
- Nutné použít další prostředky, které tyto informace poskytnou
  - Slovníky, ontologie
  - Slovník – kolekce IRI spolu s definicí kontextu, domény, oboru hodnot, …
  - Ontologie – formalizované slovníky, často pro specifickou oblast
- Serializace dat do několika formátů
  - RDF/XML, Turtle, N-triple, …

# 1. RDF - slovníky

- Jednoduché jazyky
- RDF (vocabulary)
  - Základní prvky pro popis zdrojů a relací
  - Jednoduchá sémantika
- RDF Schema
  - Rozšiřuje RDF (nyní již jedna specifikace)
  - Obsahuje prostředky pro popis skupin a souvisejících zdrojů (třídy, podtřídy) a vztahů mezi nimi (vlastnosti, podvlastnosti)
  - Domain – definice definičního oboru vlastnosti
  - Range – obor hodnot

# 1. Ontologie

- „Formální specifikace konceptualizace“
- Konceptualizace
  - Systém pojmů modelující určitou část světa
  - Abstraktní model určitě oblasti
- Formální
  - Vyjádřená ve formálně-logickém jazyce

# 1. Ontologie - OWL

- Ontol. jazyk založený na deskripční logice
- Bohatší vyjadřovací schopnosti než RDF

- Lokální omezení vlastností v rámci určité třídy
  - kardinalita
  - univerzální a existenční kvantifikace
- Matematické charakteristiky vlastností (tranzitivní, funkční, inverzní, …)
- Disjunktnost či ekvivalence tříd
- Anonymní (nepojmenované) třídy, definované určitým logickým výrazem pro jednorázové použití

# 1. Ontologie - owl

- **Odvozovací úlohy v OWL**
  - Testování splnitelnosti tříd … tím i konzistence ontologie jako logické teorie
  - Odvozování taxonomické struktury
  - Ověřování příslušnosti instance ke třídě
  - Klasifikace individua vzhledem k ontologii

# 1. Ontologie – verze OWL 1

- OWL lite
  - omezený z hlediska elementárních konstruktů
  - neumožňuje definovat kardinalitu jinou než 0,1
  - výpočtově efektivní
- OWL DL
  - Všechny konstrukce OWL a některé z RDF, ale omezené
  - Zaručen výpočetní výkon a správnost vrácených tvrzení
  - Nejpoužívanější
- OWL Full
  - Úplné sjednocení RDF a OWL, bez omezení
  - Nezaručuje výkonost a správnost výsledných tvrzení

# 1. Aplikace sem. webu

- **Ontologie x pravidla**
  - Ontologie založené na deskripční logice umožňují jen omezený okruh typů odvození
  - Zejména chybí možnost odvozovat (pro daný objekt) hodnotu jedné vlastnosti z hodnoty jiné vlastnosti
- **Aplikace sem. webu**
  - Sémantické vyhledávání na webu
  - Elektronické obchodování
  - Automatická tvorba portálů
  - Podpora vědecké spolupráce
  - Podpora výuky (e-learning)
  - NLP, text mining a web mining

# 1. Extrakce informací

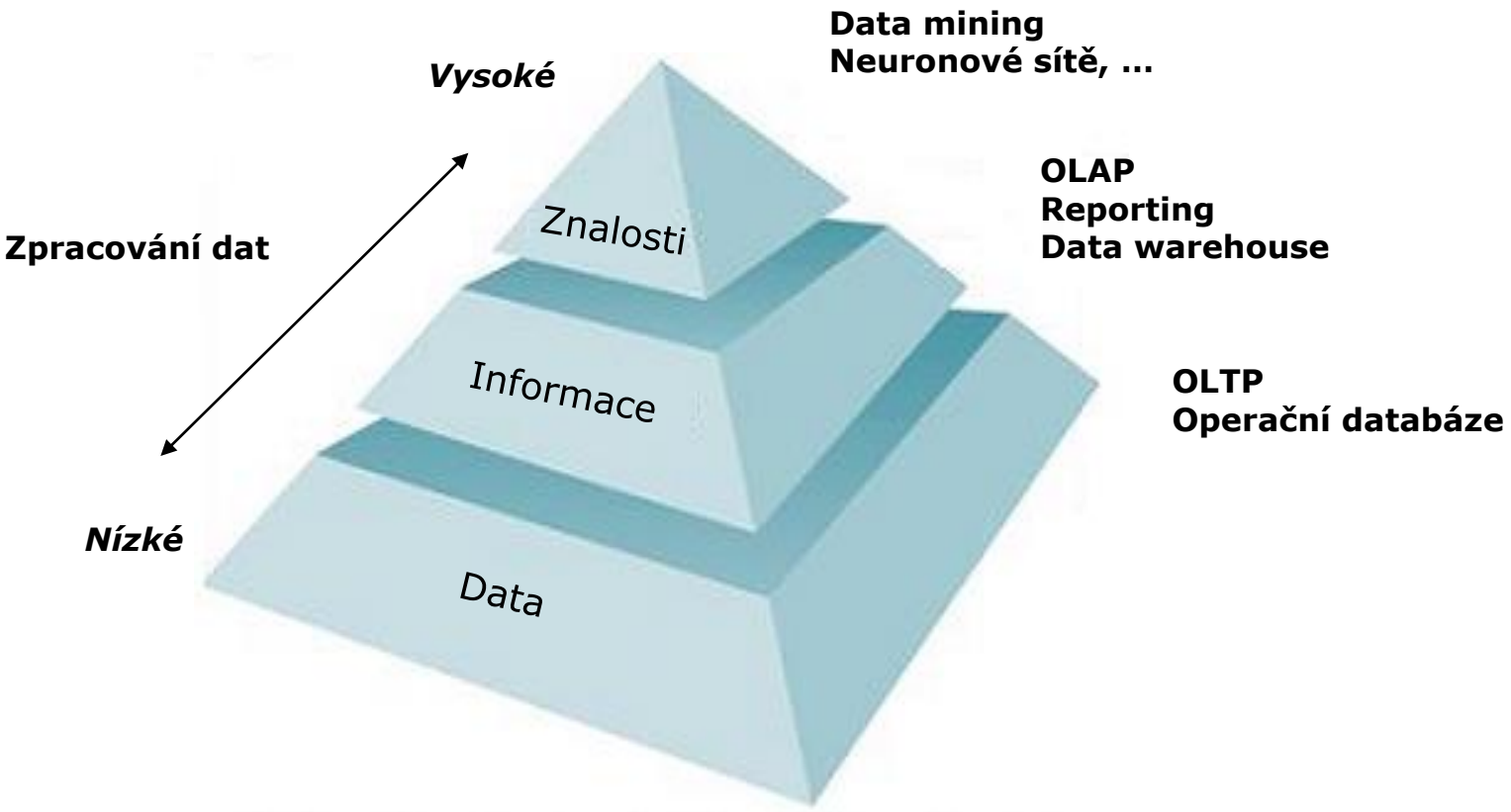
- Extrakce do šablony přirozeně evolvuje do „populování ontologie“
- Možnost využít informaci již obsaženou v ontologii (např. kardinalitní omezení)
- Při extrakci z webu možnost opřít se o strukturu HTML
  - Kromě lexikálních indikátorů relací (slovesa, předložky…) také charakteristické struktury v HTML
- Často se znovuobjevují věci známé z „lingvistické“ sémantické analýzy

# 2, 3.
# Pokročilé transakční modely směrem k workflow. Bussines Intelligence.
# Architektura BI, datový sklad, datové tržiště,ETL, OLTP a OLAP databáze.

# 2,3. Business Inteligence

- Proces transformace podnikových dat na informace a převod informací na znalosti, sloužící k podpoře podnikání
- Techniky a nástroje umožňující transformovat hrubá data do užitečných a srozumitelných informací pro potřeby uživatelů

# 2,3. Business Inteligence



# 2,3 - Historie BI

- Navazuje na rozvoj databázových systémů:
  - Transakční systémy (OLTP)
  - 60. léta – Dávkové zpracování výkazů
    - Složité nalézt a analyzovat informace
    - Nákladné a neflexibilní, přeprogramování dle nových požadavků
  - 70. léta – první manažerské aplikace (Lockheed)
    - Terminálově orientované EIS a DSS aplikace
  - 80. léta – Desktopové aplikace a analytické nástroje
    - Dotazovací nástroje, tabulkové procesory, GUI
    - Jednoduché na používání, přístup pouze k operačním databázím
  - 90. léta – Rozvoj datových skladů, integrace OLAP databází a DM technik

# 2,3 – Komponenty BI

- Produkční a zdrojové systémy
- Dočasná uložiště dat (DSA)
- Operativní uložiště (ODS)
- Transformační nástroje (ETL)
- Integrační nástroje (EAI)
- Datové sklady (DWH)
- Datové tržiště (DM)
- OLAP
- Reporting
- Managerské aplikace (EIS)

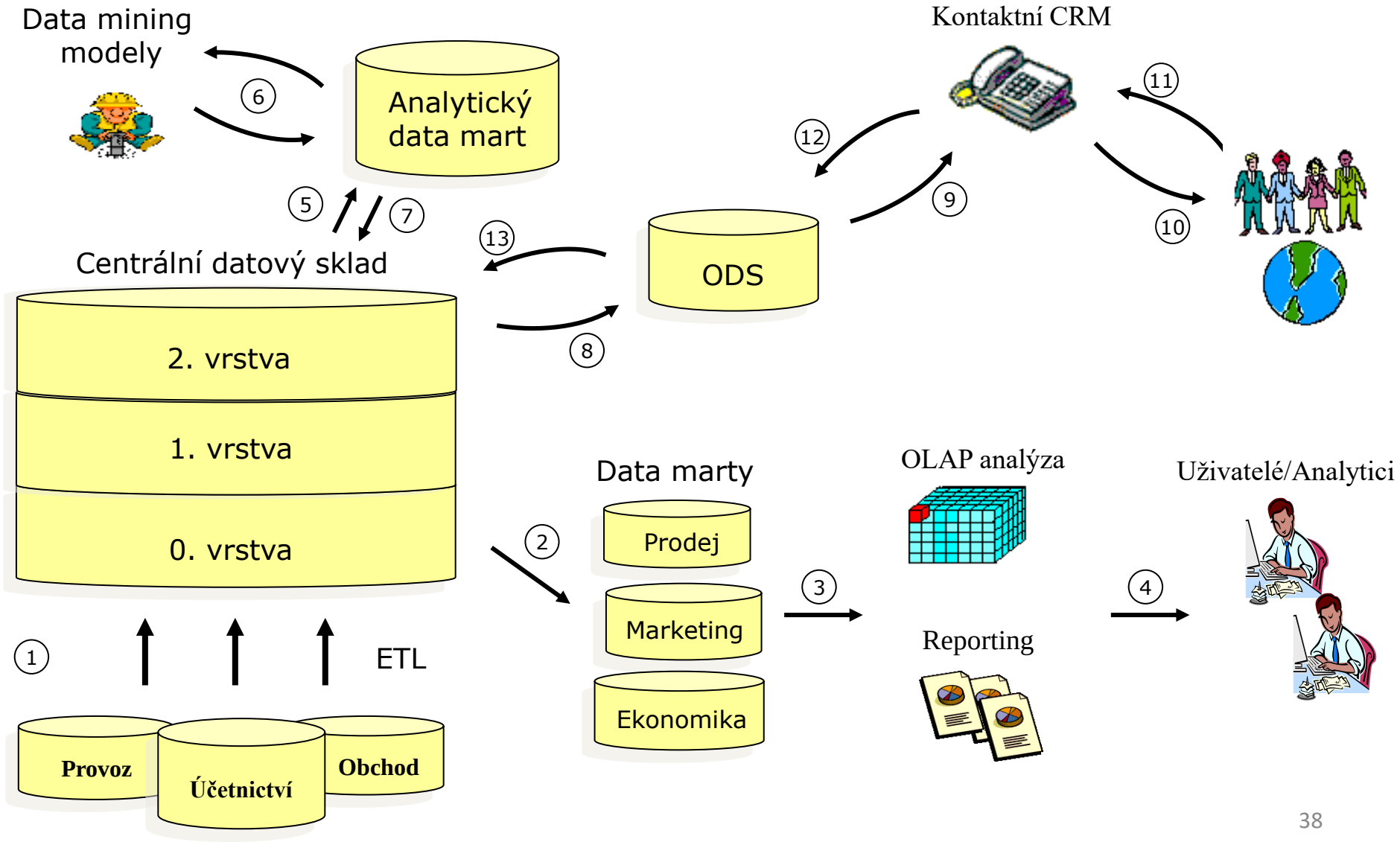
# 2,3 – Komponenty BI

- Dolování dat – data mining
- Nástroje pro zajištění kvality dat
- Nástroje pro správu metadat

- Každé řešení BI se skládá z řady vrstev, které si předávají data od původního zdroje (zdrojové DB) ke zpřístupnění výsledných analytických inform. uživateli
- Komponenty řešení BI se mohou lišit podle situace, potřeb klientů/podniku
- Neexistuje žádná jednotná struktura

# 2,3 – Architektura BI

- CRM – řízení vztahů se zákazníky

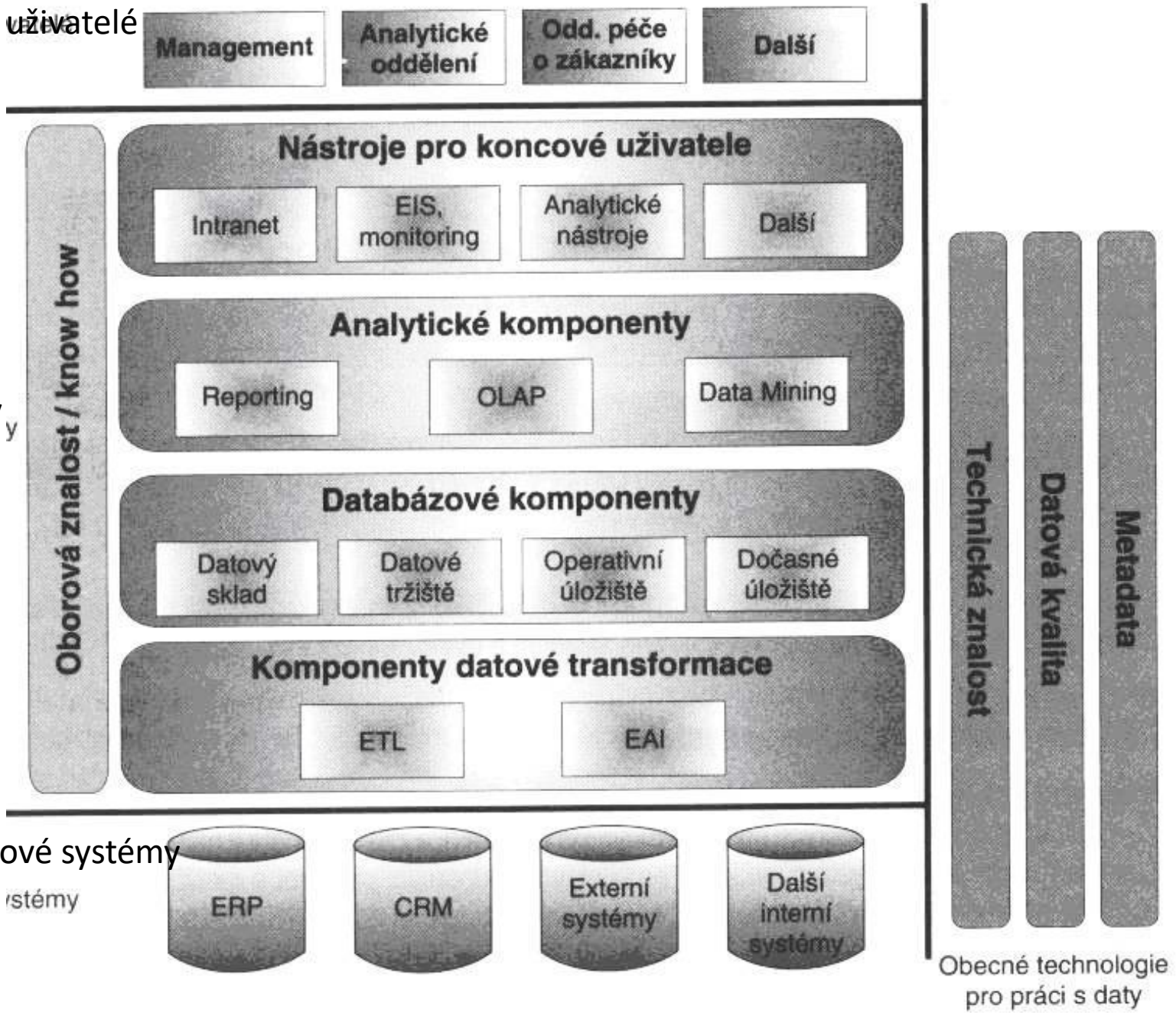


# 2,3 – Architektura BI

Koncový uživatelé

Komponenty BI

Zdrojové systémy



# 2,3 – Datový sklad (data warehouse)

- „Datový sklad je podnikový strukturovaný depozitář předmětově orientovaných,
- vzájemně provázaných,
- časově neměnných,
- historických dat používaný na získávání informací a podporu rozhodování.
- V datovém skladu jsou uložena detailní a sumární data.„
- (Bill Inmon)

# 2,3 – Datový sklad (data warehouse)

- Data jsou extrahována z produkčních DB v daném čase (většinou po tzv. uzávěrce dne), ne po každé změně
  - ukládána po časových snímcích (snapshot)
  - vytváří časově proměnnou řadu – historii
- Časově neměnný
  - Údaje se zpravidla nemění ani neodstraňují, jen přidávají
- Integrovaný
  - pochází z několika produkčních systémů podniku
  - na základě určitých pravidel spojována tak, aby poskytla koncovému uživateli celopodnikový pohled na oblast jeho zájmu.
- Subjektová orientace
  - Uchovávají se údaje podle předmětu zájmu, nikoliv podle aplikace, ve které byly vytvořeny. Subjekt např. zákazník, dodavatel, ….

# 2,3 – Datový sklad (data warehouse)

- Centrální uložiště různorodých dat firmy
  - Řešení integračních efektů; Jednotné místo uložení dat
- Určeno pro analytickou podporu rozhodování
- Zahrnuje nejen data v databázích, ale i nástroje pro extrakci dat, reporting, analýzu dat, data mining…
- Prezentace dat uživatelsky příjemným způsobem zaměřeným na střední a vyšší management
- Třívrstvá architektura
  - dočasné uložiště,
  - centrální uložiště (ROLAP,  MOLAP),
  - datové tržiště (front-end klientské vrstva. Tato vrstva obsahuje nástroje pro vyhledávání a nástroje pro vytváření přehledů, analytické nástroje a nástroje pro dolování dat.)

# 2,3 – Přínos řešení DS

- Integrace a čistota dat, integrační efekty
- Podpora úloh analytického charakteru a jejich ekonomické a mimoekonomické efekty
  - Vyšší flexibilita řízení a realizace změn
  - Zpětná vazba
- Automatizace rutinních procesů (tvorba výkazů, zpráv)
- Kontrola plnění plánů a finanční analýza
- Podpora analýzy dat:
  - Trendy, sledování a analýza časových řad
  - Poměrové ukazatele
  - Identifikace odchylek
  - Drill-down, Drill-up, Drill-across, Slice-dice

# 2,3-Operace s datovým skladem v OLAP analýze

- **Drill-down**: ukaž mi větší detail
  - Přidání sloupce z dimenze do výstupu
- **Drill-up**: ukaž mi agregaci
  - Odebrání sloupce z výstupu
- **Drill-across**: spojení dvou a více faktových tabulek se stejnou granularitou
- **Drill-around**: podobné jako drill-across, ale pro nelineární uspořádání
- **Slice-dice**: řez multidimenzionální kostkou, omezení výběru
  - Slice – výběr dimenze (zákazník, produkt, čas)
  - Dice – výběr hodnoty v dimenzi (za rok = 2004 a produkt = chleba)

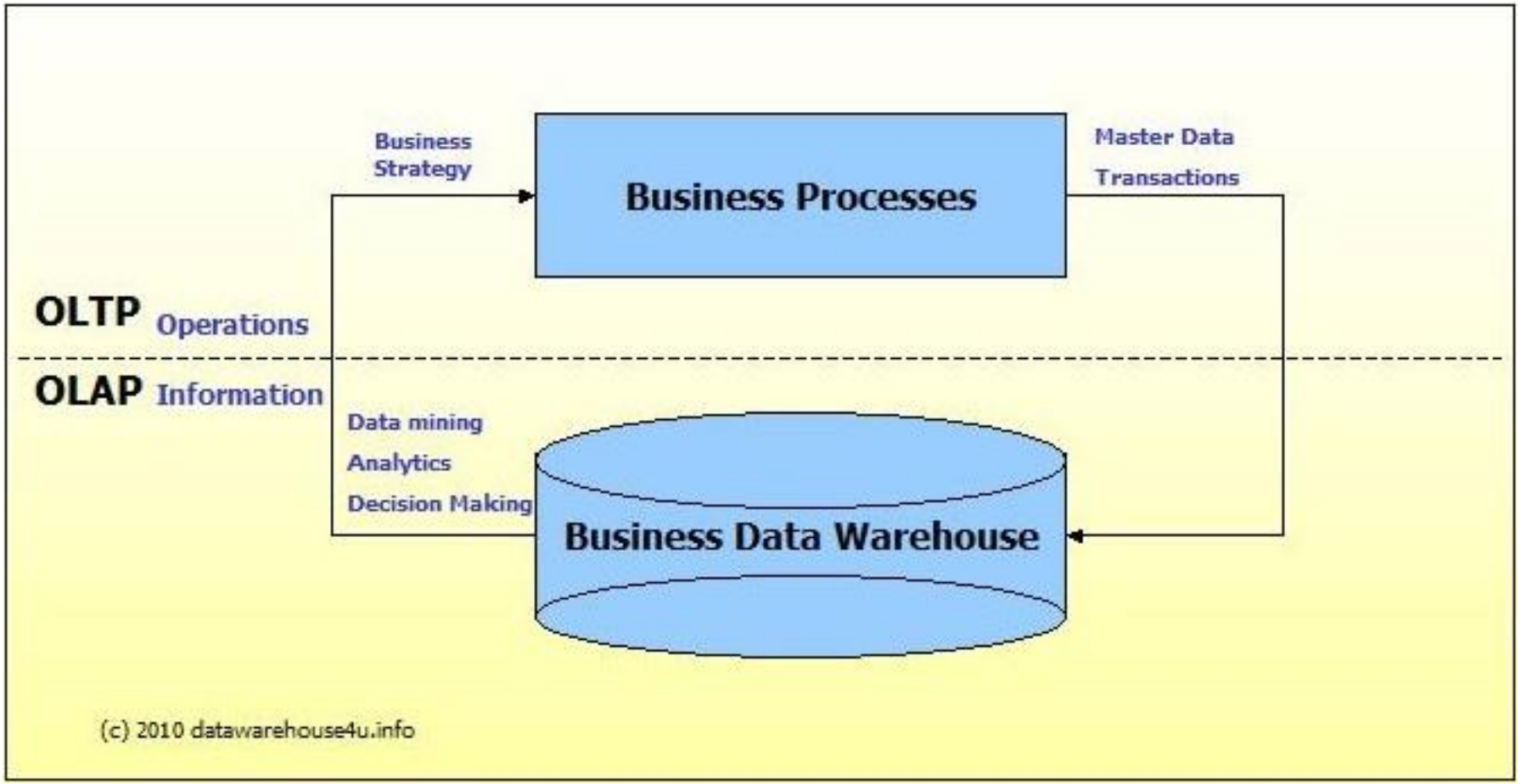
# 2,3 – OLTP – online transaction processing

- Provozní databáze
- Charakterizováno vysokým počtem krátkých transakcí (I, U, D)
- Důraz kladen na rychlé zpracování transakce,
  - udržování integrity dat ve víceuživatelském prostředí,
  - Efektivita měřena počtem transakcí za vteřinu
- Detailní aktuální data (každodenní operace)
- Schéma vysoce normalizované (obvykle 3NF)
  - Mnoho tabulek
- Podpora analýzy nízká
- Dotazy krátké a vracejí relativně málo záznamů

# 2,3 – OLTP x DW

- OLTP
  - Podpora transakcí – každodenní operace
  - Data uložená na úrovni transakcí
  - Normalizovaný datový model
  - Problém je decentralizace.  Používají se různé systém pro zpracování různých údajů.
- DW
  - Analýza i historických dat
  - Integrace dat
  - Denormalizovaný model

# 2,3. OLTP x OLAP



# 2,3 – Datové tržiště (data mart)

- Logická podmnožina DW
  - Tematicky orientovaný DW
- Část řešení DW určená pro podporu specifické analýzy (účetnictví) nebo oddělení firmy (marketing, finan. oddělení, …)
- Někdy může být DW vytvořen kompozicí jednotlivých DM

| **Vlastnost** | **Data Warehouse** | **Data Mart** |
|---|---|---|
| Rozsah | Enterprise | Oddělení |
| Obsah | Více oblastí | Jedna oblast |
| Zdroje | Hodně | Málo |
| Velikost (typicky) | 100 GB to > 1 TB | < 100 GB |
| Implementace | Měsíce až roky | Měsíce |

# 2,3 – Datové tržiště

- Budovány na základě požadavků jednotlivých částí organizace
  - Potřeba vlastních dat
  - Používání vlastních definic pojmů
  - Vlastní historie dat
  - Vlastní periodicita aktualizace dat
- Dvouvrstvá architektura
- Rychlá implementace, nižší náklady
- Volíme pokud potřebujeme rychlé řešení pro konkrétní pobočku/oddělení

# 2,3 – ETL – extraction, transformation, loading

- Kompletní proces načítání dat do datového skladu
- Zahrnuje mnoho subprocesů
  - **E**xtrakce – výběr dat
  - **T**ransformace – ověření, čištění, integrace dat
  - **L**oading – načtení dat do DW
  - Kontrola kvality; Auditování; Bezpečnost
  - Zálohování a obnova
- Proces přenosu z produkčních db je prováděn pomocí datových pump dávkově, v určených intervalech

# 2,3 – OLAP – online analytic processing

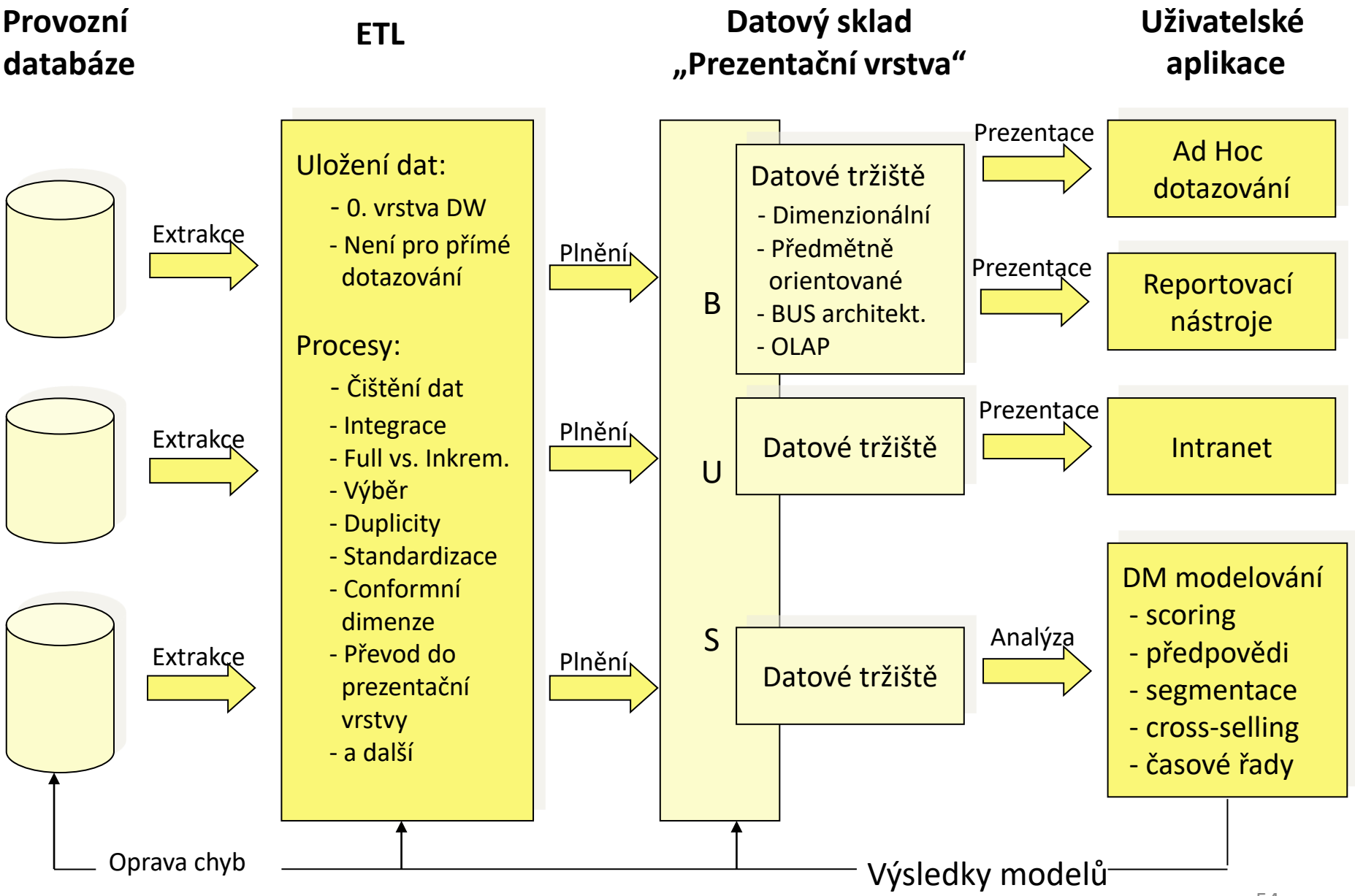
- Obecné označení pro dotazování a zobraz. dat z DW
- Multidim. uložení a analýza dat (OLAP DB)
- Relativně malý objem dotazů
- Dotazy často velmi komplexní a s agregacemi
- Efektivnost – čas odezvy
- Agregovaná, historická data z různých zdrojů v multidim. schématu
  - obvykle hvězda/vločka
- Denormalizované schéma s méně tabulkami
- Rychlost závisí na objemu dat, dávková aktualizace a složité dotazy můžou trvat několik hodin
- Pro urychlení - indexování

# 4. Přístupy k vytváření datových skladů a datových tržišť, výhody a nevýhody jednotlivých přístupů k tvorbě datového skladu. Ekonomická náročnost.

# 4. Cíle DW

- Zajistit dostupnost firemních informací
- Zajistit konzistenci firemních informací
- Vytvořit adaptivní a pružný zdroj informací
- Zabezpečit ochranu firemních informací
- Vytvořit základnu pro firemní podporu rozhodování (analytické centrum)

# 4. Architektura



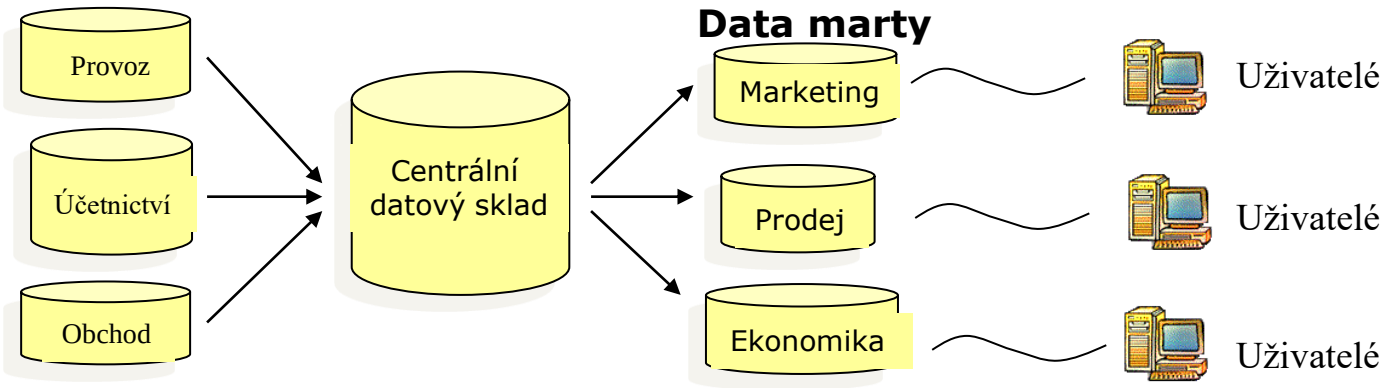
# 4. DW procesy

- ETL - Hlavní proces při tvorbě datového skladu
- Podprocesy:
  - Extrakce
  - Transformace
    - Čištění dat
    - Výběr dat
    - Integrace
    - Umělé klíče
    - Agregace
  - Načtení (Loading) a tvorba indexů
  - Data Quality Assurance

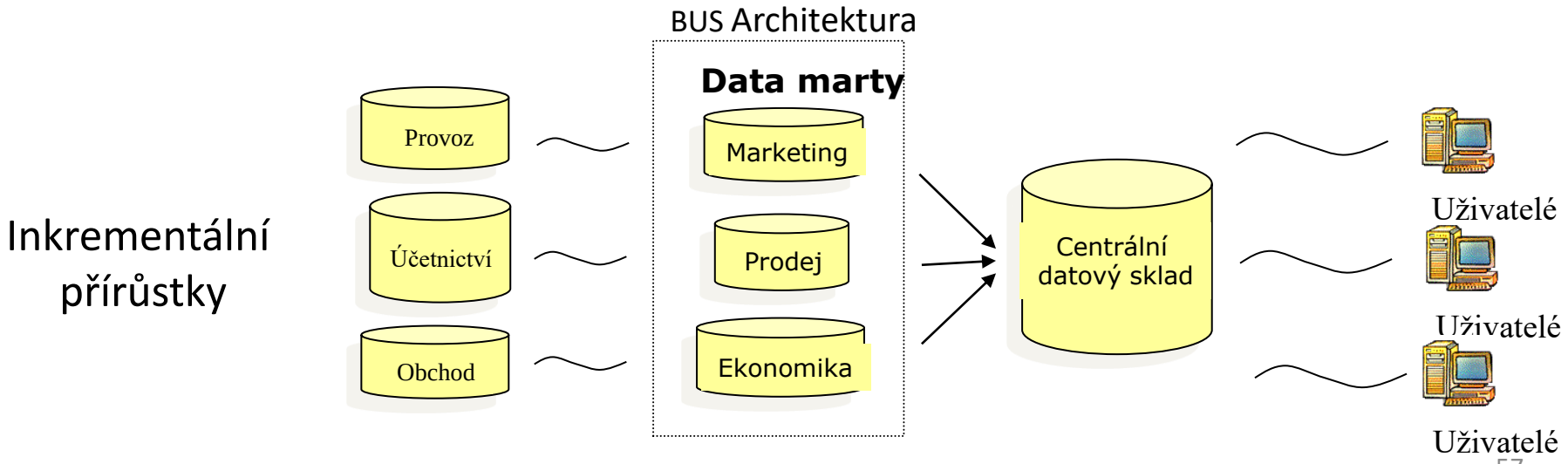
# 4. DW procesy

- Další procesy v DW
  - Publikace dat (prezentační server)
  - Update dat
  - Dotazování
  - Zpětná vazba (čistá dat do OLTP, data z DM do DW)
  - Audit dat
  - Bezpečnost
  - Zálohování a obnova

# 4. Vytváření DW




„Velký třesk“




Inkrementální přírůstky

# 4. Vytváření DW – metoda velkého třesku

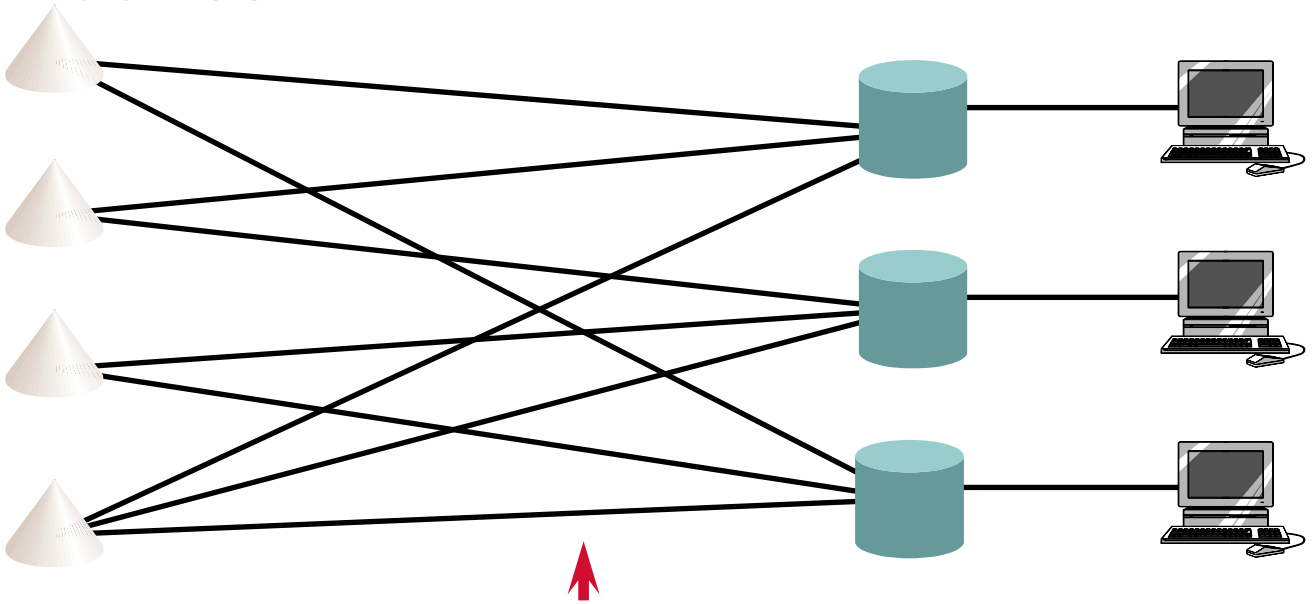
- Detailní návrh celého skladu a až poté implementace
- Vše uložíme do CDS a poté distribuujeme do jednotlivých DM
- Náročná realizace, velké nároky na zdroje
- V průběhu vývoje se můžou změnit technologie i potřeby uživatelů
- Trvá velmi dlouho než jsou viditelné první výsledky
- Tři etapy
  - Analýza požadavků
  - Vytvoření podnikového DW
  - Vytvoření přístupu buď přímo nebo přes DM

# 4. Vytváření DW – přírůstková metoda

- Evoluční metoda
- Budujeme DW po jednotlivých etapách
- Místo budování celého DW postupně přibývají jednotlivá přírůstková řešení
- Začneme budováním několika (1-2) předmětových oblastí, ty částečně impl. jako škálovatelný DW a poskytneme uživatelům pro otestování
- První subsystémy začnou fungovat krátce po zahájení → rychlejší zisk a návratnost investic
- Pokud se částečné řešení osvědčí, můžeme přidat další oblast
- Iterativní proces, udržuje neustálou spojitost mezi DW a uživateli

# 4. Nezávislá datová tržiště

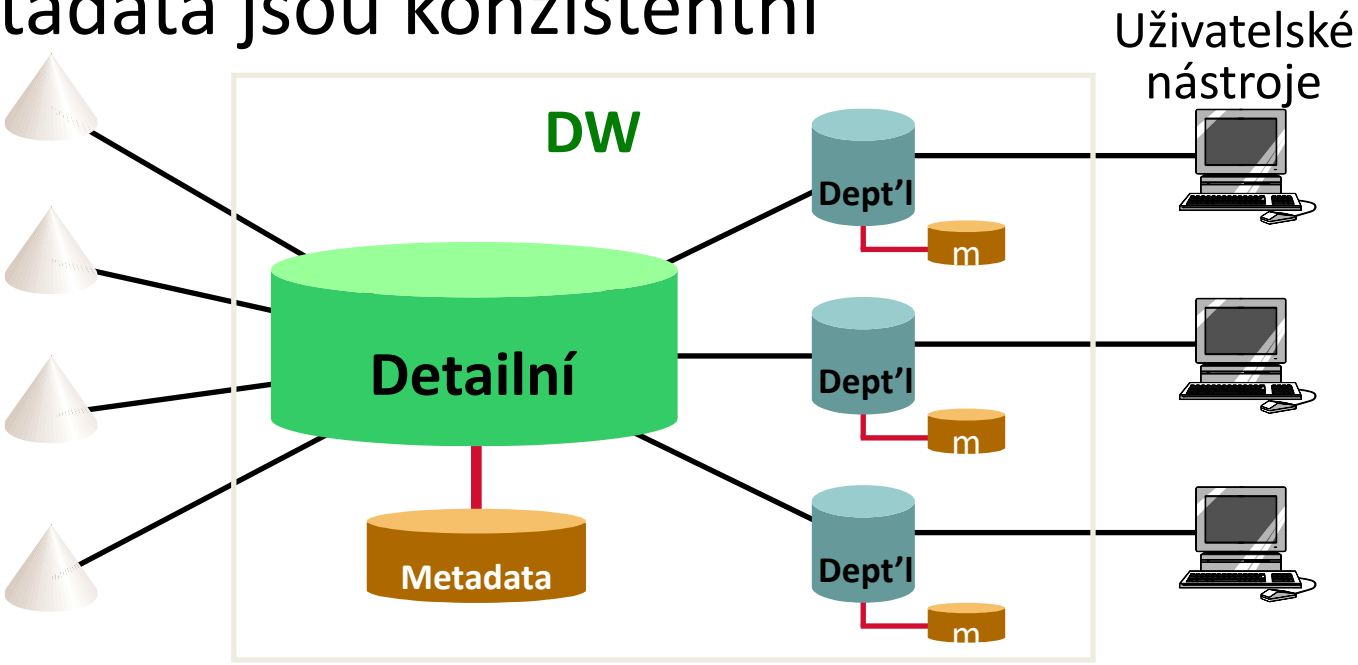
- Duplikace na ETL
- Časově i finančně náročné
- Údržba nezávislých DM je náročná a těžkopádná
- Rychlý vývoj, ale problém konzistence ETL, redundance



ETL – 3 krát !!!

# 4. Závislá datová tržiště

- Komplexní architektura
- Konzistence dat
- Detailní i agregovaná data
- Metadata jsou konzistentní




Závislé datové tržiště s odpovídajícími metadaty

# 4. Produkty pro BI

- Nástroje pro BI
  - Microsoft (MS SQL Server, Analysis Services, Reporting Services)
  - Oracle
  - Sybase IQ
  - IBM DB2, DB2 OLAP Server
  - Microstrategy
  - SPSS
  - SAS
  - SAP – Business Warehouse
- Klientské nástroje
  - ProClarity
  - Oracle Discoverer
  - MS Excel 2000
  - Business Objects
  - Cognos: PowerPlay, Impromptu
  - Brio: Brio Query
  - Quadbase – EspressReport

# 4. Agenda BDLC

- Plán projektu a projektový mng
  - Existuje poptávka pro DW, od koho, proč?
    - Poptávka jediného oddělení, informatika, mnoho oddělení
  - Ohodnotit připravenost pro projekt DW
    - Silný business sponzor
    - Pocit potřeby podpory businessu
    - Stupeň práce s informacemi dnes, ochota do budoucnosti
    - Stav IS/IT
    - Proveditelnost
  - Kritický faktor úspěchu podpora mng-u

# 4. Agenda BDLC

- Jak odstranit nepřipravenost?
  - Popsat hlavní potřeby business na konceptuální úrovni
  - Potřeby mng-u
  - Prioritizace business potřeb
  - Proof of concept (pozor na přehnaná očekávání)
- Definice rozsahu projektu – 1. Etapy
  - Řízeno obchodními potřebami ne harmonogramem
  - Spolupráce IT a business
  - Doporučeno jednoduchý obchodní problém řešitelný z jednoho zdroje dat
  - Limit na počet uživatelů (do 25)
  - Určit kritéria úspěchu realizace

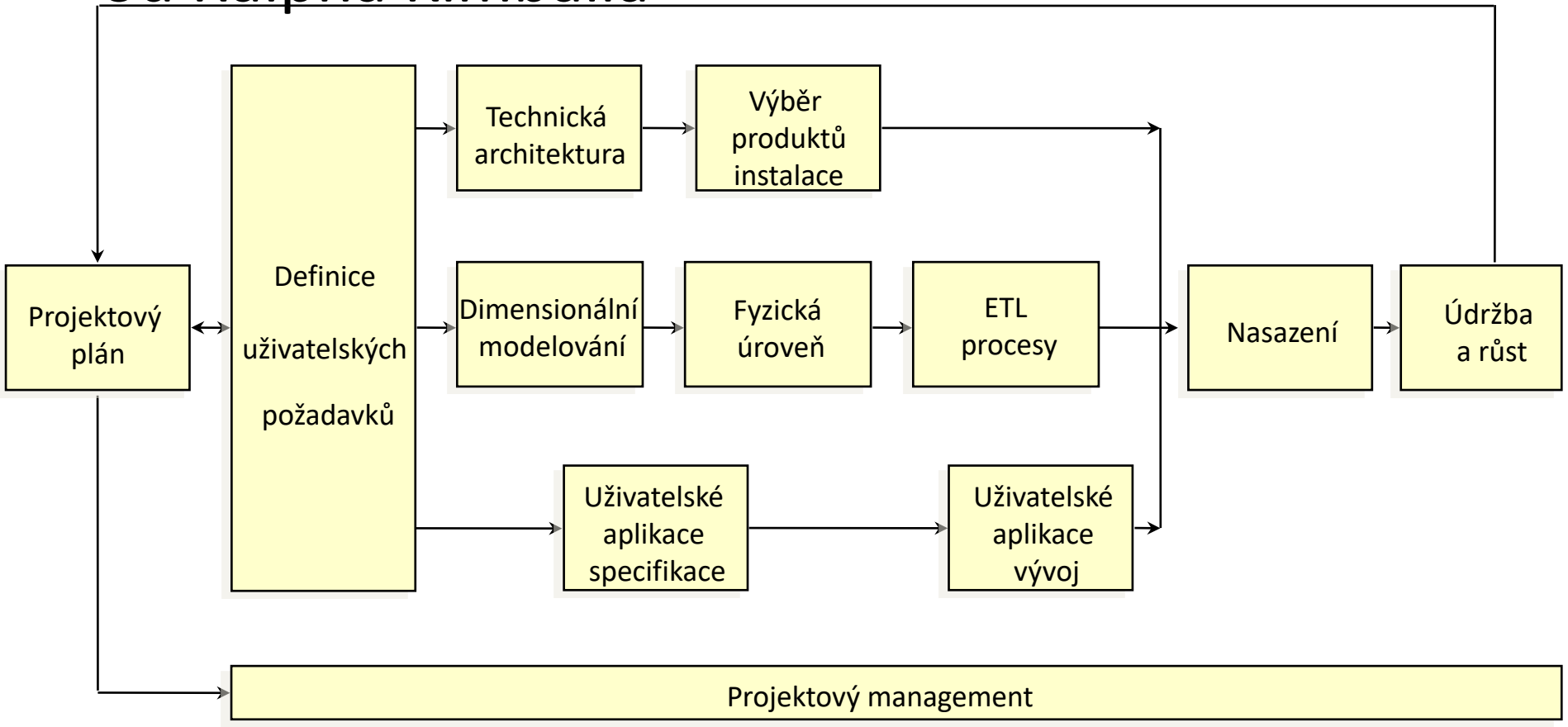
# 4. Agenda BDLC

- Největší riziko: neporozumění kvalitě a problémům v datech
- Zadokumentuj
- Zdůvodnění rozsahu a záměru projektu
- ROI
- Náklady
  - HW, SW
  - na údržbu, interní vývoj, externí vývoj, školení
  - Na podporu ze strany externích řešitelů, na další rozvoj

# 5,6,7 - modelování dat v datových skladech

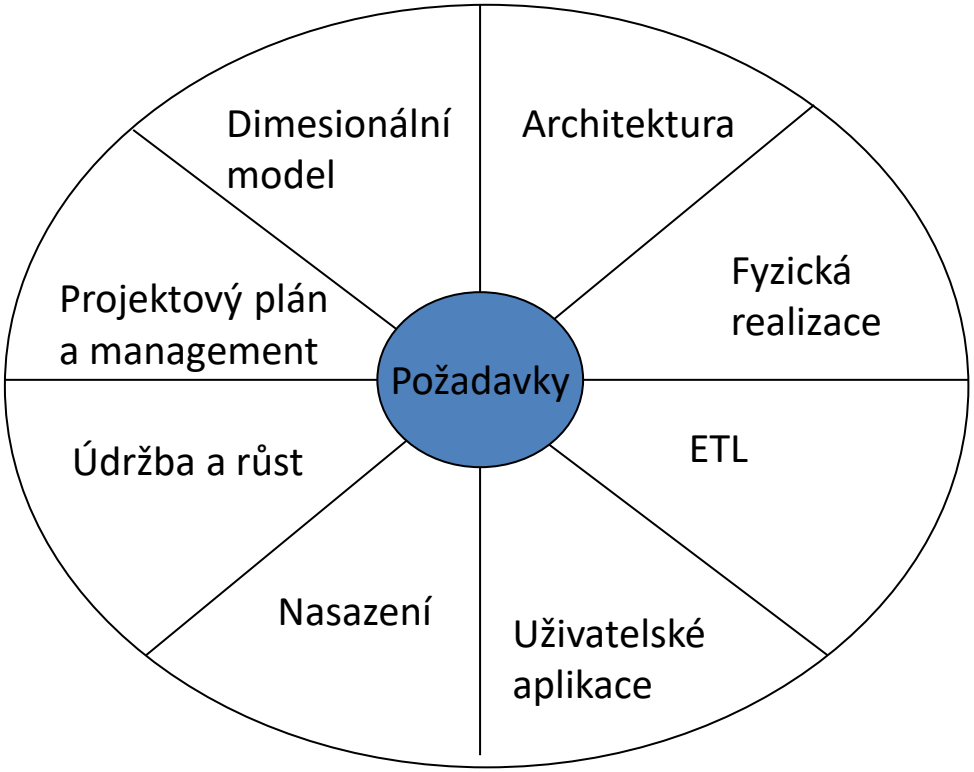
# Business Dimensional LifeCycle

- Standardní metodologie vývoje datového skladu od Ralpha Kimballa



# 5,6,7 - Business požadavky

- Uživatelské požadavky jsou kriticky důležité
- Určují
  - Jaká data budou v datovém skladu
  - Jak budou organizovaná
  - Jak často budou aktualizována
- Dopad na všechny další fáze tvorby DW



# 5,6,7 – Modely

- Konceptuální model
  - Definice základních entit DW a jejich vazby
- Logický model
  - Transformace entit do návrhů logických struktur databázových tabulek, včetně struktury atributů
- Fyzický model
  - Specifikace všech nezbytných technologických charakteristik db tabulek a jejich vazeb

- Dimenzionalitu uložených dat můžeme realizovat, buď v relačních (ROLAP – hvězda, vločky) nebo multidimenzionálních DB (MOLAP – hyperkostka, multikostka)

# 5,6,7 – Konceptuální a logická úroveň

- **Konceptuální modelování**
  1. Obsah DS – entity, atributy, ERD pro základní typy entit a typy vztahů
  2. Rozhodnutí o centrálním DS, DM, komunikace mezi nimi
  3. Přenos ze zdrojů, filtrace, transformace, odvození

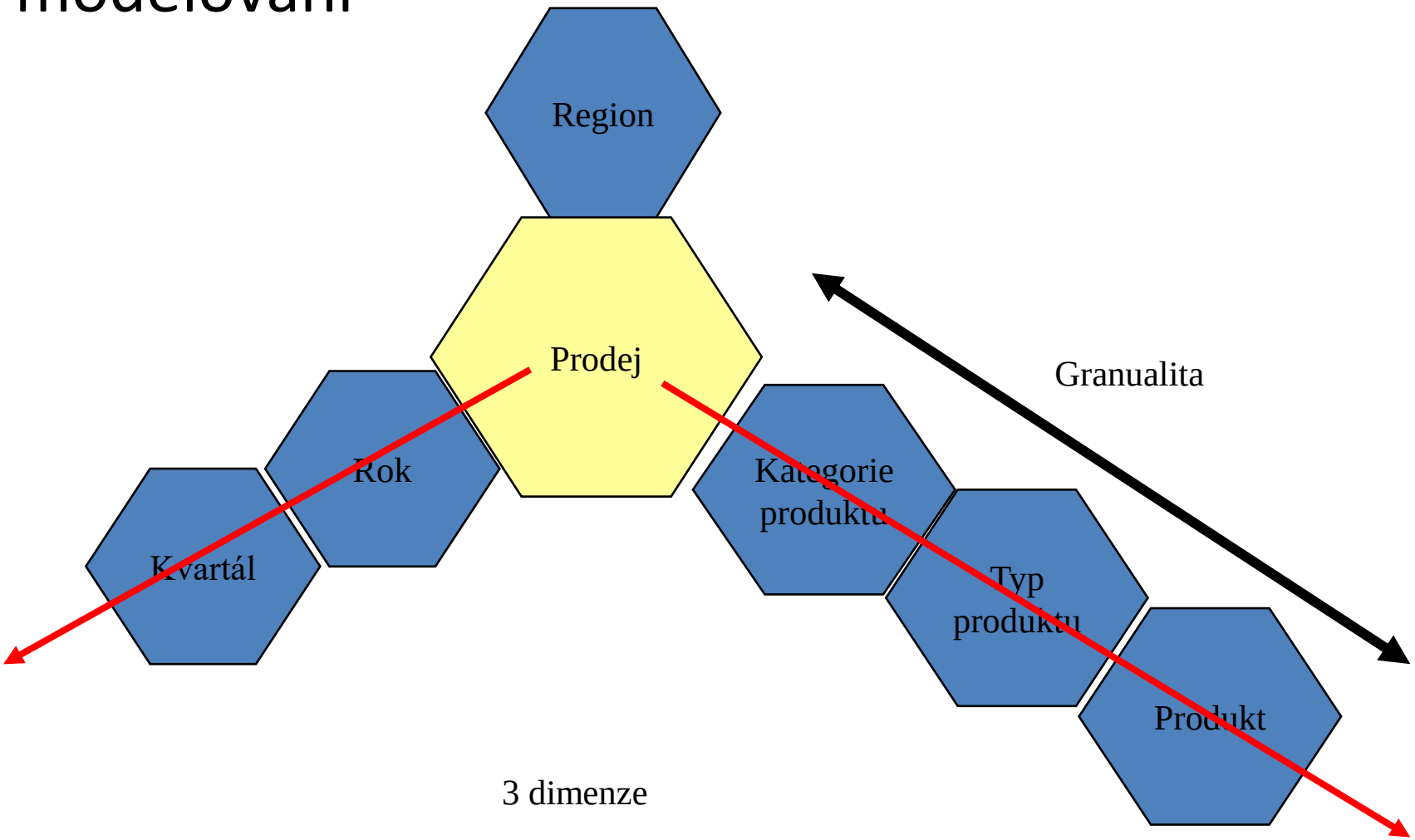
- **Dimenzionální modelování**
  4. Rozdělení atributů na dimenze, fakta, atributy
  5. Definice typů vztahů tvořících základ tabulek faktů – ERD
  6. Definování hierarchie dimenzí
  7. Určení aditivity faktů, Definice omezení dotazů

# 5,6,7 – Dimenzionální modelování

- Je rozdílné od klasického ERD datového model.
- Důraz na srozumitelnost člověku
  - Dosaženo standardní strukturou – fakta, dimenze
- Základní přístup – denormalizace, redundance
- Úkolem vytvořit základní logiku uložení nebo uspořádání
  - Určit dimenze, fakta, hierarchie, vztahy
- Denormalizace
  - Menší počet tabulek, spojení
  - Rychlejší odezva; Většina údajů v jedné tabulce
  - Indexování
- Dva základní typy tabulek: fakta, dimenze

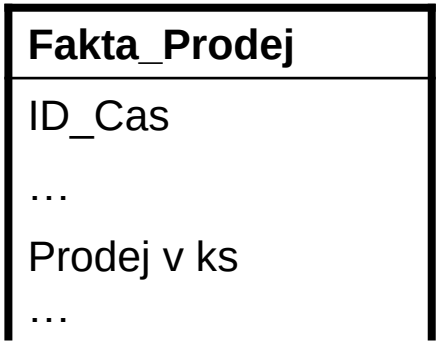
# 5,6,7 - Dimenzionální modelování

- Základní myšlenka multidimenzionálního modelování



# 5,6,7 – DM –Fakta

- Informace o sledovaných ukazatelích
- Velké tabulky s velkým objemem dat
- Sloupce
  - Ukazatele
  - Klíče do tabulek dimenzí
- Př.: množství prodaných produktů, hodnota prodaného zboží, počet zákazníků, …
- Většinou numerické hodnoty, které jdou agregovat
- Mění se v čase



# 5,6,7 – DM –Fakta

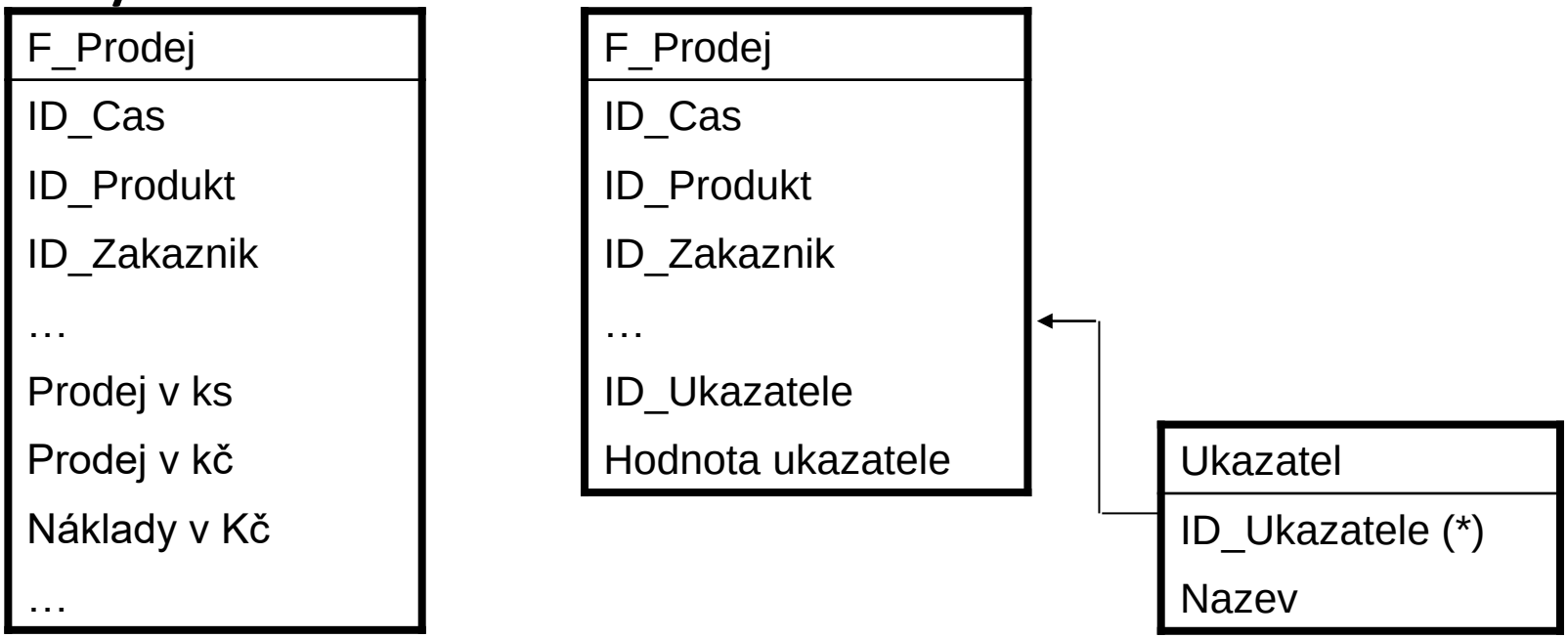
- Přiřazení dimenzí
  - Popisy, které nabývají jedné hodnoty pro jeden záznam ve faktové tabulce
  - Při více hodnotách (M:N vztah) řešit přes pomocné tabulky
- **Granularita** faktové tabulky
  - Míra podrobnosti sledovaných ukazatelů, jejich přesný význam
  - Např.: počet prodaných kusů za den daného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

# 5,6,7 – DM – Fakta - Typy ukazatelů

- **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
- **Neaditivní**
  - Neagregovatelná
  - Např. textová fakta

# 5,6,7 – DM – Fakta

- Existují dva způsoby zápisu ukazatelů do faktové tabulky:



# 5,6,7 – DM – Typy faktových tabulek

- **Transakční**
  - Zachycuje jednotlivé transakce,  jednotlivé akce v daný časový okamžik
  - 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
  - Umožňuje efektivně generovat výstupní reporty s často složitě vypočitatelnými ukazateli

# 5,6,7 – DM – 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 ve 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

# 5,6,7 – DM – Fakta – kroky tvorby tabulky

1. Výběr datového tržiště
   - Jeden datový zdroj vs. více (začít s řešením kde jeden)
2. Určení granularity dat
   - Co nejdetailnější
   - Potřeba přesně vydefinovat, určuje co bude ve fakt tabulce uloženo
   - Určení typu faktové tabulky
3. Výběr dimenzí
   - Vychází z určení gr. plus další dimenze vyhovující navržené gr.
   - Gr. dimenze nemůže být nižší než gr. faktové tabulky
4. Určení faktů (ukazatelů)
   - Vychází z gr. a typu faktové tabulky
   - Ukazatele s rozdílnou gr. (vytvořené např. z důvodu urychlení výpočtu) je třeba uložit do zvláštní faktové tabulky (např. součty pro výpočet procent z detailních ukazatelů, …)

# 5,6,7 – DM – Dimenze

- Tvoří vstupní bod do DW
  - Omezení v dotazech, hlavičky v reportech
- Hierarchické uspořádání údajů
- Umožňují zkoumat data z různých pohledů
- Typické dimenze
  - Čas, zákazník, produkt, prodejna, smlouva
- Atributy
  - Textové hodnoty (nebo se tak chovají)
  - Diskrétní, statické (příliš se nemění)
  - Slouží pro definici omezení a agregace

# 5,6,7 – DM – Dimenze

- Atributy
  - Hierarchické (kategorie – subkategorie – produkt)
  - Nehierarchické (barva)
- Hierarchie mají stromovou strukturu
- Může existovat i více hierarchií
- Sloupce: Identifikátor; Popisné atributy
- Počet dim. cca 5-15
  - Nezávislé, not NULL, popisné, standardizace, kvalitní
- Konformované dimenze
  - Mají stejný význam ve všech připojených fakt. tabulkách (v souhvězdích/sněžení, nezávislých datamartech)

# 5,6,7 - Minidimense a subdimense

- Minidimense
  - Skupina atributů je oddělena do samostatné tabulky, kde každý řádek představuje unikátní kombinaci hodnot
  - Má vazbu na tabulku faktů
- Subdimenze
  - Vypadají jako sněhové vločky, ale mají odlišnou charakteristiku
  - Váže se na tabulku dimenzí
- Minidimenze je obdoba subdimenze

# 5,6,7 - 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. obj.)
- Pro vazbu do produkčních systémů

# 5,6,7 – Fyzická úroveň - Databázové modelování

8. Transformace ERD do
   - ROLAPu (hvězda/souhvězdí, vločka/sněžení, kombinace)
   - MOLAPu (hyperkostka, multikostka)
   - HOLAP (hybrid  ROLAP a MOLAP)

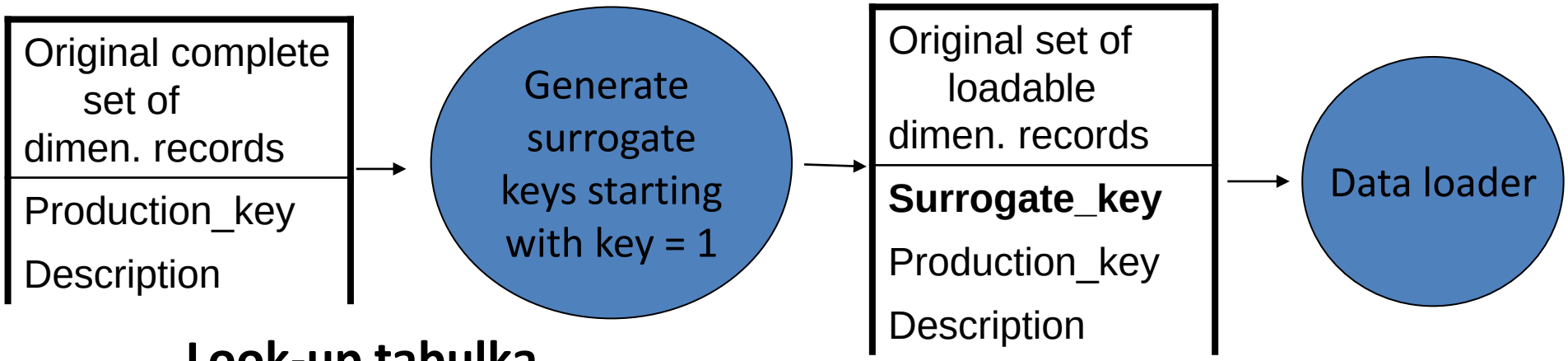
9. Řešení hierarchií dimenzí, volba modelu pro hierarchie pro ROLAP

# 5,6,7 – DBM – Surrogate Keys

- Umělé klíče - místo přirozených prim. klíčů
- Je doporučováno všechny přirozené klíče nahradit umělými
  - Všechny join do fakt tabulek přes umělé klíče
- Důvody
  - Flexibilní, nezávislé na změnách v OLTP systémech
  - Označení hodnoty „Nevím“ v dimenzi
  - SCD (Slowly Changing Dimensions)
  - Rychlejší join
  - Většinou menší nároky na místo než u přirozených klíčů
- I časová dimenze by měla mít umělý klíč
  - Nespojovat přes DATE-TIME (datový typ) atribut

# 5,6,7 – DBM – Surrogate Keys

- Umělé klíče: přiřazené každé dimenzi
  - Hodnoty 1, 2, 3, ….
- Přiřazení hodnot umělého klíče:
  - První načtení; Následné načtení; Využití look-up tabulek




**Look-up tabulka**

| Production_key | Current_surrogate_key |
|---|---|
| SKU43WERT567 | 2345 |
| SKU653TYH7889 | 4567 |

# 5,6,7 – DBM – ROLAP

## Hierarchie dimenzí v jedné tabulce faktů

- 1F – 1 tabulka faktů pro všechny hierarchické stupně
- 1D – 1 tabulka pro každou dimenzi
  - Nejméně prostoru zabere umístění celé hierarchie do jedné d-tabulky
  - Denormalizace, s velkou redundancí a umělou identifik.
- Celý DW se realizuje jako hvězda/souhvězdí se zdrojovými i agregovanými údaji o všech stupních hierarchie

# 5,6,7 – DBM – ROLAP

- Pro rozlišení hodnot i jejich hierarch. stupňů se používají dvě metody
  - Záznamy každé dimenze i jejich hierarchické stupně mají generovaný klíč
  - Záznamy každé dimenze mají samoidentifikovatelný klíč

# 5,6,7 – DBM – ROLAP

- **D-tabulka Prodejce** se skrytou hierarchií a generovanými klíči

| gen_klíč | prod_id | adresa | reprez | obec kraj | kraj | úroveň |
|---|---|---|---|---|---|---|
| 223 | Prod_123 | A | Horák | Ostrava | sever_mor | adresa |
| 239 | NULL | NULL | NULL | Ostrava | sever_mor | Obec |
| 243 | NULL | NULL | NULL | NULL | sever_mor | Kraj |
| 244 | NULL | NULL | NULL | NULL | NULL | všechno |

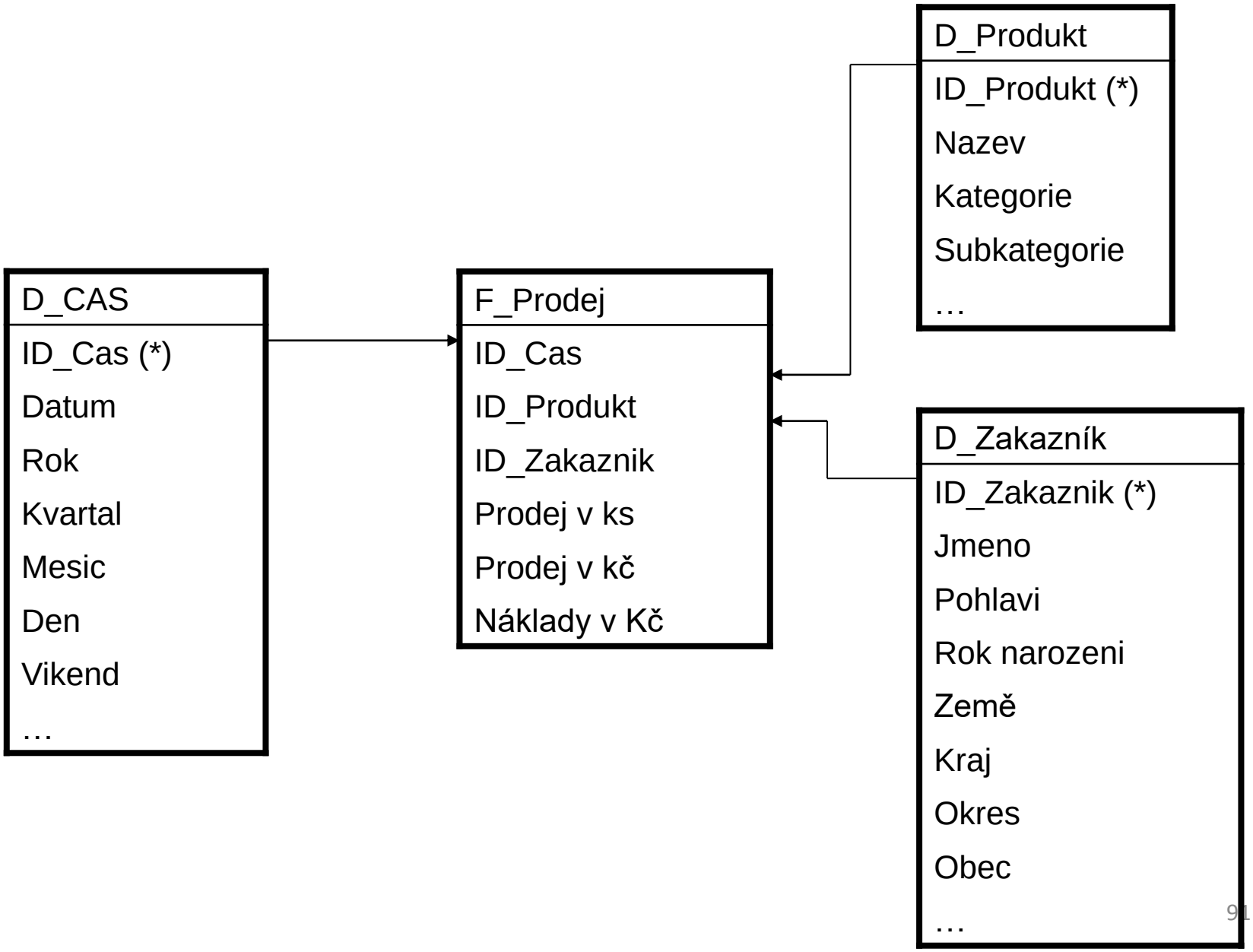
- **F-tabulka Prodej** s hierarchií podél dimenze Prodejce

| typ | prod_ID | barva_ID | poč_zákaz | množství | Cena | zisk |
|---|---|---|---|---|---|---|
| 1 | 223 | 333 | 2 | 3 | 200000 | 20000 |
| 1 | 239 | 333 | 5 | 6 | 500000 | 50000 |
| 1 | 243 | 333 | 10 | 13 | 1000000 | 100000 |

# 5,6,7 – DBM – ROLAP – schéma tabulek - hvězda

- Denormalizovaný tvar
  - Méně spojení, tabulek
- Preferované uspořádání
- Snadněji udržovatelné
- Přehlednější pro uživatele
- Rychlejší odezva
- **Souhvězdí (galaxie)**
  - Schéma s více hvězdami, které sdílejí některé dimenze

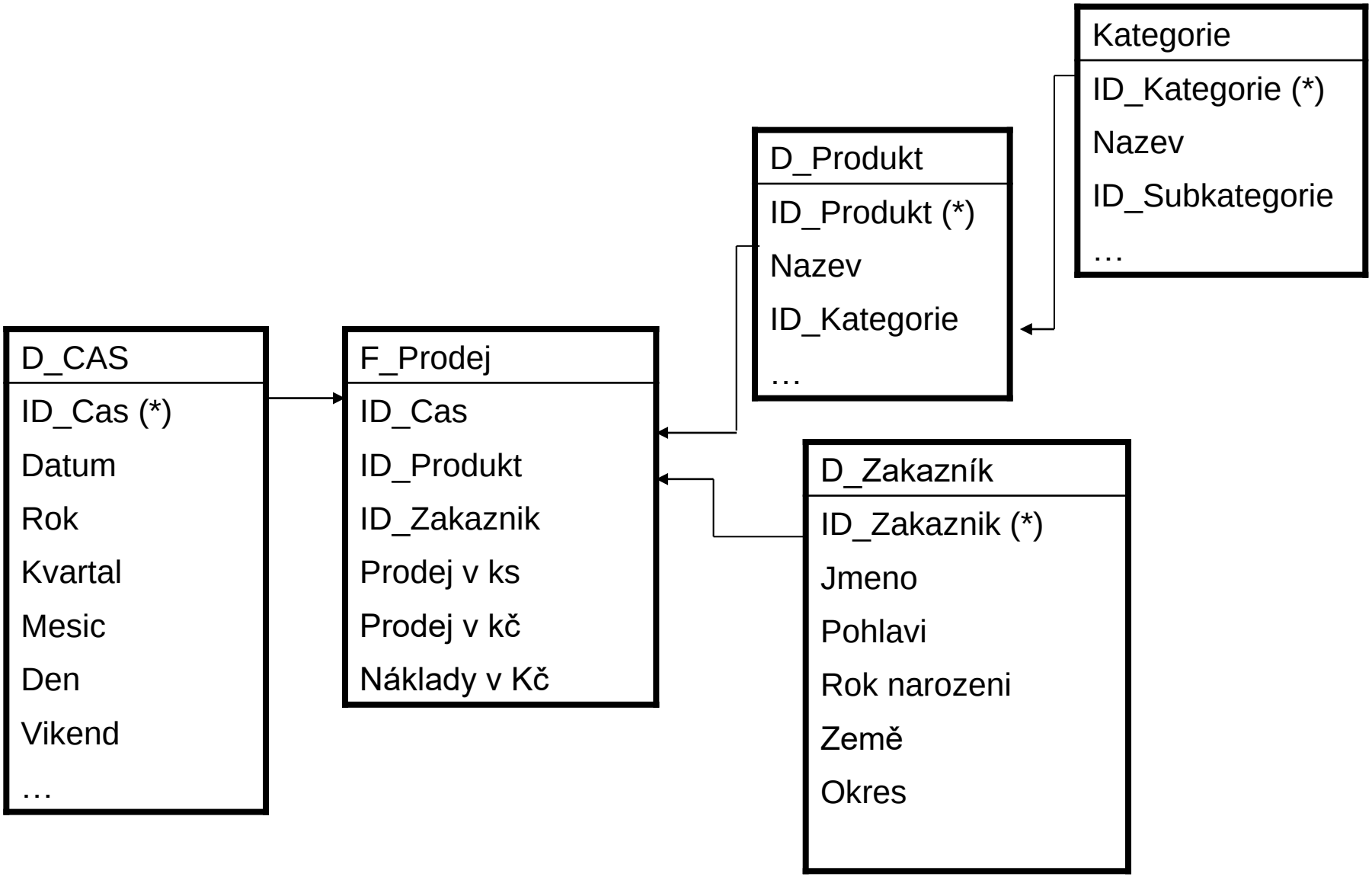
# 5,6,7 – DBM – ROLAP – schéma tabulek - hvězda



# 5,6,7 – DBM – ROLAP – schéma tabulek - vločka

- Normalizovaný tvar, relační struktura
- Méně přehledné
  - Vhodné odstínit uživatele pomocí views
- Náročnější na údržbu
- JOIN – pomalé
- Nutné pro zapojitelnost dalších fakt tabulek
  - BUS architektura
- Vhodné při častých změnách v dimenzích a hierarchické struktuře jejich prvků
- **Sněžení –** více vloček (jako u souhvězdí)

# 5,6,7 – DBM – ROLAP – schéma tabulek - vločka



# 5,6,7 – DBM – ad 9. hierarchie dimenzí ROLAP

- **Hvězdy či souhvězdí** tabulek faktů s hierarchií dimenze v 1 tabulce (F + Di) + generovaný klíč nebo samoidentifikovatelný klíč
- **Souhvězdí** tabulek faktů dle dimenzí, dimenzionální tabulka obsahuje implic. hierarchií
  - ({F1 + Fd1h1 + Fd1h2 + ...} + Di)
- **Sněžení** = sněhových vloček - rozdělení tabulky dimenze podle hierarchie dimenzí
  - ({F + Di}, {Fd1h1 + D1h1, Di}, {Fd1h2 + D1h2, Di}, ...)
  - **Souhvězdí s explicitními hierarchiemi** dimenzí a vazbou mezi stupni dimenze

# 5,6,7 – DBM – MOLAP

- **Hierarchie a jejich dimenze v MOLAP**
- Multidimenzionální DB místo relační
- Místo 2D tabulky je tzv. multikostka – m-dim. tab.
- **Pole** – N_tice **faktů** nejnižší úrovně
  - N_tice - V jedné F-tabulce bývá více faktů
- Každý **Rozměr** tabulky – jedna **dimenze**

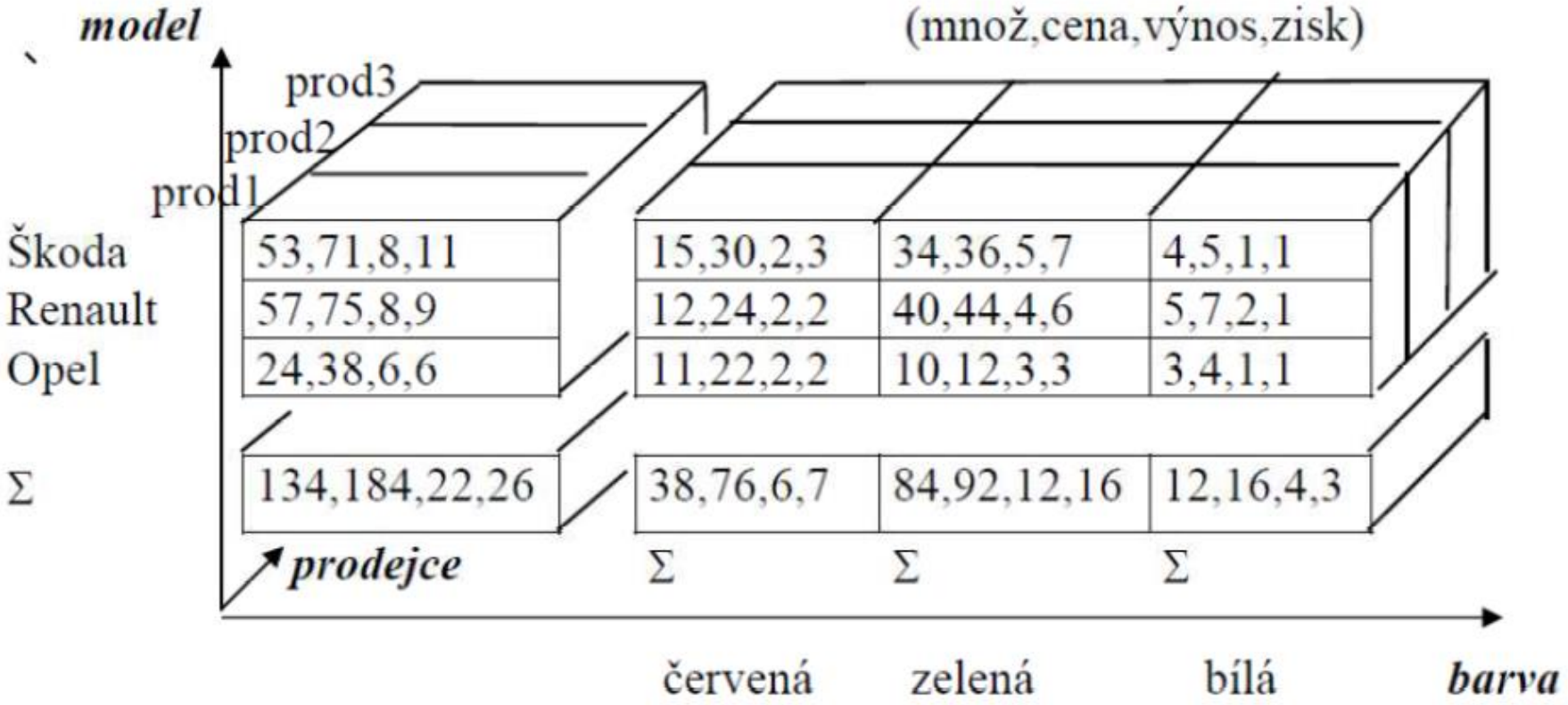
# 5,6,7 – DBM – MOLAP - Hyperkostka

- M-rozměrný prostor dimenzí nad nímž existuje n-rozměrný prostor faktů
- Obsahuje zdrojové i agregované údaje (odvozená data)
- **Dimenze** – osy kostky, každá popsaná jednoatributovým klíčem
- **N-tice faktů** – jednotlivá pole kostky
- => hodnoty n-tice faktů závisí na příslušné m-tici dimenzí

# 5,6,7 – DBM – MOLAP - Hyperkostka

- Model kostky umožňuje zaznamenat hodnoty faktů pro všechny kombinace hodnot dimenzí
- Každá kombinace nemusí být zastoupena
  - => neexistují k ní fakta -> řídká obsazenost
- Z důvodu úspory paměti se hyperkostka někde chápe jen jako logický model
  - Implementují se jenom „obsazené části“ - multikostka

# 5,6,7 – DBM – MOLAP – Hyperkostka



# 5,6,7 – DBM – MOLAP – Multikostka

- Podprostor hyperkostky, v němž nejsou neexistující kombinace dimenzí
- Hyperkostku lze chápat jako sjednocení MK – jejich „obal“
- HK je implementačně jednoduší
- MK efektivnější
- Na logické úrovni – HK, na fyzické – MK
- Blokové MK
  - Fakta i dimenze včetně času tvoří rozměr kostky
- Sériová MK
  - Pro každý fakt existuje samostatná kostka se všemi dimenzemi, fakty tak tvoří dimenzionální řady

# 8,9,10 - Metadata

# 8,9,10 - Metadata

- Metadata – data o datech
  - Data poskytující informace o různých aspektech dat, např. o struktuře (kniha -> obsah)
- Součástí DS je repozitory metadat
- MD jsou formována standardy

# 8,9,10 – Metadata – základní rozdělení

- Technické MD
  - Popis struktury skladu, umístění a obsah DM, …
- Obchodní MD
  - Vlastnictví dat, obchodní pojmy a definice, …
- Operační MD
  - Časová linie dat (historie přesouvání dat, a transformací)
  - Hodnota dat (stav – aktivní, archivovaná, vyloučená)
  - Monitorovací informace (statistiky, loggy, audity)
- Algoritmus pro sumarizaci dat
  - Zahrnuje algoritmy dimenzí, údaje o granularitě, agregaci, sumarizaci atd.
- Mapování z operačního prostředí na DW
  - Zahrnuje zdrojové databáze a jejich obsah a pravidla pro ETL (Metadata pro datovou pumpu)

# 8,9,10 – Metadata

- **Operativní DB**
  - MD koncovým uživatelům v podstatě skryta, pracují s nimi jen vývojáři a správci DB
  - Uživatelé pracují s IS skrz UI
  - Struktura DB – katalogy, data dictionary – integritní omezení, systémové tabulky, pohledy, uložené procedury, …
- **Datový sklad**
  - O datových strukturách musí být uživatelé analytici informování – předpoklad správného použití
  - Pomocí MD lze vyhledat požadovaná data a jejich interpretaci

# 8,9,10 – MD pro správu DW - zdrojová data

## MD zdrojových dat

- Pro potřeby analýzy a návrhu DW
- Rozmístění zdrojových DB na serverech
- Struktura zdrojových DB
- Struktura a popisy entit a jejich vazeb
- Definice a popis atributů, jejich datových typů, domén
  - Včetně měrných jednotek, klíčů, indexů
- Informace o vlastnictví dat a případných vazbách mezi zdrojovými daty (kdo komu poskytuje data)

# 8,9,10 – MD pro správu DW - DW

## Metadata DW

- Seznam serverů
- Rozmístění DB na serverech
- Definice tabulek dimenzí, faktů; pohledů
- Definice a popis atributů, klíčů, indexů
- Rozdělení atributů na dimenze, jejich hierarchie, fakty a popisné atributy, omezení na dotazy

# 8,9,10 – MD pro správu DW – dat. pumpa

## Metadata pro datovou pumpu

- Mapování zdrojových dat z operační DB do cílových atomických dat v primárním DW
- Pro každý atribut DW pravidla
  - pro kopírování, integrační fce, transformační pravidla, změny formátů, verifikace, odvozování, omezení na dotazy
- Měrné jednotky a konverzní koeficienty
- Zvláště pokud jsou to vzorce nebo koef. proměnné v čase
- Informace o časování převodů z DB do DW
- Obchodní pravidla, postupy, vztahy pro výpočet ekonomických ukazatelů
  - používané vzorce a postupy výpočtu

# 8,9,10 – MD pro správu DW – data a fce na pozadí DW

## MD pro data a funkce na pozadí DW

- Podpůrné dočasné datové struktury pro transformace, zobrazení
- Fce pro extrakci, transformaci, zabezpečení kvality; pořadí jejich spuštění, parametry programů
- Popis strategie plnění DW, definice dočasných podpůrných tabulek a jejich funkce

# 8,9,10 – MD pro správu DW – architektura, práva

- **Architektura DW** – pokud existují DM
  - Struktura DW a DM
  - Definice podmnožin DM z DW
  - Pořadí plnění DW a DM
- **Přístupová práva a zabezpečení DW**
  - Informace o uživatelských rolích a jejich právech
  - Informace o uživatelích a jejich rolích

# 8,9,10 – MD pro uživatele

- **Obsah DW**
  - Datové struktury v uživatelsky přívětivé formě s možností výběru
  - Dimenze a hierarchie, fakty a agregované hodnoty
- **Kvalita data**
  - Údaje o spolehlivosti dat, upozornění na data chybná, chybějící
- **Předdefinované dotazy a sestavy dotazů**
  - Katalog výstupních sestav a grafů
  - Význam jednotlivých prvků v sestavách a ve výsledcích analýz
  - Význam popisů, metod a analýz pro uživatele

# 8,9,10 – MD pro uživatele

- **Obchodní pravidla a postupy**
  - Informace o použitých vzorcích a postupech pro ekonomické ukazatele, …
- **Stavové informace**
  - Informace o aktuálnosti dat a stavu skladu
- **Pravidla pro pročištění DW**
  - Kdy je možné data odstranit, archivovat
- **Historie plnění skladu**
  - Objemy, protokoly o chybách, časové plány, doby přenosu a výpočtů, …
  - Záznam o historii je synchronizován se stavovými informacemi

# 8,9,10 – MD optimalizační

- **Definice agregací a jejich umístění**
  - Popis navigace mezi D-tabulkami a F-tabulkami rozsáhlého skladu v ROLAPu pro urychlení přístupu k požadovaným datům
- **Omezení na dotazy**
  - Pro rychlejší plnění DS, menší kapacitu, rychlejší přístup k datům
- **Statistiky DS**
  - Sledování četnosti různých typů dotazů nad datovým skladem
  - Zpětná vazba pro správce skladu
  - Identifikace často/málo sledovaných dat -> úprava DW

# 8,9,10 - MD jako základ pro autom. podpůrných procesů

- **metadata pro extrakce a transformace**
  - přiřazování zdrojových dat cílovým může sloužit jako podklad pro generování skriptů extrakce, integrace a transformace
- **kvalita dat**
  - uživatelé mohou zadávat přípustné hodnoty pro různé atributy, to slouží k odhalení chyb s případnou následnou automatickou opravou,
- **Generování dotazů**
  - zaznamenaná struktura dat slouží ke generování uživatelských SQL dotazů,
- **koncové nástroje**
  - generování nástrojů pro zobrazení struktury tabulek nebo obsahu sumačních tabulek.

# 8,9,10 - Standardizace

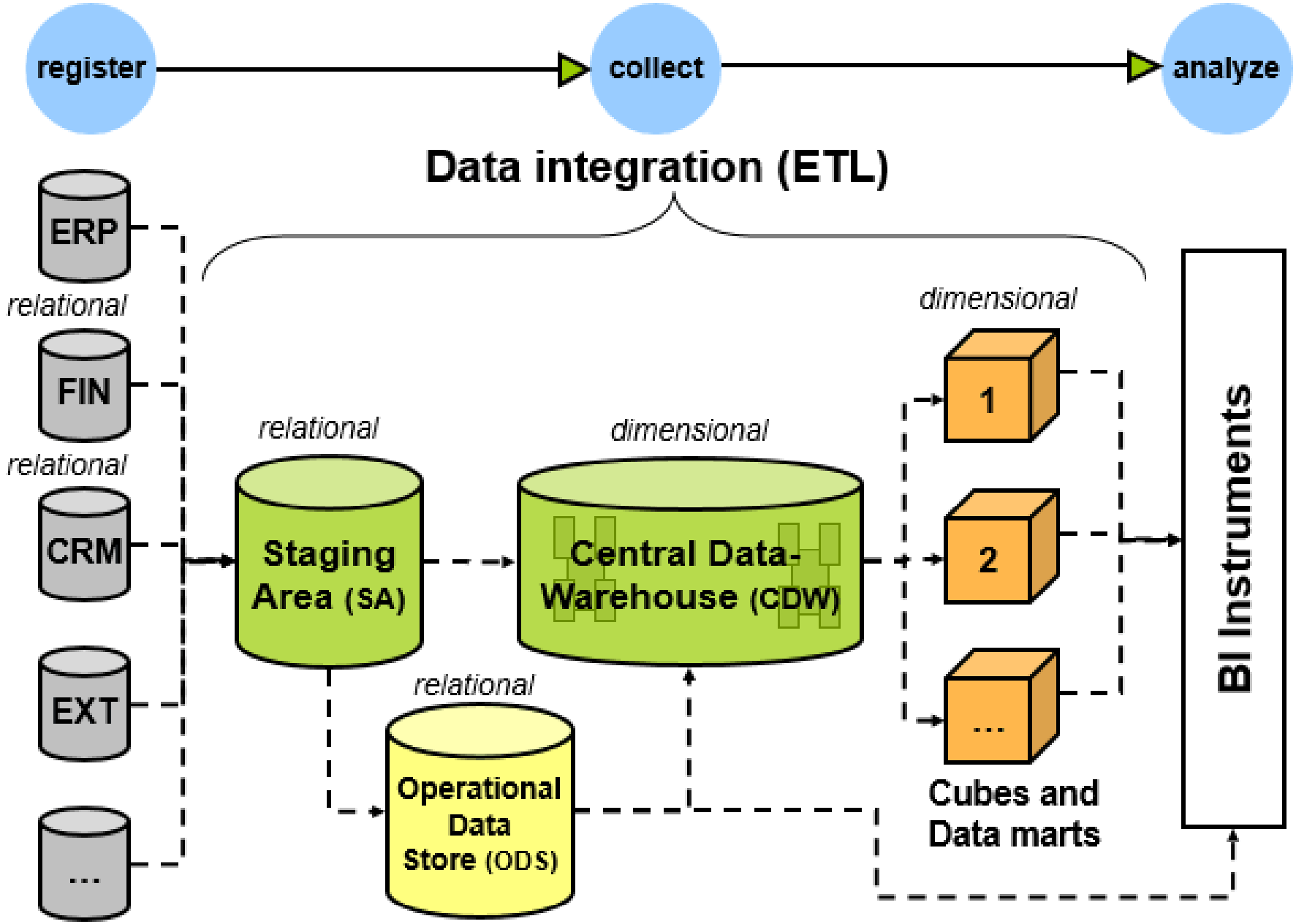
- Obecně použitelné skladiště, tzv. repozitáře
  - Specializované DB pro data o systému; Jednotný přístup
  - *Platinum Repository, Microsoft Repository, Unisys UREP, …*
- Standardy pro výměnu dat
  - *CWMI - Common Warehouse; IBM, Oracle, Unisys, ...*
  - *Metadata Interchange od OMG (Object Management Group)*
- Otevřené API rozhraní produktů
  - Většina dodavatelů řešení poskytuje otevřené rozhraní pro přístup produktů třetích stran a aplikací k jejich metadatům
  - *Hyperion Integration Server, IBM Meta Data Interchange Language*

# 11. Datová pumpa, proces extrakce, transformace a vložení dat, metody čištění dat.

# 11. Vrstvy datového skladu

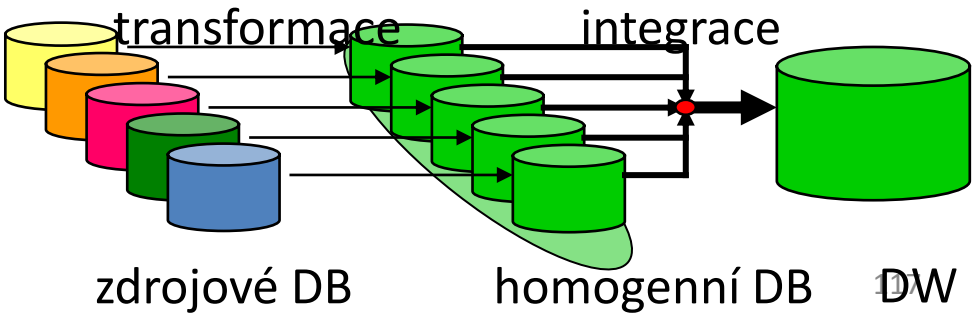
| Vrstva | Popis | ETL náročnost |
|---|---|---|
| 0. vrstva | V nulté vrstvě se uchovávají data z jednotlivých provozních databází. Jedná se většinou o kopie provozních dat 1:1. Data neslouží přímo pro analýzy, ale jako vstup pro další vrstvy. | Převod dat v zásadě 1:1, základní transformační kroky. |
| 1. vrstva | Data jsou uložena v datovém modelu datového skladu (tabulky fakt a dimenzí). Na data jsou aplikována integritní omezení. Tato vrstva slouží pro analýzy. Data jsou očištěná a konzistentní. | Náročné transformace a čištění dat, mapování dat z 0. vrstvy na datový model datového skladu. |
| 2. vrstva | Speciálně připravená data pro podporu speciálních aplikací. V podstatě se jedná o jednotlivá datová tržiště. | Náročné transformace z 0. a 1. vrstvy (speciální algoritmy). |

# 11. ETL


register
collect
analyze
Data integration (ETL)
ERP
relational
FIN
relational
CRM
relational
EXT
...
relational
Staging Area (SA)
dimensional
Central Data-Warehouse (CDW)
relational
Operational Data Store (ODS)
dimensional
1
2
...
Cubes and Data marts
BI Instruments

# 11. ETL – datová pumpa

- ETL – extraction, transformation, loading
- Dva problémy
  - Transformace dat z různorodých zdrojů
  - Integrace dat do datového skladu
- Proces transformace je možné provádět
  - Lokálně
    - v operačním prostředí
    - Zatěžuje – jen u méně zatížených systémů
  - Vzdáleně (stage area)
    - V dočasném uložišti
    - V prostředí DW



# 11. ETL – Extrakce

- PDB -> 0. vrstva DW
- Výběr dat z produkčních databází
- Údaje jsou v různých nehomogenních prostředích a formátech
- Někdy je nutné zpracovat i externí zdroje
  - Analýzy konkurentů, koupená data o zákaznících, volně dostupné údaje z internetu
- K dispozici různé nástroje, postupy, technologie
- Extrakce nesmí ovlivňovat výkon zdrojového systému
- Při správně navrhnuté etapě ETL máme k dispozici metadata (o místě, typu, právech a struktuře zdrojových údajů) pro všechny fáze této etapy

# 11. ETL – Extrakce

- Způsob získání dat
  - Aktualizační notifikace
    - Zdroj může posílat notifikace o změnách a jaké změny byly provedeny
  - Inkrementální extrakce
    - Některé systémy nejsou schopny poskytnout notifikace, ale jenom určit, které záznamy jsou změněny
    - Tyto změny je nutné identifikovat a propagovat dále
    - Nemusíme být schopní správně zpracovat smazaní záznamu
  - Plná extrakce
    - Některé systémy nejsou schopny identifikovat změny
    - Je nutné zachovat předchozí kopii a porovnat ji s novou
    - Zpracovává i smazané záznamy

# 11. ETL – Transformace

- 0. vrstva -> 1. vrstva DW
- Čištění, filtrace, validace, aplikace business pravidel, integrace, transformace do formátu DW
- Převod do stejných dimenzí se stejnými jednotkami
- Spojení dat z několika zdrojů, generování umělých klíčů, řazení, generování agregací, provedení výpočtů ukazatelů a aplikace validačních pravidel

# 11. ETL – Čištění dat

- Nejdůležitější krok, zajišťuje kvalitu dat DW
- Kvalita dat silně závisí na kvalitě zdrojů
- Standardizace/unifikace/normalizace dat
  - Unikátnost identifikátorů
    - Např. Male/Female, M/F, Man/Woman => Male/Female
  - Prázdné/null hodnoty
    - Do standardního tvaru: např. „NotAvailable“
  - Převod telefonních čísel/adres/dateTime/… na jeden formát (může být problém)
  - Jednotná konvence názvů
- Referenční integrita – db může být nekonzistentní

# 11. ETL – Čištění dat

- Kódování textů
- Formáty čísel a řetězců
  - Např. psč, rodná čísla, čísla a datumy uložené jako řetězce
- Různé měny, jednotky – sjednotit
- Detekce anomálií
- Rozdělení hodnot, které slučují více údajů
- Chybějící údaje
  - Malé množství ignorovat, doplnit z jiného zdroje, …
  - řešit podle situace

# 11. ETL – Vložení

- 0./1. vrstva -> 1/.2. vrstva DW
- Vlastní plnění DB DW (0.->1.) viditelných uživatelům
  - Plnění jednotlivých DM (1.->2.)
- Při prvním plnění velké objemy dat
  - Následně iterativně v určený intervalech
- Před provedením plnění je vhodné vypnout indexy a omezení kvůli výkonu, po provedení zase zapnout
  - Referenční integritu zajistí ETL nástroje
  - Aktualizovat indexy

# 12. Technologie implementace datového skladu, plnění datového skladu, indexovací technologie, paralelismus

# 12. Technologie pro implementaci DW

- Vysoké objemy dat v DS a komplikovanější charakter jejich využívání oproti operativním databázím vedou i k jiným fyzickým modelům DB i k jiným technologiím přístupu k datům
- Používají se buď technologie specifické pro DS nebo se aplikují technologie známé i jinde
- Základní požadavek – maximální zrychlení přístupu k datům
- Dosahuje se několika způsoby (i kombinacemi)
  - Inkrementálním plněním skladu i výpočtem agregací
  - Předpočítáním a uložením předpokládaných agregací
  - Rozdělením datového skladu na menší datová tržiště
  - Použitím speciálních indexovacích technologií
  - Využitím paralelního přístupu k datům

# 12. Inkrementální plnění skladu

- Data plníme z ODB periodicky
- Perioda může být různě dlouhá, od denního cyklu až po velmi dlouhý interval
- U kratšího cyklu je nutné dobu plnění – práci datové pumpy – optimalizovat
- Nejdelší čas – výpočet agregovaných hodnot
- Vhodný inkrementální způsob výpočtu agregačních funkcí
  - Suma nová = suma předcházející + suma přírůstku
  - Minimum nové = min (minimum předcházející, minimum přírůstku)
  - Sumy, … počítat průběžně

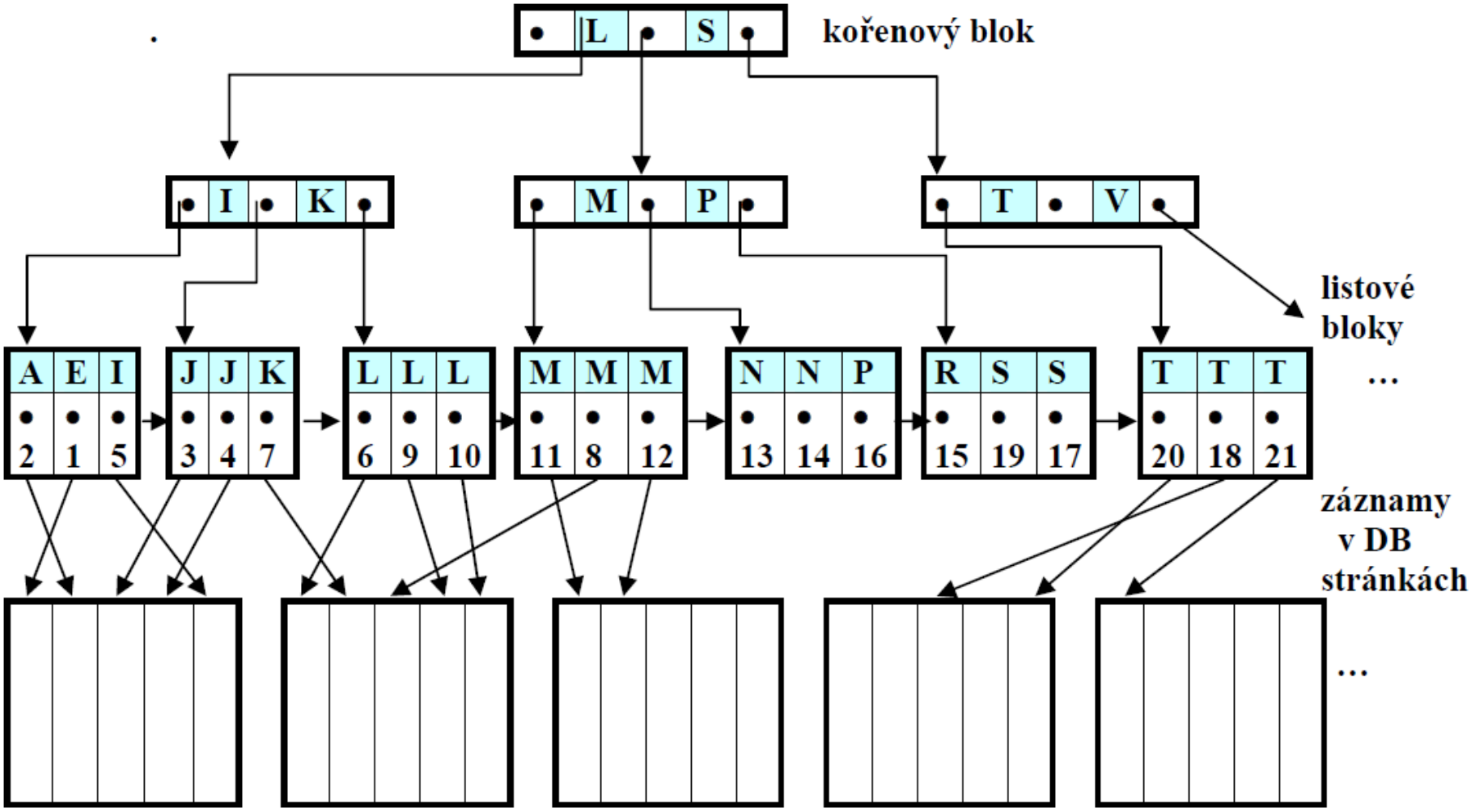
# 12. Indexovací technologie - B/B+ stromy

- B strom
  - Defaultní ve většině relačních databází
  - Indexuje se >= 1 sloupců pomocí B-stromu
  - víceúrovňová stromová struktura
  - Nejnižší úroveň obsahuje bloky listů na každý řádek indexované tabulky
- B+ strom
  - Listy provázány ukazateli
  - umožňující sekvenční procházení tabulkou,
  - dotazování na intervaly,
  - třídění dle indexovacího klíče bez procházení celým stromem

# 12. Indexovací technologie - B/B+ stromy

- Výhodné u sloupců s vysokou kardinalitou (např. jména), protože velikost indexu je nezávislá na kardinalitě sloupce
- Naopak u sloupců s nízkou kardinalitou (např. pohlaví) neefektivní
- Jednotlivé stromy jsou nezávislé – nemůžou spolupracovat na úrovni indexu
- Data v indexu jsou řazena podle klíčů, id řádků jsou neseřazená -> více I/O operací a výpadků stránek
- Efektivnější jiné struktury

# 12. Indexovací technologie - B/B+ stromy



# 12. Indexovací technologie – Bin. index. matice

- Pro sekundární atributy s malou kardinalitou
- V DS a OLAP
- V indexovém souboru jsou **všechny varianty atributu** a označení jejich **umístění** 1 bitem v binární posloupnosti s délkou úměrnou kardinalitě indexovaného sloupce
- Při hledání se projdou všechny indexy daných sloupců a vyhledají 1 u zvolených hodnot. K nim se pak určí pořadí záznamu v datové tabulce
- Vhodné je ukládat indexy transponovaně
  - Pracujeme s celými binárními vektory

# 12. Indexovací technologie – Bin. index. matice

- Např. Oracle, sybase
- Snadná realizace kombinovaných dotazů (log. spojky)
- Vhodné pokud se hodnoty příliš nemění
- Menší velikost oproti jiným strukturám

# 12. Indexovací technologie – Bin. index. matice

**Zam**

| adr | osob | ... | plat | dan |
|---|---|---|---|---|
| 1 | | | 2000 | 30 |
| 2 | | | 4000 | 10 |

**I_plat**

| 2000 | 3000 | 4000 | 5000 |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 0 | **1** | 0 |

**I_dan**

| 10 | 20 | 30 |
|---|---|---|
| 0 | 0 | **1** |
| 1 | 0 | 0 |

| atribut | hodnota | pořadí záznamů 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **dan** | **10** | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | **20** | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | |
| | **30** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | |
| **plat** | **2000** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | **3000** | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | |
| | **4000** | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | |
| | **5000** | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | |

| | | |
|---|---|---|
| **dan = 30 ∧ plat = 4000** | **1** | **1** |

# 12. Indexovací technologie – Join indexy

- Tabulky faktů se spojují s řadou dimen. tabulek
  - JOIN operace je pomalá
- Pomocná tabulka skládající se z několika sloupců,
- Obsahuje index a adresy příslušných záznamů ve 2 nebo více spojovaných tabulkách podle indexovaného spojení.

D_Oddelení

| adr_D | id_odd | mesto |
|---|---|---|
| 1 | 17 | OS |
| 2 | 13 | OP |
| 3 | 7 | BR |
| 4 | 2 | FM |

F_Prodej

| adr_F | id_odd | datum | mnoz |
|---|---|---|---|
| 32 | 2 | 3.3.02 | 3200 |
| 33 | 13 | 1.3.02 | 2500 |
| 34 | 7 | 1.3.02 | 9800 |
| 47 | 17 | 1.3.02 | 3200 |
| 55 | 17 | 2.3.02 | 6500 |
| 65 | 7 | 2.3.02 | 7400 |

Join index: Oddel [*]Prodej

| id_odd | adr_D | adr_F |
|---|---|---|
| 2 | 4 | 32 |
| 7 | 3 | 34 |
| 7 | 3 | 65 |
| 13 | 2 | 33 |
| 17 | 1 | 47 |
| 17 | 1 | 55 |

# 12. Indexovací technologie – Kombinovaný index

- Spojením principů bitmapových indexů a indexů pro spojení dostaneme kombinovaný index.
- Je obdobou předcházejícího indexu, ale tabulka faktů je spojena s dimenzionální tabulkou tak, že k popisnému sekundárnímu atributu je zkonstruován bitmapový index.
- Tak je možné přistupovat k faktům spojeným s dimenzí a omezit hodnoty atributu pomocí bitových operací

# 12. Indexovací technologie – Kombinovaný index

| adr_D | adr_F | OS | OP | BR | FM |
|---|---|---|---|---|---|
| | ... | | | | |
| 4 | 32 | 0 | 0 | 0 | 1 |
| 2 | 33 | 0 | 1 | 0 | 0 |
| 3 | 34 | 0 | 0 | 1 | 0 |
| | ... | | | | |
| 1 | 47 | 1 | 0 | 0 | 0 |
| | ... | | | | |
| 1 | 55 | 1 | 0 | 0 | 0 |
| | ... | | | | |
| 3 | 65 | 0 | 0 | 1 | 0 |

# 12. Indexovací technologie – R stromy

- Speciální modifikace B stromu, zabezpečující efektivnější přístup k rovinným (2D) nebo prostorovým (3D) objektům v relační nebo objektově-relační databázi – vytvořeno pro prostorová data.
- Proti B stromům jsou v listech kromě id_řádků i informace o ohraničení příslušného objektu.
- V blocích vyšších úrovní se uchovávají informace o ohraničení sjednocení objektů nižší úrovně. Pro případ 2D (rovinné objekty) jde například o souřadnice [x, y] pro pravý dolní a levý horní roh jednoho objektu či sjednocení několika objektů.

# 12. Indexovací technologie – R stromy - Rastrové indexy (mřížkové indexy)

- Používanější varianta R stromů - rastrové indexy (také označované mřížkové indexy).

- Mapovaný prostor je rozdělen mřížkou, její jednotlivé bloky jsou očíslovány a je možno je řadit. Vytvořením druhé hustší mřížky se bloky první úrovně opět rozdělí a tak se vytváří varianta B stromu vhodná pro přístup k prostorovým datům.

# 12. Paralelismus

## 1. rozdělení databáze do paralelně přístupných částí (partitioning)

- **horizontální dělení**
  - pomocí **selekce** tabulky se vytvoří disjunktní množiny řádků, ty se umístí do zvláštních fragmentů
- **vertikální dělení**
  - tabulka faktů je **projekcí** rozdělená na víc fragmentů, ty se pomocí operace spojení spojují pomocí redundandního primárního klíče;

# 12. Paralelismus

## 2. symetrický multiprocesing / masivně paralelní procesing (SMP/MPP)

- Paralelní zpracování jedné aplikace na více procesorech současně rozdělením úlohy do vláken
- U DS má smysl paralelně zpracovávat především dva oddělené typy procesů – datovou pumpu a OLAP dotazy, nebo samostatné dotazy různých uživatelů

# 13. Dolování dat (Data mining) - popis dolování dat, úlohy dolování dat, použité aplikace dolování dat, techniky dolování dat- statistické metody, metody umělé inteligence. Získávání znalostí z komplexních dat.

# 13. Data mining

- Získávání znalostí z dat
- Stále větší množství dat uložených v databázích
- Neustále generujeme data
  - Obchodní a bankovní transakce
  - Biologická, astronomická data
- Ukládáme stále více dat
  - DB technologie jsou stále rychlejší a levnější
  - DB systémy jsou schopny pracovat se stále rozsáhlejšími daty
- Data jsou stále rozsáhlejší, ale vyvodit z nich užitečné závěry je stále složitější
  - Velké množství nákupů v supermarketech
  - Miliony hovorů denně

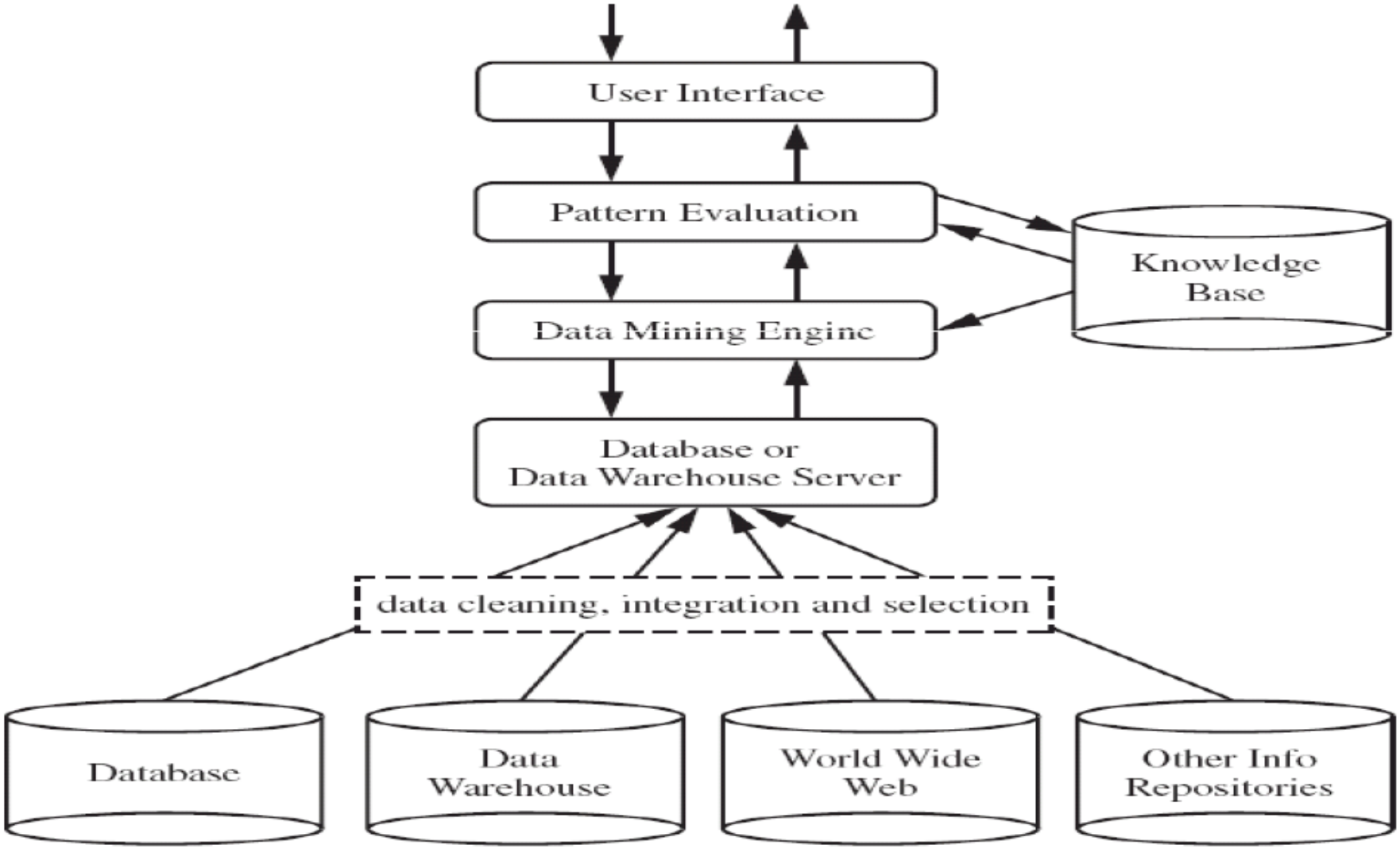
# 13. Data mining

- „Netriviální proces identifikace nových, platných, potenciálně použitelných a snadno pochopitelných vzorů v datech.“ (Frawley)
- DM
  - Hledání nových vzorů, znalostí, které v datech nejsou explicitně uvedeny
  - Znalostí je dosahováno aplikací sofistikovaných alg.
- OLAP
  - Soubor operací (drill-down, roll-up, …) poskytujících různé pohledy na data
  - Výsledků je dosahováno pomocí sumačních a předdefinovaných operací

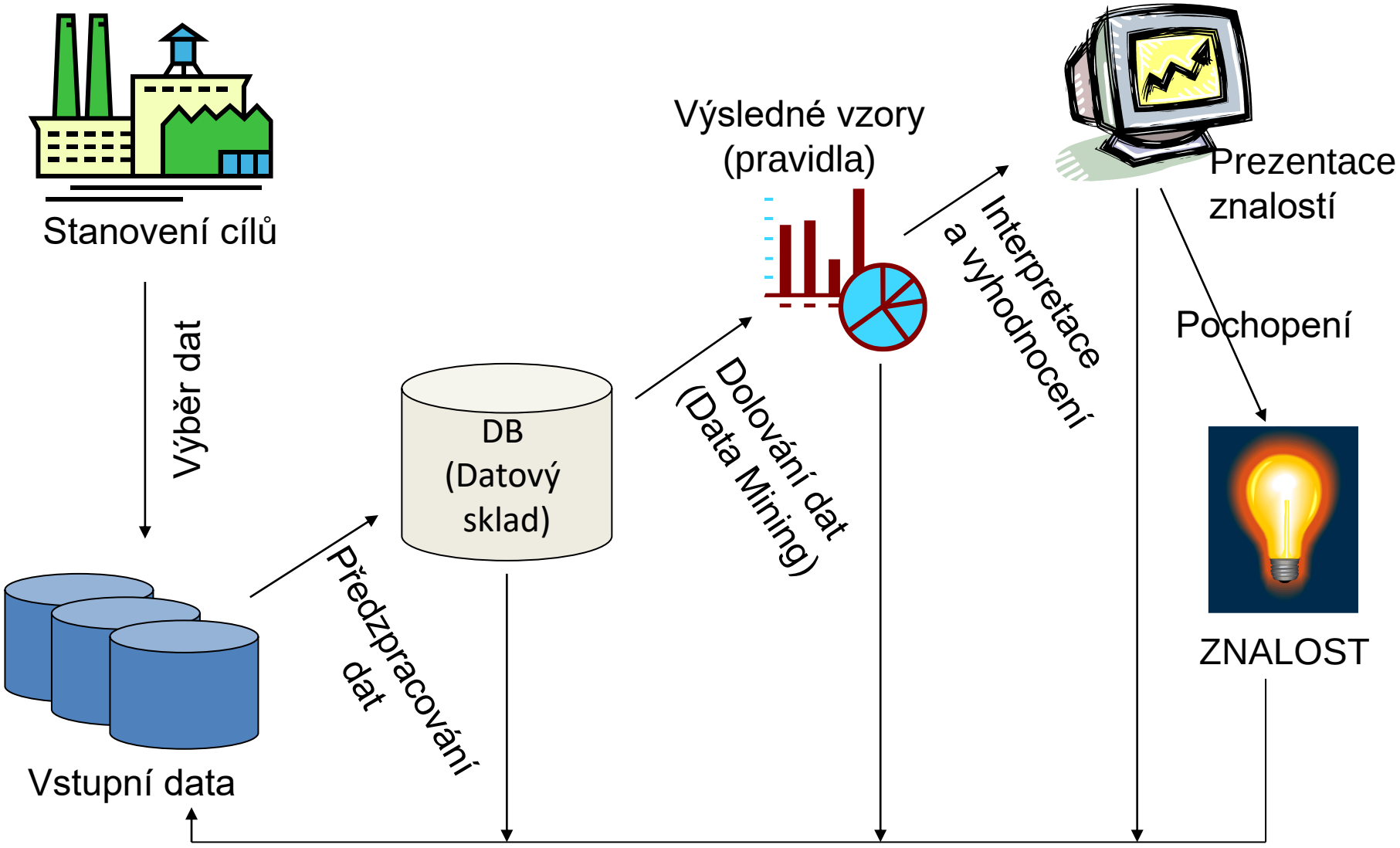
# 13. Data mining

- Základní cíle
  - Predikce dat, popis dat
- Související obory
  - Umělá inteligence, vizualizace, statistika, databázové systémy

# 13. DM - architektura



# 13. Proces získávání znalostí z dat



# 13. Stanovení cílů

- Jaký typ znalosti chceme získat
- Nad jakými daty budeme proces ZZ provádět
- Je problém řešitelný, použitelnost v praxi
- Tvar a forma prezentace
- Vhodnost dané dolovací metody

# 13. Výběr zdrojů dat

- **Typy dat z hlediska zaměření**
  - Demografická data (pohlaví, věk, …)
    - Levné, často neúplné
  - Behaviorální data (nákupy, prodeje)
    - Dražší, nejcennější
  - Psychografická (průzkum veřejného mínění, …)
    - Pomáhají při analýze chování zákazníka
- **Typy DB z hlediska obsahu**
  - Zákaznická DB (údaje o zákaznících, jejich aktivitách)
  - Transakční DB (většinou anonymní údaje o zákaznících)
  - DB historie nabídek (DB o oslovování zákazn. kampaní)
  - Datový sklad

# 13. Výběr zdrojů dat

- **Externí data**
- **Typy dat z hlediska formátu**
  - Relační, transakční, objektově orient. DB
  - Multidim. DB
  - WWW
  - Textové dokumenty
  - Prostorová, časová data, …

# 13. Předzpracování - důvody

- Data v podobě v jaké jsou uloženy v DB většinou nejsou vhodná pro analýzu a modelování
- Mohou být nekompletní, nekonzistentní, obsahovat chybná data
- Nekvalitní data => nekvalitní výsledek

- Fáze
  - Čištění, integrace, redukce, transformace

# 13. Předzpracování – čištění/validace

- Cílem je zajistit kvalitu a najít chybné hodnoty
- Kontrola datových typů, rozsahu atributů, konzistence, porovnání s ostatními instancemi, …
- **Chybějící hodnoty**
  - Není snadné rozeznat proč data chybí (chyba, údaj nebyl k dispozici, …)
  - Odstranění záznamu; Ruční doplnění
  - Nahrazení průměrem/modem atributu / *K* nejbližších instancí
  - Nahrazení 0/konstantou/“unkown“
  - Regresní nebo klasifikační model pro predikci hodnoty

# 13. Předzpracování – čištění/validace

## Identifikace outliers a vyhlazení dat

- **Binning**
  - seřazení, seskupení a případné vyhlazení na základě mediánu/průměru/hranic/… binu
- **Shlukování**
  - podobné hodnoty jsou organizovány do skupin, ostatní jsou chybné
- **Regresní metody**
  - vyhlazení na základě regresní funkce
- **Clustering**
  - nalezení outliers
- **Kombinace lidské a počítačové kontroly**
  - Detekce podezřelých hodnot a kontrola

# 13. Předzpracování – čištění/validace

- **Nekonzistentní data**
  - Vznikají při vkládání dat do DB
  - Při integraci (např. různé názvy atributů)
  - Řešení – ruční oprava, opravné rutiny

# 13. Předzpracování – Integrace

- Integrace více zdrojů do jednoho
- Redundance
  - Identifikace objektu - Objekty můžou mít různé názvy
  - Odvozená data
  - Korelační analýza – např. $\chi^2$ test
- Mapování ekvivalentních entit
- Detekce a řešení konfliktů hodnot atributů
  - Různé kódování, měrné jednotky, …

# 13. Předzpracování – Transformace

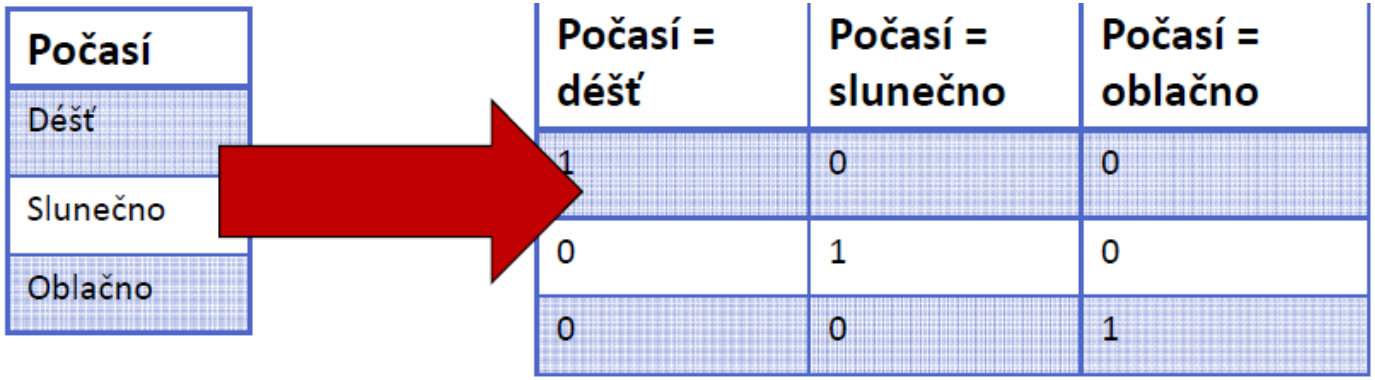
- Transformace do vhodného formátu
- **Slučující techniky** – agregace (numerické atrib.)
  - Sumační operace, …
- **Generalizace** (nominální atr.)
  - Data nižší úrovně nahrazeny vyšší
  - Např. ulice -> město
- **Normalizace** (často [0,1], [-1,1])
  - Přepočítání do daného intervalu (oboru hodnot)
  - Min-max: (v-min)/(max-min)
  - Stan. odchylka, průměr: (v-mean)/(stDev)
- **Přidávání odvozených atributů**
  - Kombinace atrib., které mohou zlepšit model

# 13. Předzpracování – Transf. - Diskretizace (binning)

- metody s učitelem/bez učitele
- Rozdělení spojité veličiny do intervalů
  - Počet intervalů musíme zvolit
- Ekvidistantní
  - Equal-interval binning (equiwidth): Rozdělení do intervalů se stejnou šířkou
  - Equal-frequency binning (equiheight): intervaly se stejným počtem hodnot
- Do hloubky – formování hierarchie
  - Rekurzivní redukce dat nalezením a nahrazením hodnot na nižší úrovni vyšší úrovní (viz generalizace)
  - Např. věk rozdělen do skupin: dítě, dospívající, …

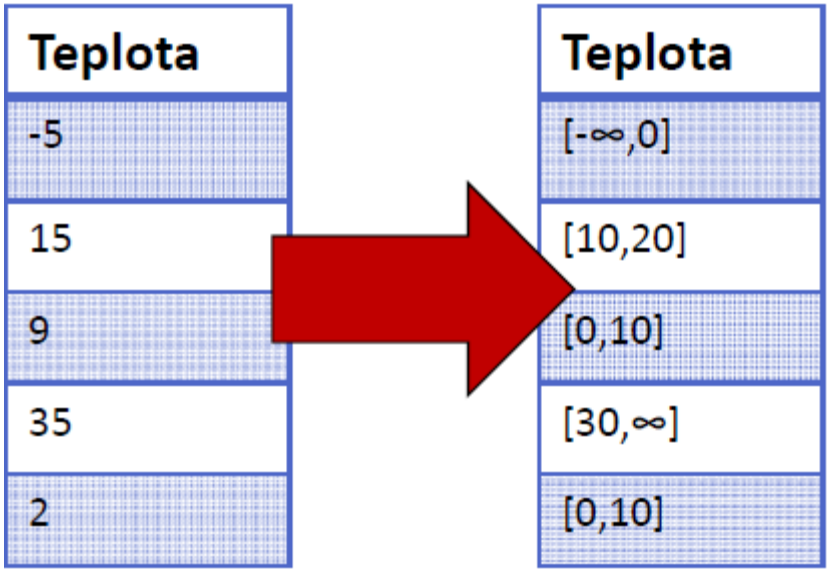
# 13. Předzpracování – Transf. - Konverze typu atributu

- Binominální (2 stavy), nominální (zobecněné binominální), ordinální (množina stavů, existuje uspořádání na množině), numerické atr.
- Řada metod nepodporuje všechny typy atributů
- Nominální -> Binominální
  - Pro každou z různých hodnot nom. atributu vytvoříme nový atribut (kódování 1 z N)
  - Pokud je hodnot hodně rozdělíme je do skupin



# 13. Předzpracování – Transf. - Konverze typu atributu

- Numerické -> Nominální
  - Diskretizace (viz slajd 152)



# 13. Předzpracování – Redukce dat

- Redukce počtu atributů
  - Datová kostka – agregace (roll-up, slice/dice oper.)
  - Odstranění irelevantních atributů
- Redukce hodnot
  - Binning (diskretizace – slajd 152)
  - Clustering
  - Agregace, generalizace
- Komprese – zmenšení objemu
- Vzorkování
  - Výběr reprezentativní podmnožiny hodnot

# 13. Dolování dat

- Aplikace zvoleného alg. na předzpracovaná data dle typu znalostí a dat
- Dělení dle typu znalostí
  - Asociační pravidla
  - Shlukování
  - Klasifikace
  - Predikce

# 13. Dolování – Asociační pravidla

- Proces získávání asociačních pravidel z databází hledá zajímavé vztahy mezi velkým množstvím datových položek.
- Př. analýza nákupního košíku – jaké zboží je opakovaně nakupováno současně
- Původně pro transakční data

- Pravidla ve tvaru: A => B (A, B množiny dat)
  - předpoklad => závěr, IF předpoklad THEN závěr
- Interpretace
  - Jestliže transakce obsahuje položky z A, pak také pravděpodobně obsahuje i položky z B

# 13. Dolování – Asociační pravidla

- **S – podpora** (support): S(A=>B) = P(A U B)
  - Pst, že se v databázi vyskytují A i B
- **C – spolehlivost** (confidence): C(A=>B) = P(B|A)
  - Podmíněná pst, že se vyskytuje v transakci množina B, za předpokladu, že se tam vyskytuje i A
- **Silné asociační pravidlo**
  - Pravidlo, které má $S > s_p$ a $C > c_p$ , kde $s_p$, $c_p$ je mez
- **Frekventovaná množina**
  - Množina, která má $S > s_p$
- **Základní postup**
  - Výpočet FM: na základě minimální podpory
  - Generování SAP z FM: na základě min. spolehlivosti

# 13. Dolování – Asociační pravidla – Apriory alg.

- Založeno na postupném generování kandidátů na frekvenční množiny(mn. u nichž není rozhodnuto)
- **1. vyhledáme FM položek *l*; (Li – množina FM)**
  - Apriorní vlastnost (znalost)
    - jestliže množina *k* není frekventovaná, tak ani *k+1* není FM
  - Vyhledáme FM množiny s kardinalitou 1 (L1), L1 použijeme k sestavení L2, … (pro každou Li je třeba nový průchod daty),
  - Pro položky $L_{i-1}$ provádíme spojovací fázi, výsledná množina je kandidát $C_i$ na $L_i$
  - Pro $C_i$ provedeme vylučovací fázi, výsledná mn. je $L_i$

## 13. Dolování – Asociační pravidla – Apriory alg.

- **Spojovací fáze**: Spojíme dvě stejně velké množiny, které se liší jenom jedním prvkem (pokud budou lexikograficky seřazené, tak $l_i < l_j$)
- **Vylučovací fáze**: Vyloučíme množiny, jejichž některá podmnožina není FM (podmnožiny i počty výskytů můžeme uchovávat v hashovacím stromu/tabulce)

- **2. Generování silných asociačních pravidel**
  - Založeno na výpočtu spolehlivosti
  - C(A=>B) = P(B|A) = S(A U B)/S(A);  S = počet výskytů
  - Pro každou FM *l* vygenerujeme všechny podmnožiny *s* a testujeme C(s=>(l-s)) = S(l)/S(s) >= $c_p$
  - Je-li nerovnost splněna, vypíšeme AP: **s=>(l-s)**

# 13. Dolování – Víceúrovňová asociační pravidla

- Důvod: málo silných asociačních pravidel
- Položky se sdružující do skupin (konceptů), musí být definována tzv. konceptuální hierarchie položek
- Potraviny -> (nápoje -> (džus V cola) V pečivo)

# 13. Dolování – Asociační pravidla v relačních db

- Kategorické atributy
  - Mají konečný počet hodnot
  - Lze na ně použít známé modifikované metody pro transakční data, např. algor. Apriori
- Kvantitativní atributy
  - Nemají konečný počet hodnot
  - Nutnost diskretizace (pak již lze považovat za kategorický):
    - Základní problém asociačních pravidel v relačních datech
  - Metody diskretizace
    - Základní (do hloubky, ekvidistantní)
    - Pokročilé
      - Postupné spojování menších intervalů ve větší
      - Shlukovací metody – jsou hledány shluky hodnot ležící blízko sebe, ty pak vytvoří interval

# 13. Dolování - Sekvenční vzory

- Podobné jako frekventované množiny, ale hraje zde roli čas
- Př. zákazník si koupí výrobek1 a později i výrobek2
  - SV („výrobek1“, „výrobek2“)
- Důležité pořadí položek ve vzoru

# 13. Dolování - shlukování

- Učení bez učitele
- Roztřídění objektů do skupin, které nejsou předem stanoveny
- Rozdíly uvnitř shluků musí být minimální, mezi shluky maximální
- **Rozdělovací metody**
  - Rozdělení na předem daný počet shluků
  - K-means
- **Hierarchické metody**
  - Automatické vytvoření hierarchie
  - Top-down (divise cl.), bottom-up (HAC)
  - Ukončení dělení/spojování při splnění podmínky
- Další (např. neuronové sítě, …)

# 13. Dolování – Shlukování – K-means

- Shlukované objekty lze chápat jako body v euklidovském prostoru
- Vzdálenost – např. euklidovská metrika
- Počet shluků *k* je předem dán nebo získáme minim. cenové funkce
- Shluky jsou definovány centroidy $\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$
- Optimalizační kritérium – stanovení optim. *k*
  - Minimalizace:
    - $\mu_c^{(i)}$ – index centroidu ke kterému je přiřazen vzorek $x^{(i)}$

$$\frac{1}{m} \sum_{i=1}^{m} ||\mathbf{x}^{(i)} - \mu_{c^{(i)}}||^2$$

# 13. Dolování – Shlukování – K-means

- Kroky
  - Výpočet shluků – přiřazení objektu k nejbližšímu centroidu
  - Aktualizace centroidů pro nové shluky
- Inicializace
  - Náhodný výběr *k* hodnot z trénovací množiny
  - Prvních *k* hodnot
  - Výběr heuristikou
- Vždy konverguje, ale nevíme jak dlouho to potrvá
- Konvergence neznamená, že konverg. k optimu

# 13. Dolování – Shlukování – hierarchické metody

- **(TD) Divise clustering**
  - Všechny obj. v jednom clusteru
  - Rekurzivně dělíme

- **(BU) Hierarchical Agglomerative Clustering**
  - Vytváří hierarchii ve formě bin. stromu
  - Každý objekt ve svém clusteru a rekurzivně spojujeme clustery, které jsou si nejpodobnější
  - Podobnost: single-link, complete-link, centroid, group-average)

# 13. Dolování – Shlukování - Příklady aplikací

- Marketing – možnost identifikace skupin zákazníků, použití cílených reklam
- Plánování města – iden. skupin domů podle typu, ceny, polohy
- Studie zemětřesení – epicentra zemětřesení dle jejich vlastností
- Pojištění – hledání potenciálních zákazníků s vysokým povinným ručením
- Geografie – hledání shluků pozemků na základě jeho typu

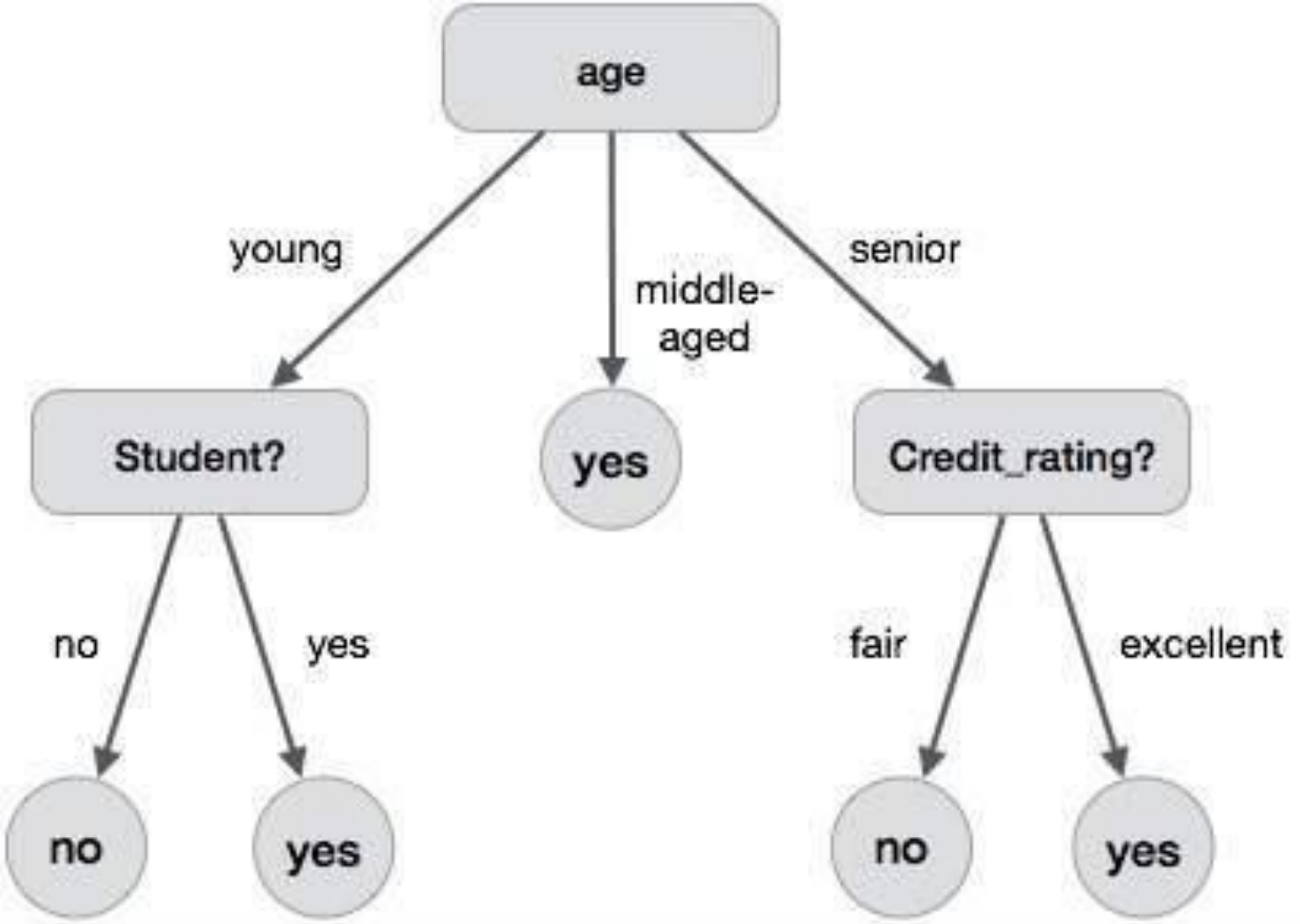
# 13. Dolování – Klasifikace

- Rozdělení objektů do předem známých skupin
- Učení s učitelem
- Daná trénovací množina zařazených objektů
- Podle pravidel (rule-based)
  - Ve tvaru IF podmínka THEN závěr
    - Podmínka – test jednoho nebo více atributů
    - Závěr – předpověď třídy
  - Př.: IF age = youth AND student = yes THEN buy_computer = yes
  - Lze převést na rozhodovací strom
- Statistické/pstní metody
  - Naive bayes, rocchio, kNN, …

# 13. Dolování – Klasifikace – Rozhodovací stromy

- Znalosti reprezentovány v podobě stromu
  - Uzly – třídy dat;  Větve – struktura třídy
- Postup metodou rozděl a panuj
  - Trénovací množina se postupně dělí tak, aby v jedné množině převládala data jedné třídy
- Vhodné pro kategorická data
- 1. Alg. konstrukce stromu (pro trénovací data)
  - Zvolit kořen – jeden atribut
  - Data podle něj rozdělit na podmnožiny – nové uzly
  - Pokud existuje uzel, který není homogenní -> krok 1
- 2. Klasifikace na základě vytvořeného stromu

# 13. Dolování – Klasifikace – Rozhodovací stromy



# 13. Dolování – Predikce

- Proces určení dodatečných/chybějících hodnot
- Klasifikace – natrénovaní, určení třídy
- Regrese
  - Pomocí modelu jsou dopočítány numerické atributy spojitého charakteru
  - Trénování: cílem je získat parametry regresní funkce, která co nejvíce odpovídá daným datům (lineární, polynomiální, …)
  - Alg.: Generalized Linear Models, Support Vector Machines

# 13. DM a datové sklady

- DM hraje důležitou roli v prostředí datového skladu
- Společné znaky
  - Velké množství dat, většinou na detailní úrovni – ale ne vždy jsou tam všechna data
  - Data Mining nejlépe pracuje s integrovanými a vyčištěnými daty
  - Máme-li datový sklad, není potřeba investovat do HW pro data mining

# 13. Využití DM

- Členění (segmentace) zákazníků
  - Cíl: porozumět zákazníkovi a jeho chování
- Analýza nákupního košíku
  - Nalezení závislostí mezi různým zbožím, které si zákazník koupí
- Management rizik
  - Odhalení rizikových zákazníků (např. u pojišťoven)
- Detekce podvodů
  - Např. hledání extrémních útrat na kreditní kartě
- Odhalování zločinnosti
  - Odhalení potenciálních neplatičů půjček…
- Predikce požadavků
  - Předpověď zájmu zákazníků o různé zboží…

# 13. Dotazovací jazyky pro data mining

- Data mining by měl být interaktivním procesem
- Základ pro uživatelské rozhraní
- Standardizace
- Součásti dotazu pro data mining
  - Relevantní data
  - Typ znalosti
  - Doménová znalost
  - Metriky zajímavosti
  - Vizualizace/prezentace získaných znalostí

# 13. Dotazovací jazyky pro DM – součásti dotazu

- **Relevantní data**
  - Jméno databáze/datového skladu
  - Databázové tabulky/kostky
  - Podmínky pro selekci dat
  - Relevantní atributy nebo dimenze
  - Kritéria pro seskupování dat
- **Typ získávané znalosti**
  - Asociační pravidla, shlukování, klasifikace, …

# 13. Dotazovací jazyky pro DM – součásti dotazu

- **Doménová znalost**
  - Typické využití
    - Konceptuální hierarchie
  - Stromová hierarchie
    - město – kraj – země – světadíl
  - Seskupovací hierarchie:
    - Např.: (15-39) – mladý; (40-59) střední věk
  - Hierarchie založená na pravidlech
    - nízký_zisk(X) = cena(X) = p AND náklady(X) = q AND p-q < 50$
  - Hierarchie odvozená z operace
    - emailová adresa: hagonzal@cs.uiuc.edu – login – ústav – univerzita – země

# 13. Dotazovací jazyky pro DM – součásti dotazu

- **Metriky zajímavosti**
  - Jednoduchost – počet prvků pravidla, velikost rozhodovacího stromu
  - Použitelnost – např. podpora a spolehlivost
  - Jedinečnost – odstranění podobných znalostí
- **Prezentace/Vizualizace**
  - Různé formy reprezentace – grafy, tabulky…
  - Reprezentace konceptuální hierarchie
  - Vizualizace různých typů znalostí

# 13. Získávání znalostí z komplexních dat

- **Prostorové databáze**
  - Nutnost předzpracování…
  - Příklad asociačního pravidla
- **Multimediální databáze**
  - Konstrukce vektoru rysů
  - Histogramy
  - Identifikace objektů v obrázku
- **Časová a sekvenční data**
  - Obsahují sekvence hodnot a událostí závislých na čase
  - Použití v meteorologii, lékařství (krevní tlak), burza (inflace, ceny akcií)

# 13. Získávání znalostí z komplexních dat

- **Textové databáze**
  - Velké kolekce dokumentů…
  - Hledání podobných kolekcí dokumentů obsahujících zadaná slova
  - Asociační pravidla založená na klíčových slovech
  - Klasifikace dokumentů
- **World-Wide-Web**
  - WWW dokumenty
  - Databáze s informacemi o přístupu…
- **Objektové databáze**
  - Lze použít upravené metody pro získávání znalostí z relačních dat
- **XML**

# 14. Problém zpracování velkých dat, vizualizace.

# 14. Big data

- **Volume** – množství dat vznikajících v rámci provozu firem roste exponenciálně každý rok,
- **Velocity** – rychlost s jakou data vznikají a potřeba jejich analýzy v reálném čase vzrůstá díky pokračující digitalizaci většiny transakcí, mobilním zařízením a vzrůstajícímu počtu internetových uživatelů
- **Variety** – různorodost typů dat vzrůstá, například nestrukturované textové soubory, semi-strukturovaná data (XML), data o geografické poloze, data z čidel, videa apod.
- **Veracity -** nejistá věrohodnost dat v důsledku jejich nekonzistence, neúplnosti, nejasnosti a podobně. Vhodným příkladem mohou být údaje čerpané z komunikace na sociálních sítích. Je nutné je čistit, vzájemně propojovat a korelovat jinak se snadno vymknou kontrole

# 14. Big data

- **Objem** - dat je příliš mnoho na to, abychom je uměli zpracovat. Možné řešení je při získávání dat provádět analýzu a vybírat jen relevantní. Další možnost je např. užití distribuovaných výpočtů či gridů.
- **Rychlost** - data potřebujeme zpracovávat téměř v reálném čase. Hlavně nároky na odezvu. Často se řeší pomocí NoSQL databází, kde v rychlém zpracování nehledíme na celek, ale zpracováváme jen některé jeho podstatné informace.
- **Nestrukturovanost** - stojíme před problémem, jak vyhledávat v databázích multimediálních dat jinak, než pomocí metadat či textových popisků (třeba pomocí porovnávání se vzory ve znalostní databázi – identifikujeme ve filmu Eiffelovu věž a víme, že se odehrává v Paříži atp.).
- **Nehomogenita a nekonzistence** - třeba v případě analýzy dat ze sociálních sítí je problém, že každá vypadají trochu jinak.

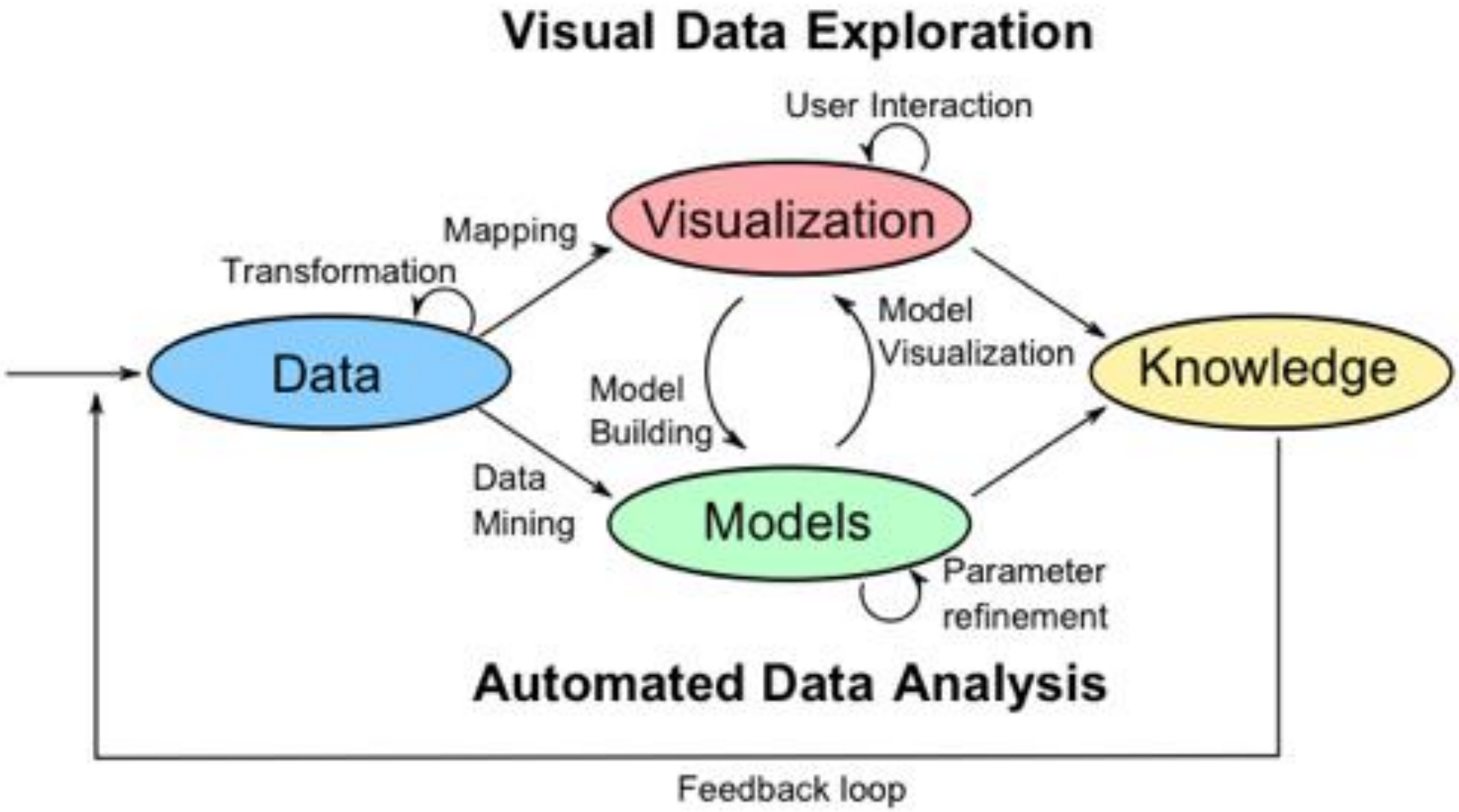
# 14. Problém zpracování velkých dat

- Velké objemy dat
- Trpí efektivita zpracování
- Standardní analytické nástroje selhávají
- Potřeba inteligentnějších a efektivnějších nástrojů a metod podporujících analytický proces

# 14. Visual Analytics

- VA kombinuje automatické analytické techniky s interaktivní vizualizací, pro efektivní porozumění, odvozování a rozhodování na základě velkých a komplexních datových množin
- Zprůhlednění celého analytického procesu
- Vizualizace informací a interakce s daty
  - Lepší přehled a jednoduší rozhodování
- Iterativní proces
  - Získání dat, předzpracování, reprezentace informací, interakce, vyvozování

# 14. Visual Analytics - Proces



# 14. VA - Proces

- Předzpracování

- Volba mezi vizuální a autom. metodou analýzy
- Střídání vizualizačních a analytických metod
- Vylepšování na základě verifikace předchozích (mezi)výsledků
- Postupné vylepšování modelu umožňuje dříve odhalit problémy
  - Chyby v předzpracování; Chyby ve zdrojových datech
  - Nevhodný postup analýzy;
  - => Kvalitnější a důvěryhodnější výsledky

# 14. VA - Proces

- Znalosti mohou být získány:
  - Vizualizací
  - Analytickými metodami
  - Interakcí analytika s vizualizacemi a modely
- Poznatky získané při vizualizaci jsou užitečné při dalším směrování analýzy

# 14. VA - Proces - vizualizace

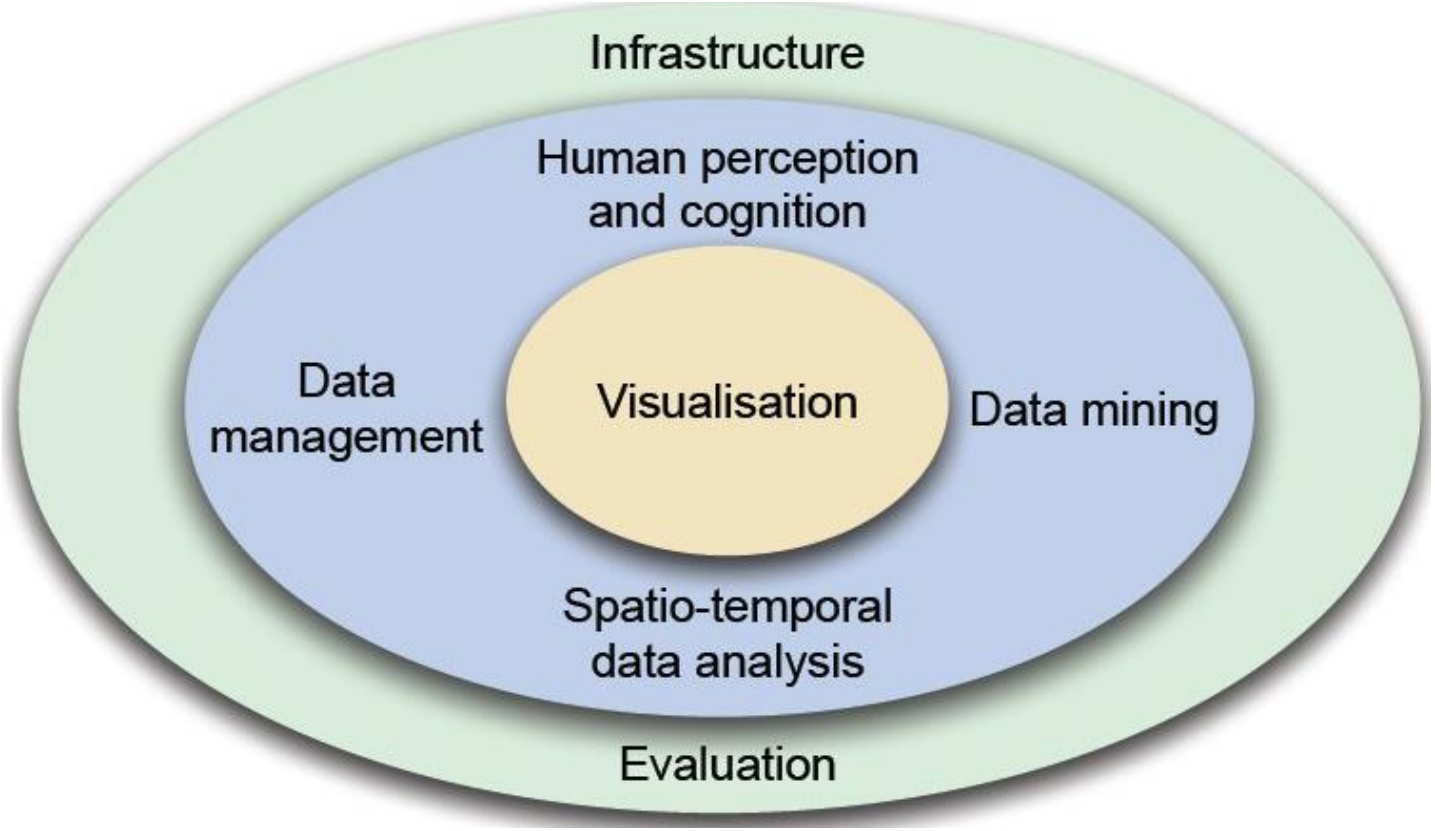
- Jak vhodně prezentovat data?
- Menší data
  - „overview first, zoom/filter, details on demand“
- VA
  - Není vhodné – obtížené vytvořit přehled, mohli bychom přijít o důležité informace
  - „Analyse first, show the important, zoom/filter and analyse further, details on demand“

# 14. VA - Proces – automatické metody

- DM metody
- Výstup je model
- Možnosti interakce s daty
- Přehlednější úprava parametrů metod
- Vizualizace modelu – jednoduší vyhodnocení výsledků

# 14. VA - Building Blocks of Visual Analytics Research

- Integruje několik vědních disciplín
- Slouží k zobrazení: Dat; Výsledků analýz
- Zpřehlednění procesů v ostatních oblastech



# 14. VA - Základní součásti

1. Vizualizace
2. 
    a) Správa dat
    b) Data mining
    c) Prostorová a časová analýza
    d) Aspekty vnímání a poznávání
3. 
    a) Infrastruktura
    b) Evaluace

# 14. VA - Základní součásti – 1. vizualizace

- „grafická reprezentace myšlenek a dat“
- **Analýza dat**, tři hlavní přístupy**:**
  - Prezentace, konfirmační/explorační analýza
- **Prezentace** – výběr vhodných technik uživatelem
- **Konfirmační analýza**
  - Deduktivní přístup
  - Vstup hypotéza o datech, kterou ověřujeme
  - Pomocí vizualizace potvrdíme/vyvrátíme
- **Explorační analýza**
  - Induktivní přístup
  - Není přímo daná hypotéza, ale hledáme potencionálně užitečné informace a vztahy v datech
  - Důležitá interaktivita a vizualizace

# 14. VA - Základní součásti – 1. vizualizace – vědecká data

- Hlavně 3D data
- Biologie, technika, meterologie, kosmologie, …
- S cílem reprezentovat data, často dočasná, jako fyzikální entity
  - Vizualizace toků, vykreslování objemů, …
- Senzor, simulace, laboratorní testy
- Lze jednoduše mapovat do 2D/3D prostředí

# 14. VA - Základní součásti – 1. vizualizace – informace

- Metody pro vizualizaci abstraktních dat
  - Bussiness data, demografická data, sociální sítě
- Velké objemy, často stovky dimenzí
- Různé datové typy (numerická, textová, multimediální)
- Data nelze snadno mapovat do 2D/3D prostředí
- Standardní grafové techniky jsou u velkých multidimenzionálních dat neefektivní
  - Velmi důležitá interaktivita vizualizace
  - Vyvinuty nové speciální typy vizualizace a techniky
  - Např.: metoda paralelních souřadnic, treemapy, …

# 14. VA - Základní součásti – 2a. Správa data

- Efektní a kvalitní správa dat je klíčovou částí VA
- Poskytuje data k analýze
- Integrace heterogenních dat
- Čištění a validace – chybějící/nepřesná data
- Vyžadována rychlá odezva (skoro real-time) a automatizace při přístupu k datům
- Nové formy zdrojů dat – streamovaná data, senzorové sítě, automatická extrakce informací z velkých kolekcí dokumentů, …
- Techniky správy dat stále více využívají inteligentní techniky analýzy dat a také vizualizaci pro optimalizaci procesů a informování uživatele

# 14. VA - Základní součásti – 2b. Data Mining

- Automatické metody pro extrakci informací
- Odhalení struktury bez předchozích znalostí
- Klasifikace, rozhodovací stromy, … (uč. s učitelem)
- Shlukování, asociační pravidla, … (uč. bez učitele)
- **Visual DM**
  - Interaktivní vizualizace
  - Rozhraní umožňující vizuální prezentaci zkoumaných dat
  - Prezentace výsledků analýz

# 14. VA - Základní součásti – 2c. Prostorová, časová analýza

- **Prostorová data**
  - Data, která mají základ v reálném světě
    - geografické měření, GPS
    - Hlavně ta, která se dají vynést do grafu nebo zobrazit na mapě
  - Hledání vztahů a zajímavých vzorů
  - Využití efektivnosti datových struktur
  - Podobnostní funkce
- **Časová data**
  - Hodnoty se mění v čase
  - Hledání vzorů, trendů, korelací v čase

# 14. VA-Základní součásti–2d. Aspekty vnímání a poznávání

- Reprezentuje lidskou stránku
- Vizuální vnímání – prostředek, kterým člověk interpretuje své okolí
- Poznávání – schopnost tyto informace pochopit a vyvodit závěry
- Poznatky z těchto oblastí jsou důležité při návrhu
  - Uživatelských rozhraní
  - Multimodálních interakčních technik
    - Interakce člověka s počítačem užitím více vstupních a výstupních zařízení

# 14. VA - Základní součásti – 3a. Infrastruktura

- Efektivní propojení všech procesů, funkcí, služeb
- Rozdílné technologie využívané v jednotlivých oblastech
- Velká interaktivita klade požadavky na kvalitu infrastruktury
- Většina VA systémů je vyvíjena na míru
- Často in-memory DB místo klasických DBMS
- **Aplikace**
  - Fyzika, astronomie (vizualizace toků, dynamika tekutin)
  - Bussiness data (finanční trhy)
  - Biologie, medicína;
  - Bezpečnost …

# 14. VA - Základní součásti – 3b. Evaluace

- Vyvíjí se velké množství nových technik a metod
- Je potřeba vyhodnotit efektivitu, přínos a vzájemnou kvalitu
- Dobré vyhodnocení může odhalit potenciální problémy
- Výzkum a vývoj je díky velkému množství specifických oblastí roztříštěn, což komplikuje použití jednotných evaluačních metod

# 14. VA: Shrnutí

- VA se tedy zabývá čtyřmi oblastmi
- Data
  - velké množství různorodých typů dat s různou kvalitou
- Uživatelé
  - vyhovět uživatelským požadavkům a zjednodušit a zpřehlednit analýzu
- Design
  - kvalitní návrh systému
- Technologie
  - využití moderních a efektivních technologií

# 14. VA: Shrnutí – Data

- Velké množství dat
- Ukládání, získávání a přenos
  - Distribuované databáze, cloudy
- Náročnost zpracování
- In-memory úložiště
  - Lépe vyhovuje požadavkům
- Různorodá data
  - Nekvalitní, chybějící, nekompletní a chybové

# 14. VA: Shrnutí – Data

- Složitost integrace dat z více zdrojů
- Potřeba transformovat data do jednotného formátu
- Nové typy dat
- Nové zdroje dat: streamovaná data
  - Velké dávky nebo neustále
- Analýza finančních toků
- Potřeba zpracovat data v reálném čase

# 14. VA: Shrnutí – Uživatelé

- Uživatel by měl mít přehled o průběhu
- Odkud se data berou
- Jaké operace s daty byly provedeny
  - Čištění, analýza, vizualizace
- Chápání nedostatků v datech a výsledcích
- Omezení chybné interpretace výsledků
- Většina DM metod je neintuitivních a vyžadují odbornou znalost
  - Vhodná úroveň abstrakce, reprezentace dat na obrazovce

# 14. VA: Shrnutí – Design

- Aplikace moderních teoretických a praktických znalostí
- Hodně technologií a pro daný problém je potřeba zvolit správné techniky
  - Analytické metody, typ vizualizace
- Unifikovaný model
  - Rychlejší a spolehlivější návrh a implementace

# 14. VA: Shrnutí – Technologie

- Je potřeba ukládat mezivýsledky
- Analytik má neustále přehled o průběhu analýzy a řídí její průběh
- Analytik může požadovat data za jeden den, stejně jako za celý rok

# 15. CI (Competitive Intelligence), charakteristické vlastnosti, geneze a souvislosti CI, interní versus externí zdroje dat, Key Intelligence Topics (KIT).

# 15. Competitive Intelligence

- Zjišťování, sledování a vyhodnocování konkurenčního prostředí (firmy, organizace) s cílem odhalit slabé a silné stránky konkurence, rozpoznat její strategické záměry.
- Zahrnuje analýzu a syntézu dat, resp. informací, které se transformují do strategických znalostí, shromažďování informací o konkurenci a sledování subjektů firemního okolí (trh, stát, právo a legislativa, politické a demografické souvislosti).

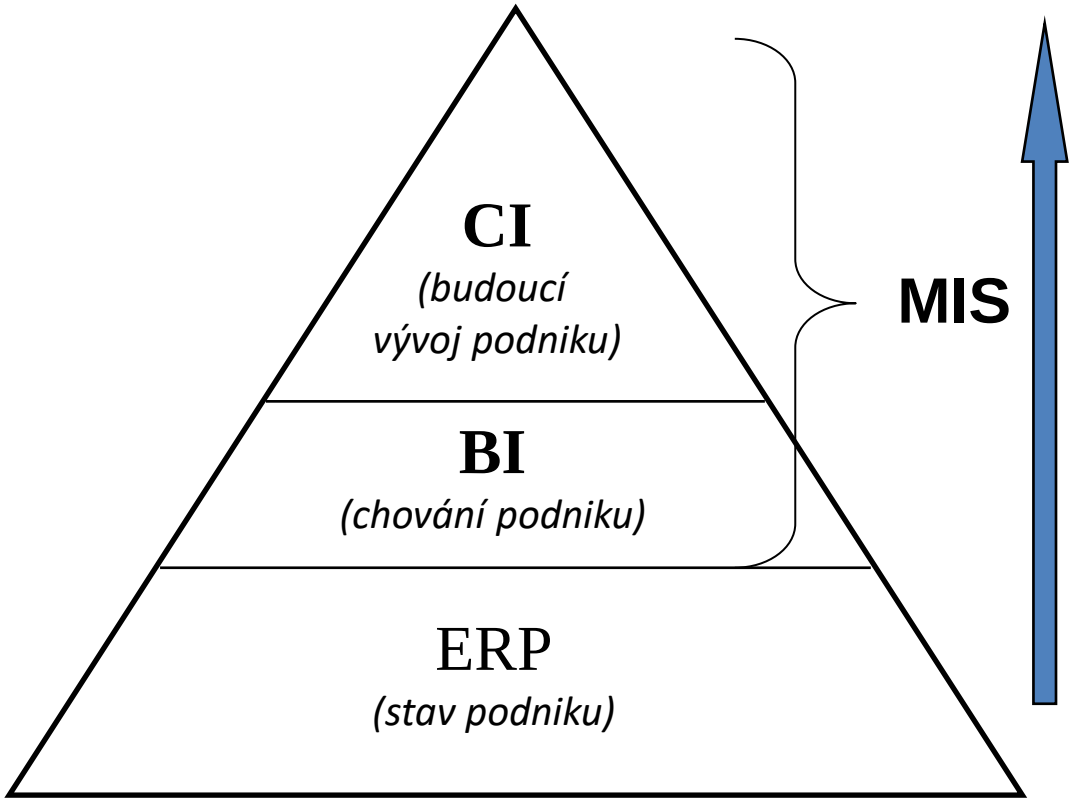
# 15. CI- Charakteristické vlastnosti

- Postaveno na procesu řízení, sběru, analýzy, distribuce informací a znalostí, které podporují rozhodování, zejména na strategické úrovni
- Tyto činnosti jsou prováděny za účelem zvýšení konkurence schopného postavení, snížení rizika hrozeb z okolí a zmapování možných příležitostí
- Zaměření CI je hlavně na externí prostředí, ve kterém se organizace nachází
- Etická a legální činnost
- CI je o naslouchání co se kde děje, o vytváření znalostí a propojování těch co znají s těmi co rozhodují

# 15. Geneze a souvislosti CI

- Zpravodajské služby, policie, vojsko
- Strategický mng., marketing, psychologie
- Informační věda a knihovnictví
- Znalostní mng.

- Systematický vývoj CI můžeme sledovat od 80. let 20. století, kdy si Spojené státy americké uvědomily sílu japonského trhu a zjistily, že mu nejsou schopny konkurovat.
- Uvědomění si konkurence cizích trhů vedlo nejprve ke vzniku oboru CI, pak k následnému zřízení samostatných oddělení CI ve velkých firmách.




*Probíhá **selekce a agregace** interních informací*
***Roste***
*- neurčitost*
*- význam externích informací.*
*- závažnost rozhodnutí*



215

# 15. CI - Interní zdroje podniku

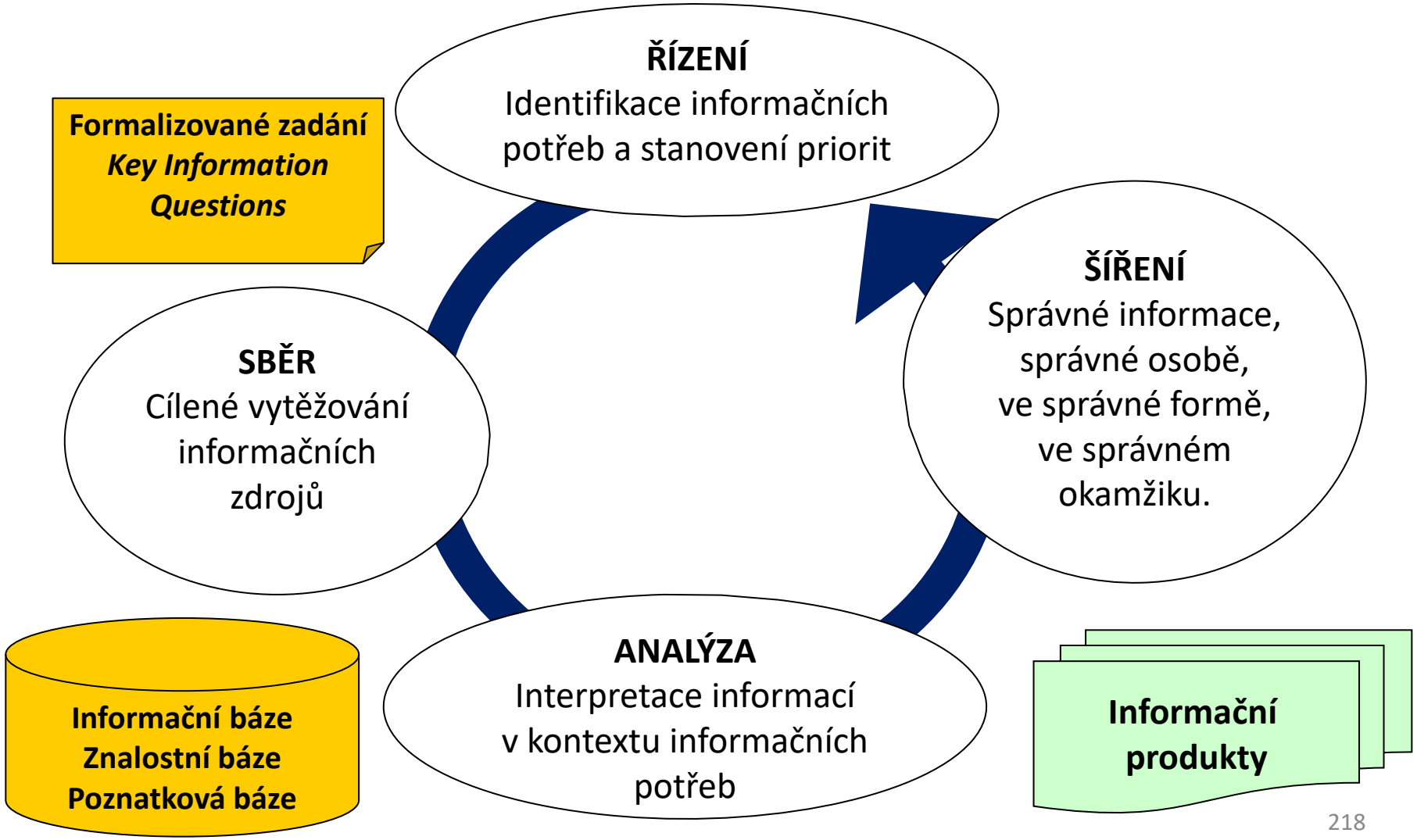
- Strukturovaná
  - Podnikové databáze, IS, CRM systémy, …
- Nestrukturovaná
  - Textové soubory, e-maily, zápisy, …
- Výhody
  - Podnik má přehled o svých informačních zdrojích
- Nevýhody
  - Soustředí se pouze na to co se děje uvnitř podniku
  - Nebo na jeho bezprostřední okolí

# 15. CI – Externí zdroje podniku

- Strukturovaná
  - Odborné databáze, specializovaní dodavatelé, katalogy, ceníky, …
- Nestrukturovaná
  - Web, RSS kanály, emaily, sociální sítě, …
- Výhody
  - Poskytují další stupeň informační podpory pro mng-é. rozhodování
    - Komplexnější vnímaní podniku v konkurenčním prostředí
- Nevýhody
  - Objem dat
  - Nutné specializované nástroje, hlavně u nestruktur. dat

# 15. Zpravodajský cyklus CI

- Cyklus spouští informační potřeba



# 15. Zpravodajský cyklus CI

1. Řízení
   - Identifikace informačních potřeb a stanovený priorit
   - Čím se zabývat, proč to dělat, co se získanými znalostmi
2. Sběr
   - Cílené vytěžování informačních zdrojů
   - Třídění, porovnávání, ověřovaní spolehlivosti
3. Analýza
   - Interpretace informací v kontextu informačních potřeb
4. Distribuce
   - Komplexní analýza je převedena na syntézy v podobě správně interpretovaných výstupů, které slouží mng. v rozhodování
   - Efektivní distribuce spočívá ve třech atributech
     - Obsahu – výsledek analýzy nebo nové relevantní informace
     - Formě – srozumitelnost pro konkrétního uživatele
     - Aktuálnosti

# 15. Key Intelligence Topics (KIT)

- V první fázi zpravodajského cyklu musí být definovány konkrétní oblasti zájmu a účel pro jaký se ZC bude provádět
- **Rozhodovací témata**
  - Váží se k nějakému plánovanému rozhodnutí
  - Jasně definován obsah i termín
- **Předmětová témata**
  - Týkají se určitých subjektů (konkurenti, partneři, stát, banky, …)
  - Slouží k předvídání chování těchto subjektů a dopadu na organizaci

# 15. Key Intelligence Topics (KIT)

- **Varovná témata**
  - Součástí systému včasného varování
  - Pomocí těchto témat se obyčejně průběžně monitorují zadané indikátory
    - Umožnění rozpoznání budoucích hrozeb
    - Zachycení možných příležitostí

# Příklad 1 1. Metriky: Co budeme měřit

**VÝSLEDEK (1)**

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

***Počet zapsaných pokusů:*** *Dle definice může mít různý význam:*

*a) Počet studentů zapsaných na předmět*

*b) Počet studentů zapsaných na předmět v jednom semestru*

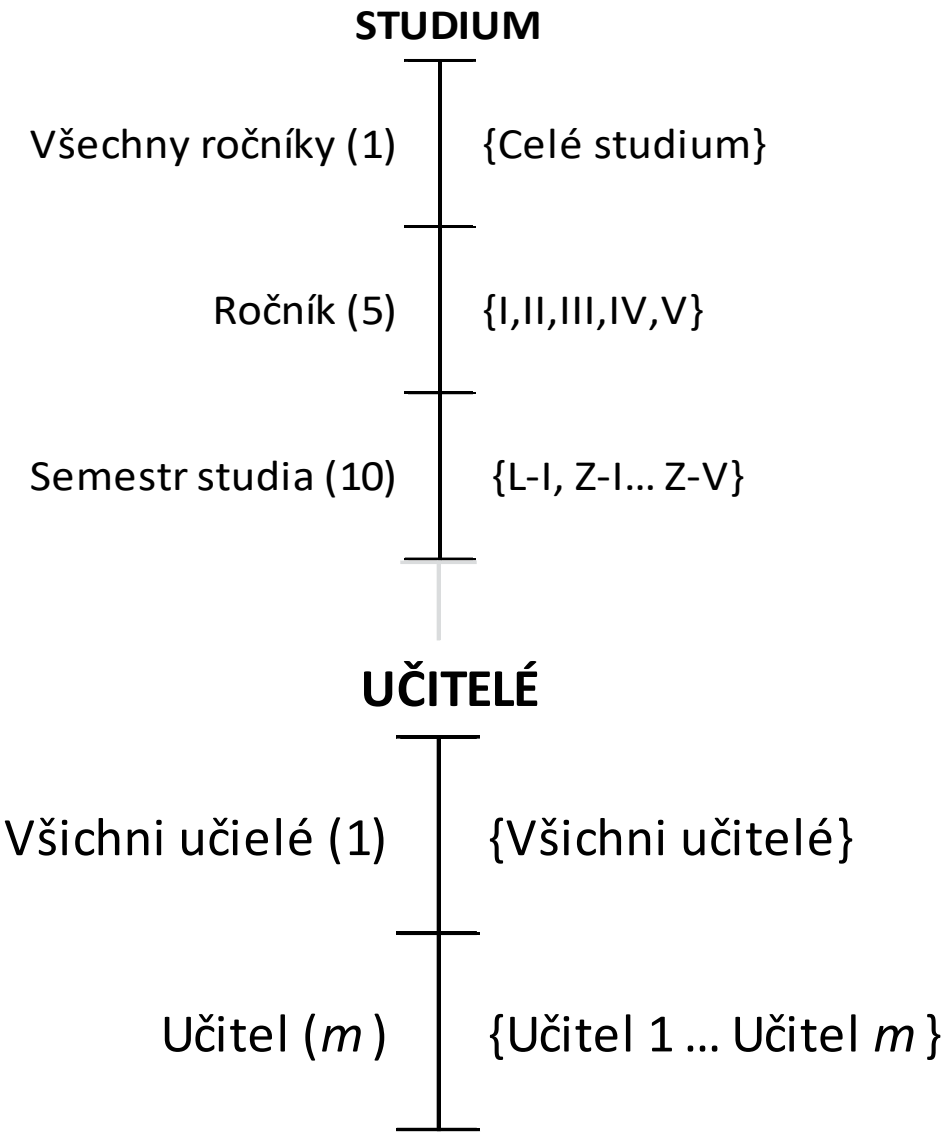
- *V tomto případ by to mohla být pně aditivní metrika*
- *Záleží ovšem ještě na hierarchiích (viz dále)*

***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?*

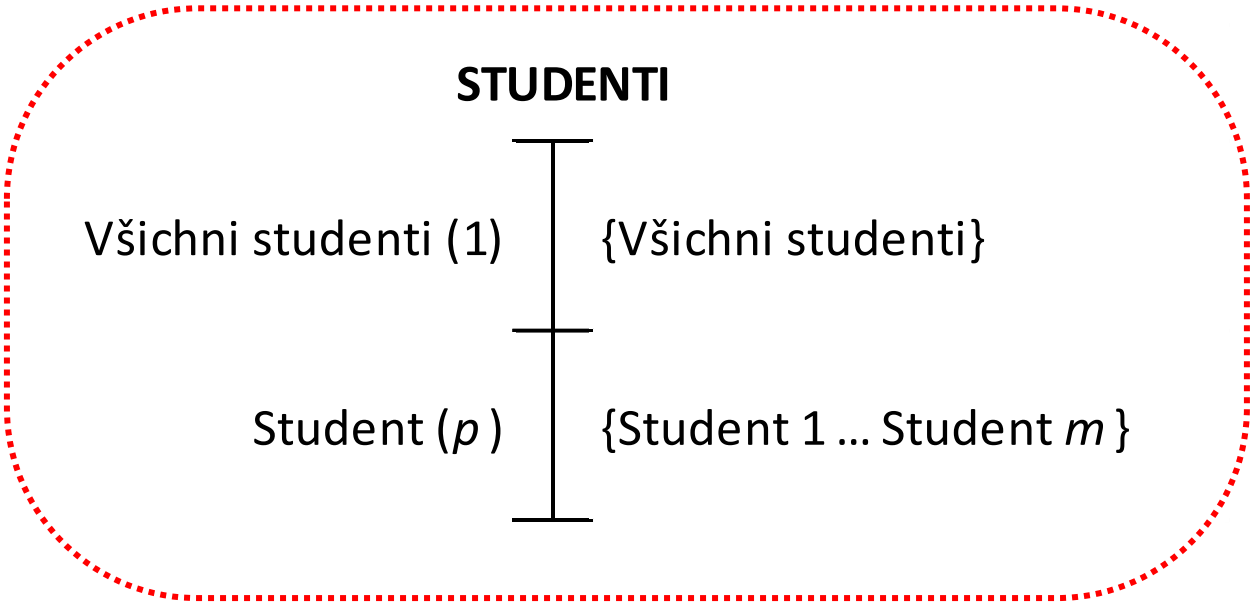
# Příklad 1 2. Dimense – podle čeho budeme sledovat

**VÝSLEDEK (2)**



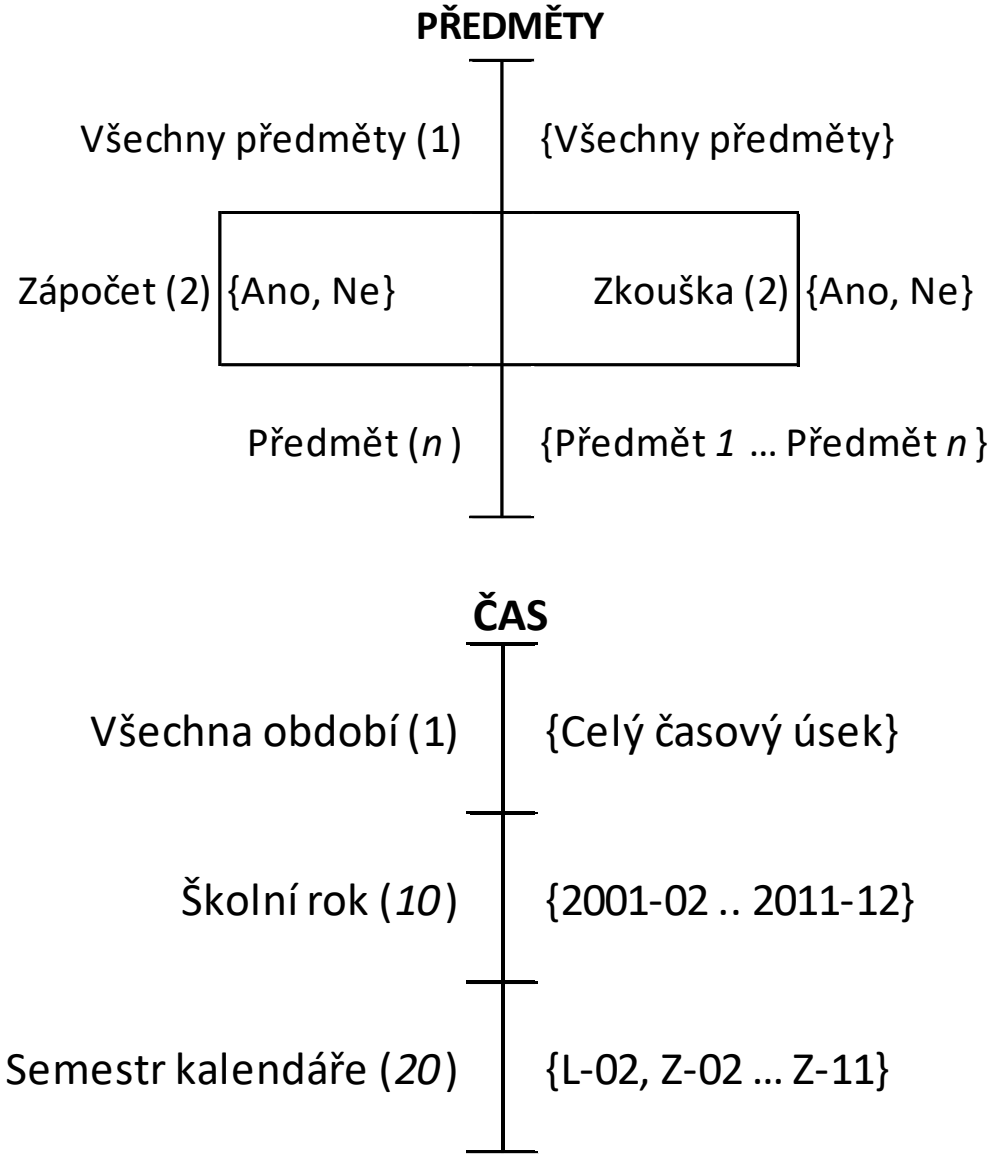
# Příklad 1 2. Dimense – podle čeho budeme sledovat

**VÝSLEDEK (2)**



# Příklad 1 2. Dimense – podle čeho budeme sledovat

**VÝSLEDEK (2)**

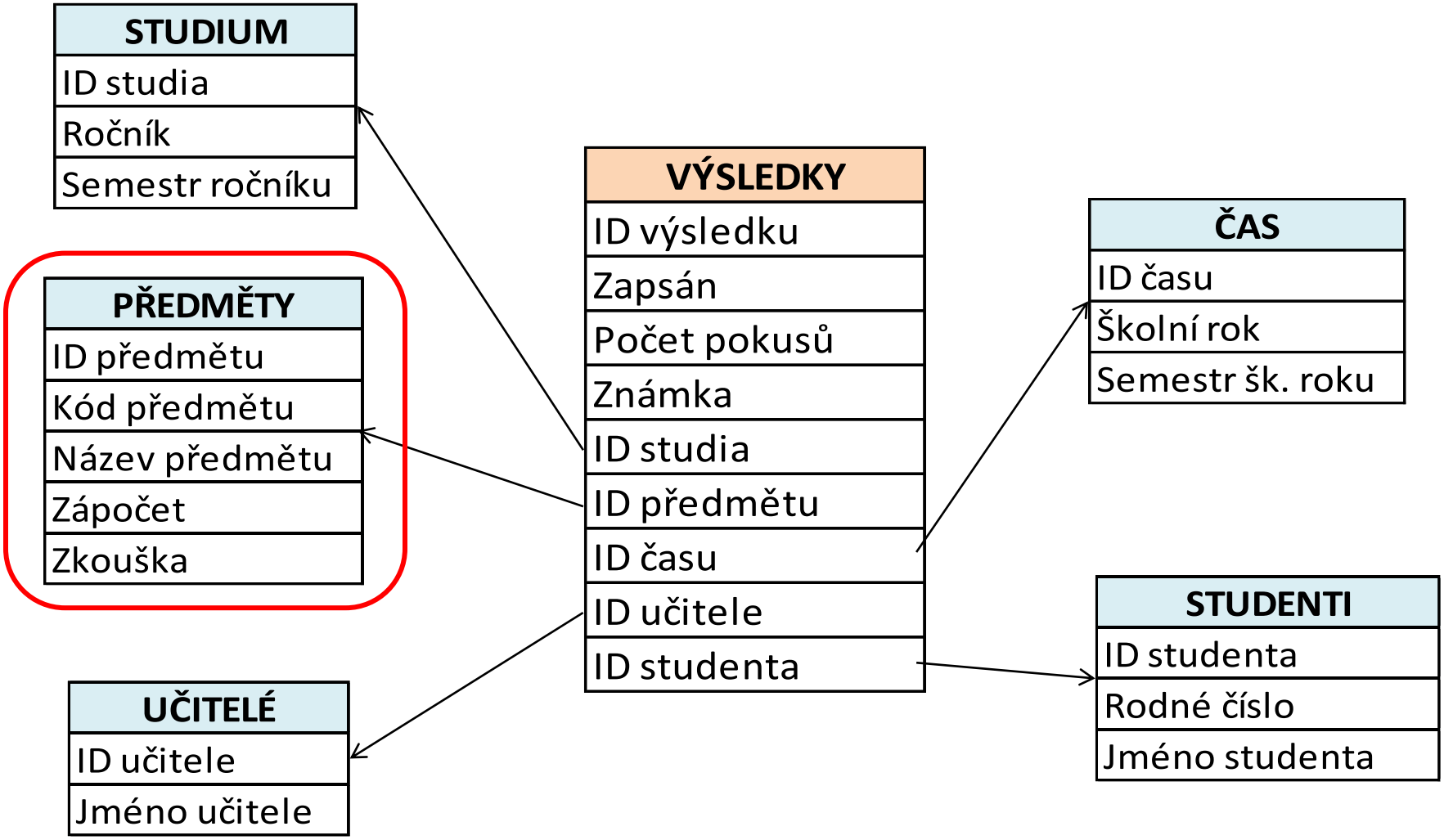


*Pozn.: Dimense STUDENTI přidána, aby bylo možno sledovat průměrné známky*

- *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*

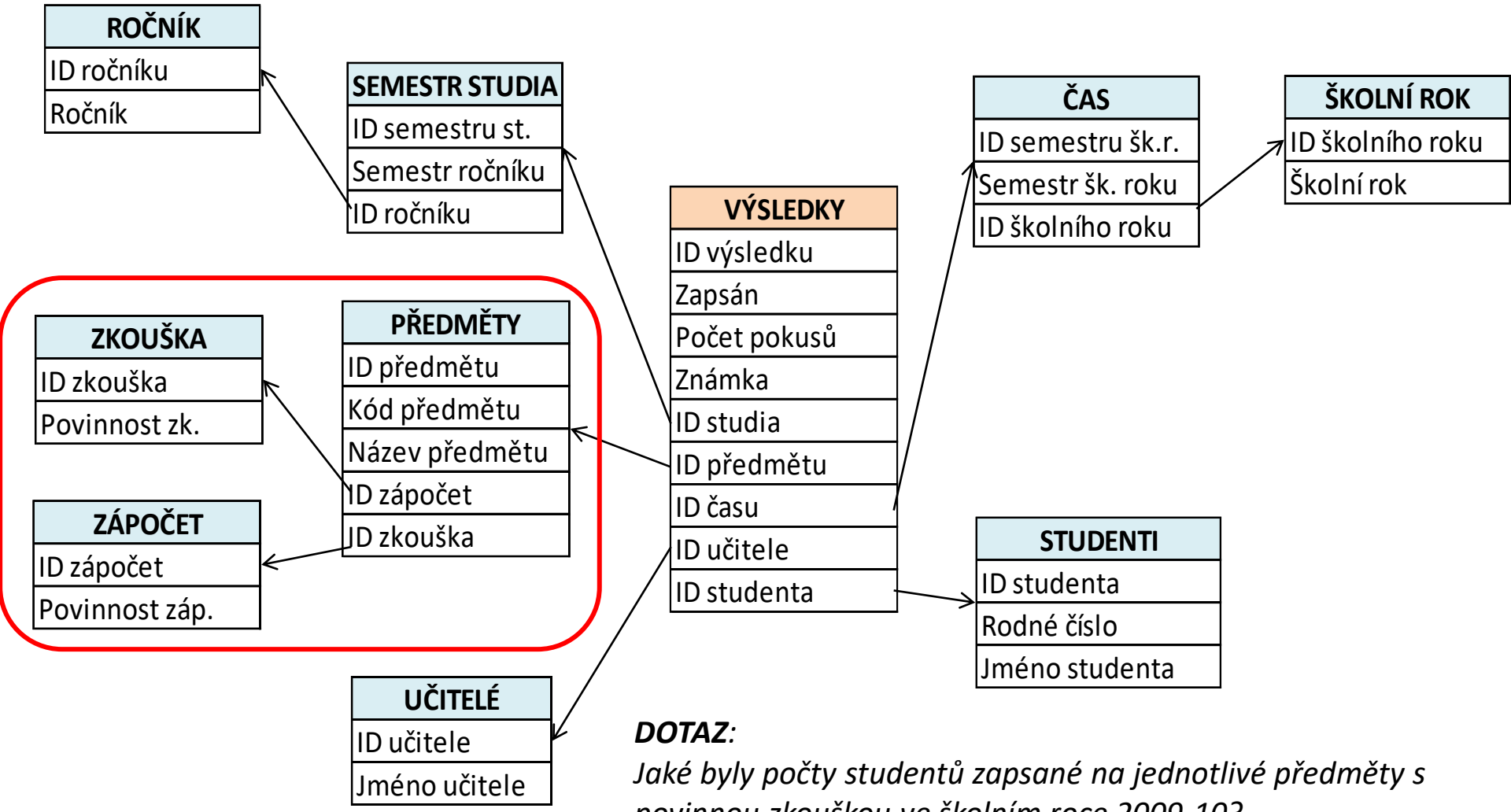
# Příklad subdimense – 1

**HVĚZDA MAXIMÁLNÍ DENORMALISACE (BEZ SUBDIMENSE)**



# Příklad

**VLOČKA – MAXIMÁLNÍ NORMALISACE**




***DOTAZ:***

*Jaké byly počty studentů zapsané na jednotlivé předměty s povinnou zkouškou ve školním roce 2009-10?*

# Příklad

**SUBDIMENSE**