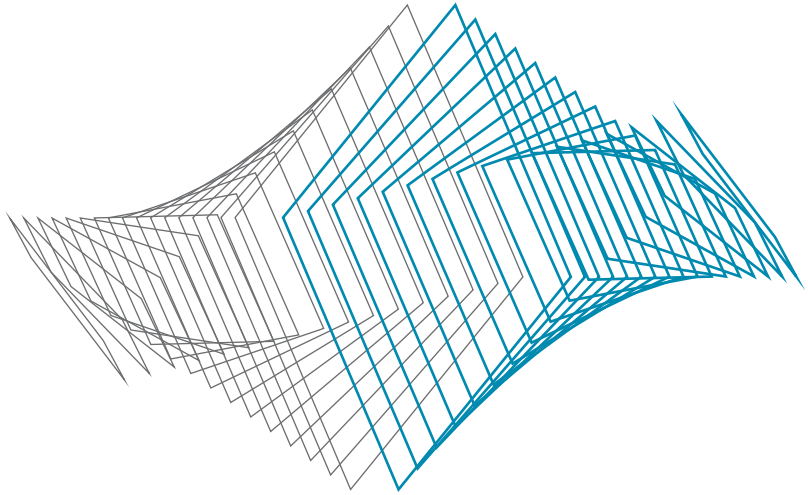
2011

PODNIKOVÉ INFORMAČNÍ SYSTÉMY

VŠEM

Business Intelligence

Jan Žižka



VYSOKÁ
ŠKOLA
EKONOMIE
A MANAGEMENTU

Business Intelligence (BI)

(Podnikové zpravodajství)

Úvod do základů dvou hlavních směrů:

1. Ukládání dat do ***datových skladů***, které následně slouží jako zdroj dat pro analytické aplikace.

2. Analýza obsahu datových skladů, a to zejména nejmodernějším způsobem prostřednictvím ***dolování znalosti*** z dat.

Zaměření business intelligence

BI se svým principem a vývojem zaměřuje především na to, jak moderní společnosti, instituce, organizace a podniky ***eticky a legálně*** **shromažďují** co největší množství **údajů (dat)** o svých zákaznících, akcionářích, konkurentech, podnikatelském prostředí, apod.

BI ***získává data*** z nejrůznějších disponibilních zdrojů.

BI z takto nahromaděných dat „***doluje***“ libovolnou užitečnou ***znalost*** pomocí nejpokročilejších analytických metod.

Definice business intelligence

Business Intelligence (BI) je zastřešující odborný termín pro určitou kombinaci vhodných softwarových i hardwarových nástrojů, databází, analytických prostředků a metod, aplikací a metodologií za účelem získání znalosti skryté v údajích popisujících reálný svět okolo nás, a to v oblastech spojených s podnikáním, ekonomikou a jimi ovlivňovanými institucemi libovolného charakteru.

Proces business intelligence vychází z transformace vstupních dat na informaci, která je dále přeměněna na znalost použitelnou pro rozhodování zakončené vybraným typem akce.

Data, informace, znalost, metaznalost

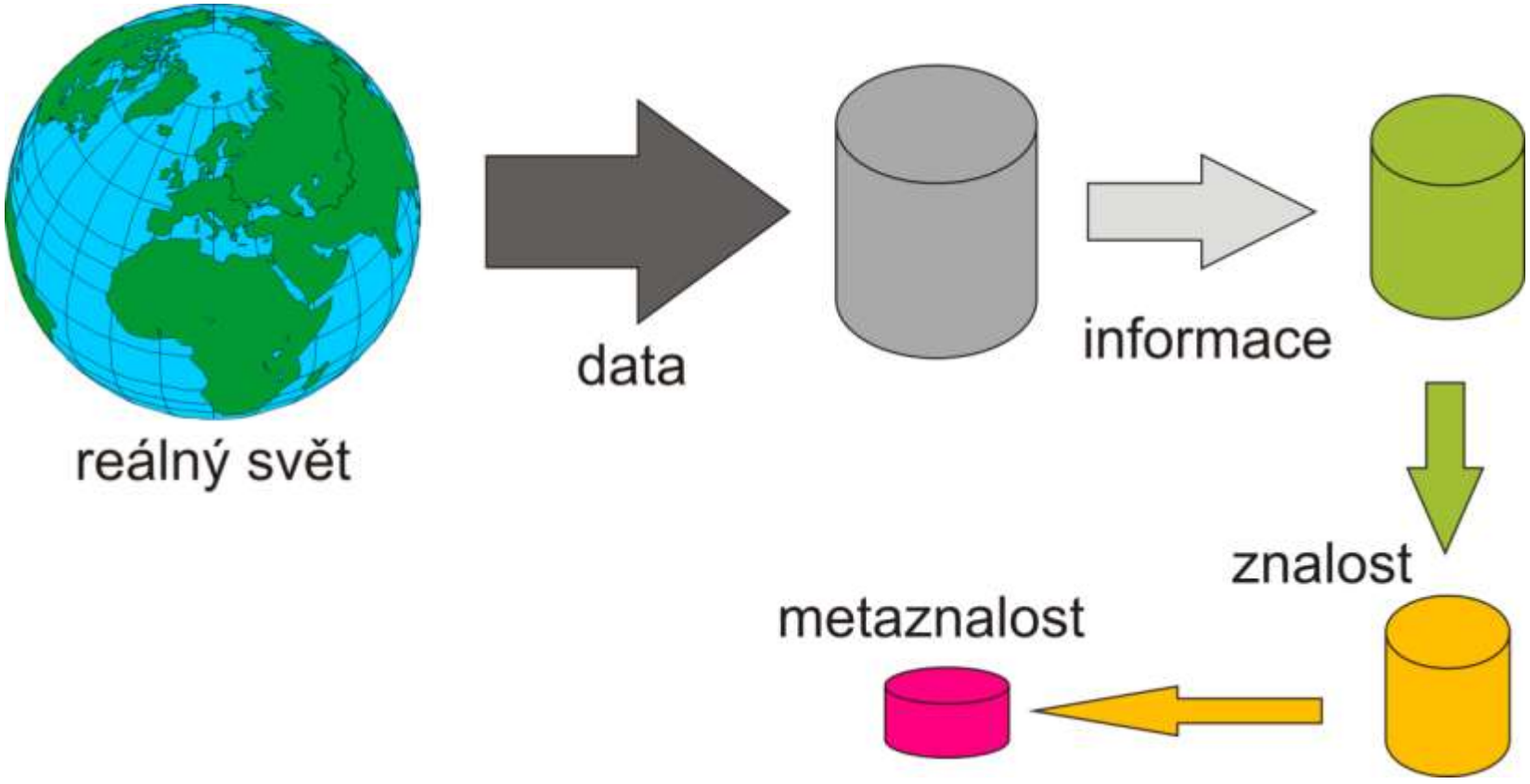
Data jsou údaje získávané libovolným pozorováním a měřením okolního reálného světa. Hodnoty údajů mohou být číselné (numerické) i nečíselné (nenumerické, tj. nominální, vyjmenované).

Informace je ta část dat, která je relevantní k řešení nějakého stanoveného konkrétního problému.

Znalost je zobecněná informace.

Metaznalost („nadznalost“) je znalost o znalosti.

Data, informace, znalost, metaznalost



Data a jejich typy

Data mohou být jednoduchá (např. počet kusů) nebo složitá (např. barevné video se zvukem), a různého typu:

numerická:

- binární
- celočíselná
- reálná numerická

nominální:

- slovní
- symbolická
- aj.

Znalost a metaznalost

Znalost ***induktivní:*** vzniká zobecněním údajů o (mnoha) konkrétních případech. Induktivně získaná znalost může pak být použita jako znalost deduktivní, pokud je dostatečně kvalitní (experimentální a empirické prokazování kvality).

Znalost ***deduktivní:*** obecná znalost, kterou lze použít na individuální konkrétní případy. Deduktivně získaná znalost je např. odvozena (vzorec) a matematicky dokázána.

Metaznalost umožňuje vybrat pro řešení konkrétní úlohy vhodnou znalost (obdoba výběru odborníka).

Datové sklady: DWH, Data Warehouse

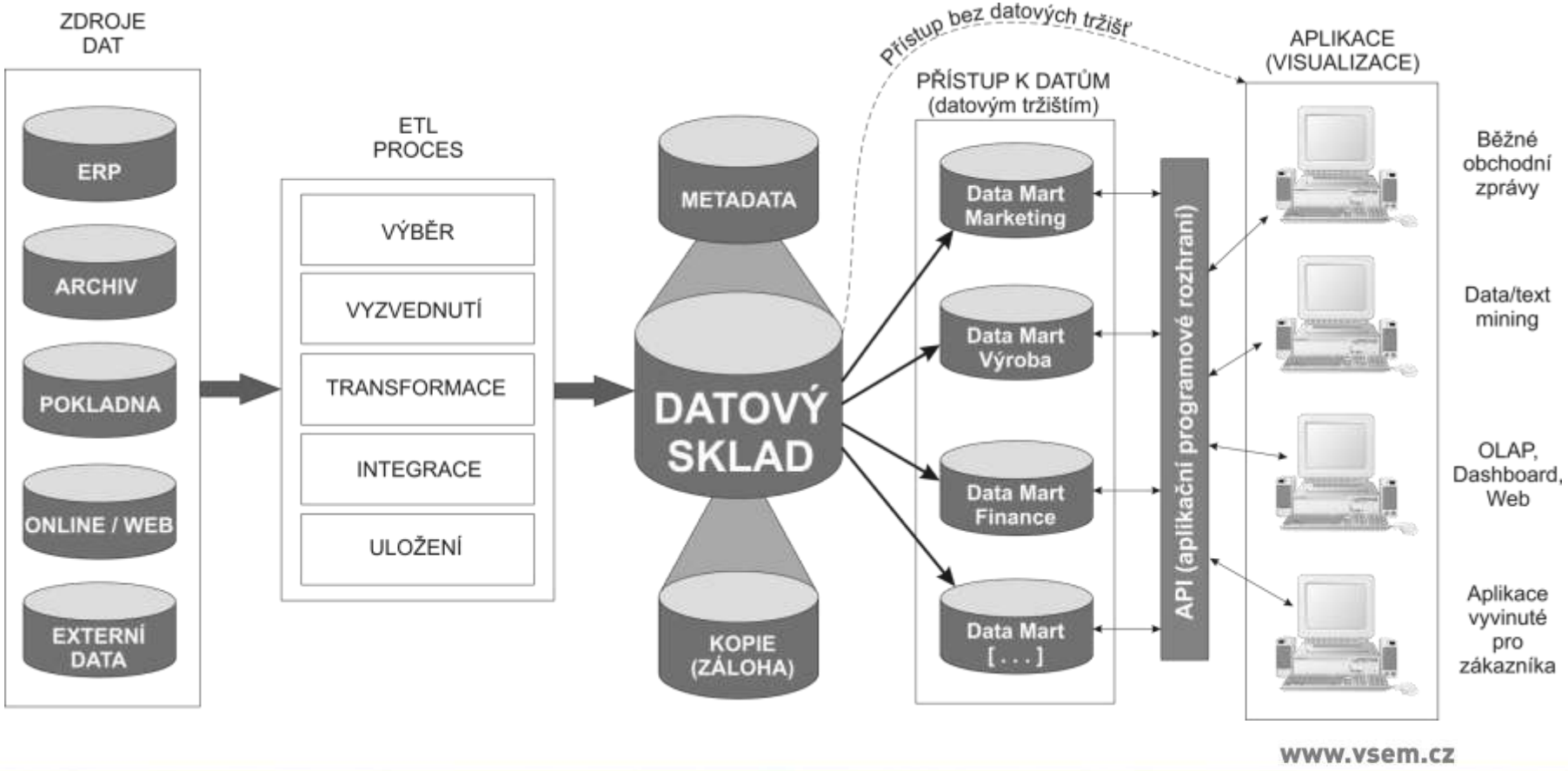
Datová tržiště: DMa, Data Marts, oddělené, navzájem nezávislé menší datové sklady pro jednotlivé pobočky, oddělení, závody, apod.

Centrální podnikový datový sklad: EDW, Enterprise Data Warehouse, jeden obvykle velmi rozsáhlý (co do objemu dat) integrovaný datový sklad pro celý podnik či organizaci.

Primárním účelem datových skladů je sloužit jako **zdroj dat a informace pro analýzy**.

Datové sklady **uchovávají historii**.

Prostředí datových skladů



Architektura datových skladů

Základem je ***vícevrstvá architektura***.

V praxi převládají architektury ***dvouvrstvé*** (klient-pracovní stanice; aplikační a databázový server) a ***třívrstvé*** (klient-pracovní stanice; aplikační server; databázový server).

Někdy postačuje i ***jednovrstvá*** architektura (pracovní stanice s daty uloženými na interním disku).

Typ architektury je dán objemem dat, aplikačním zaměřením, nebo možnostmi a potřebami datového skladování.

Tři hlavní části datového skladu

1. Vlastní ***datový sklad***, který obsahuje data a potřebný provozní software.

2. Software ***v pozadí*** (back-end) pro získávání dat převzatých z minulosti („zděděných“) nebo z externích zdrojů. Úkolem tohoto software je data konsolidovat, sumarizovat, a nakonec je uložit do datového skladu v potřebném formátu a požadované kvalitě.

3. Software ***v popředí*** (front-end) umožňující uživatelům přístup do datového skladu a analýzu disponibilních dat.

Třívrstvá architektura

Klientská pracovní stanice: software v popředí.

Aplikační server: software v pozadí.

Databázový server: vlastní datový sklad.

Data z datového skladu jsou uložena tak, aby tvořila multi-dimenzionální databázi pro ***multidimenzionální analýzu*** a presentaci, resp. umožnila vytvoření speciálně zaměřených kopií z částí datového originálu uloženého v datovém skladu, a následně tyto kopie přenést do datových tržišť.

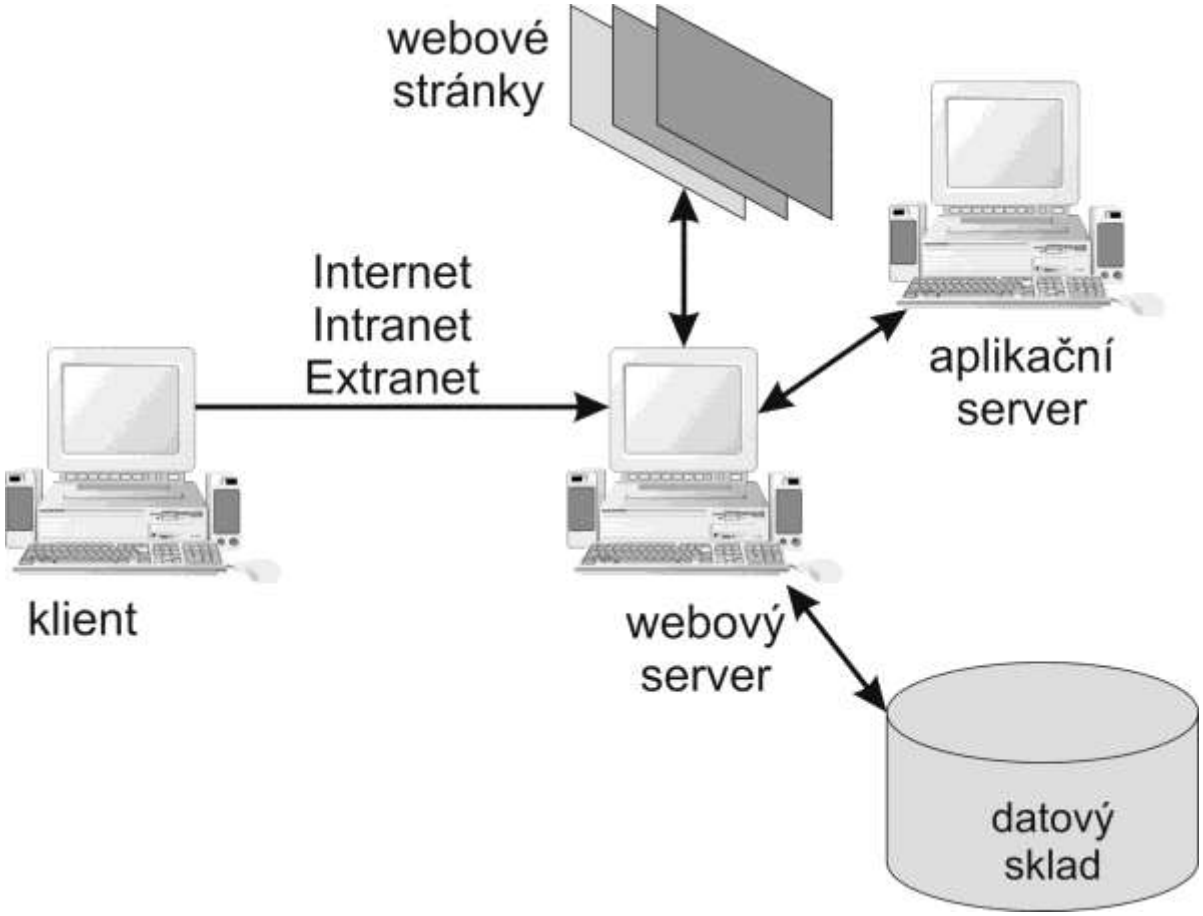
Dvouvrstvá architektura

Aplikační server je fyzicky na tomtéž hardware jako samotný datový sklad.

Výhoda je v ekonomicky nižších nákladech na vybudování vlastního datového skladu a jeho uživatelského prostředí.

Na druhé straně však mohou vznikat značné výkonnostní potíže tehdy, když je rozsáhlý datový sklad používán aplikacemi kladoucími vysoké nároky na intenzívní zpracování dat pro podporu rozhodování nebo hluboce zaměřené analýzy z oblasti business intelligence.

Webová architektura (je v principu třívrstvá)



Uživatelské výhody webového skladu

- Snadné používání.
- Pohodlný přístup k datům.
- Nezávislost na konkrétním operačním systému.
- Nízká cena realizace.
- Externí i interní uživatelé datového skladu vidí data shodně.

Požadavky na vytvoření datového skladu

- Jaký systém řízení báze dat (SŘBD) použít?
- Bude využito paralelní a/nebo distributivní zpracování?
- Budou využity migrační nástroje pro naplnění datového skladu?
- Jaké nástroje se budou používat pro podporu výběru a analýzy dat?

Deset faktorů návrhu datového skladu

- Vzájemná informační nezávislost mezi jednotkami podniku
- Informační potřeby vrcholového řízení podniku
- Naléhavost a potřeba vybudování datového skladu
- Charakteristika úloh koncového uživatele
- Zdroje a omezení
- Strategický výhled na datový sklad před jeho vytvořením
- Kompatibilita s již existujícími systémy
- Pochopení a přijetí potřeby ze strany vlastního IT personálu
- Technické problémy
- Sociální a politická hlediska

Integrace dat

Umožňuje přístup k datům řadě softwarových nástrojů z oblasti ETL a následných analýz:

- přístup k datům – získat k dispozici v potřebný čas potřebná data z libovolného zdroje;
- federalizace dat – sjednocení podnikatelských pohledů napříč mnoha různými datovými úložišti;
- zachycení změn – identifikace, zachycení a předání změn vzniklých v podnikových datových zdrojích.

Integrační technologie pro integraci dat

- EAI (Enterprise Application Integration) – integrace podnikových softwarových aplikací.
- SOA (Service-Oriented Architecture) – servisně zaměřená architektura.
- EII (Enterprise Information Integration) – integrace podnikových informací.
- ETL (Extraction, Transformation, and Load) – vyzvednutí dat ze zdroje, jejich transformace a uložení do cíle.

EAI

Poskytuje prostředek pro přesun dat ze zdrojových systémů do datového skladu.

Obsahuje funkce pro integraci, a to zejména aby umožnil sdílení své funkcionality (nikoliv dat) napříč systémy, což vede pozitivně k flexibilitě a opakované použitelnosti.

EAI usnadňuje získávání dat přímo do datového skladu, a to téměř v reálném čase, nebo předávání rozhodnutí do systémů OLTP (On-Line Transaction Processing, online zpracování transakcí).

EII

Zdroj nástrojů pro integraci informací v reálném čase, a to z množství různorodých zdrojů jako jsou relační databáze, webové služby a multidimenzionální databáze.

EII slouží pro přesun dat ze zdrojů tak, aby bylo vyhověno požadavkům na informace.

Nástroje EII využívají předdefinovaná metadata k získání pohledů na integrovaná data tak, že se koncovým uživatelům jeví jako relace.

ETL

ETL (Extract-Transform-Load) představuje moderní a netradiční přístup k ukládání dat do datových skladů a tvoří vlastně jádro datového skladování.

ETL je v současnosti prakticky nepostradatelnou integrální složkou jakéhokoliv projektu zaměřeného na data.

ETL procesy zaberou až 70 % času věnovaného na datově orientovaný projekt.

Tři hlavní funkce ETL

První funkce ***Extract*** je zaměřena na tzv. extrakci, tj. čtení dat z jedno-ho či více zdrojů).

Druhá funkce ***Transform*** je transformace, což představuje provedení potřebných změn, aby vyzvednutá data mohla být uložena do cíle, kterým je datový sklad nebo jiná databáze.

Třetí funkce ***Load*** má za úkol uložení vybraných zkonvertova-ných dat do datového skladu.

Hlavní požadavky na výběr ETL nástroje

- Schopnost číst/zapisovat z/do neomezeného počtu typů architektur datových zdrojů (univerzálnost nástroje).
- Automatizované převzetí a předání metadat současně s vlastními daty.
- Znalost historie přizpůsobování se otevřeným standardům (adaptabilita nástroje vzhledem k možnému budoucímu vývoji sdílených otevřených standardů).
- Snadno použitelné komunikační rozhraní pro vývojáře a provozního uživatele.

Vytvoření datového skladu: *Přímé výhody*

- Koncoví uživatelé mohou provádět analýzy různě.
- Je k dispozici sloučený pohled na společná data v rámci celého podniku, tzv. „jediná pravda“.
- Je k dispozici kvalitnější a aktuálnější informace. Datový sklad odlehčuje celkovou zátěž systému.
- Výsledkem je zvýšení výkonnosti celého systému
- Přístup k potřebným datům je značně zjednodušen.

Vytvoření datového skladu: *Nepřímé výhody*

- Zlepšení obchodních a podnikových znalostí.
- Výhody v konkurenčním prostředí oproti ostatním, kteří datový sklad používají buď méně dokonale nebo vůbec.
- Zlepšení služeb a spokojenosti zákazníků.
- Usnadnění a podpora složitého procesu rozhodování.
- Podpora reformování podnikatelského procesu pro dosažení lepších výsledků.

Vytvoření datového skladu: *Požadavky*

- Jasně definovat účel podnikání a činností.
- Získat podporu od koncových uživatelů z oblasti vedení podniku.
- Uvážlivě stanovit racionální časové rámce a rozpočet projektu.
- Respektovat to, co od projektu oprávněně očekává vedení podniku.

Kimballův model: *Jednotlivá datová tržiště*

- Metoda „zdola-nahoru“.
- Orientace na subjekty, jednotlivá oddělení či pobočky.
- „Zmenšená“ verze EDW zaměřená na individuální části.
- Používá dimenzionální modelování počínaje definicí a vytvořením tabulek.
- Realizuje postupně jedno tržiště za druhým podle konkrétních potřeb.

Inmonův model: *Centrální datový sklad*

- Metoda „shora-dolů“, využívající zavedené postupy a technologie pro vytváření databází (např. entity-relation-ship diagramy ERD, relační databázové systémy, apod.).
- Aplikace metody vývoje datového skladu ve spirále.
- Nevyžaduje předem existenci nebo vytváření datových tržišť.
- Poskytuje konsistentní a komplexní pohled na celý podnik.

Je lepší Kimballův nebo Inmonův model?

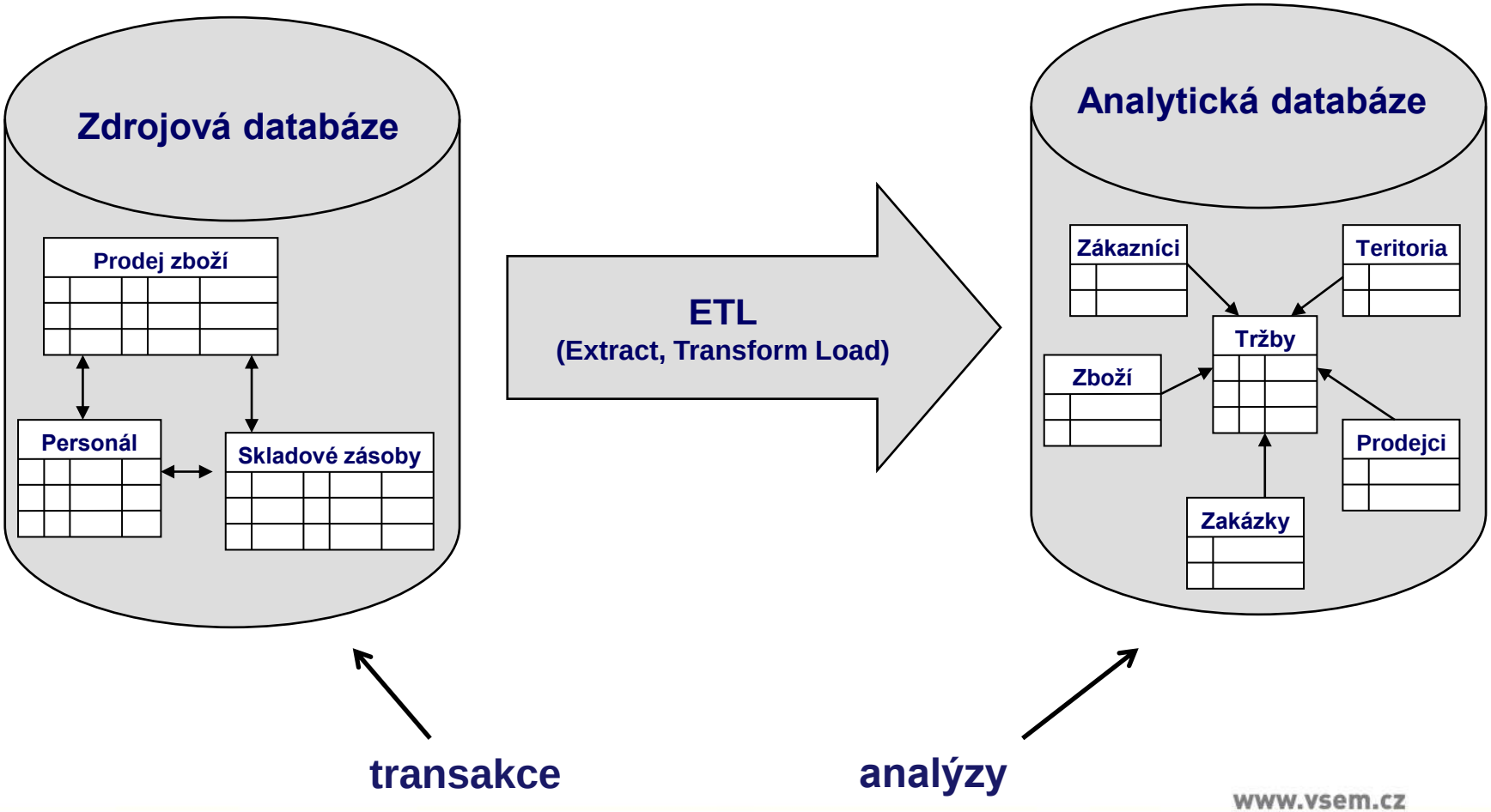
Každý z nich je velmi dobrý.

Každý má ale zároveň nedostatky.

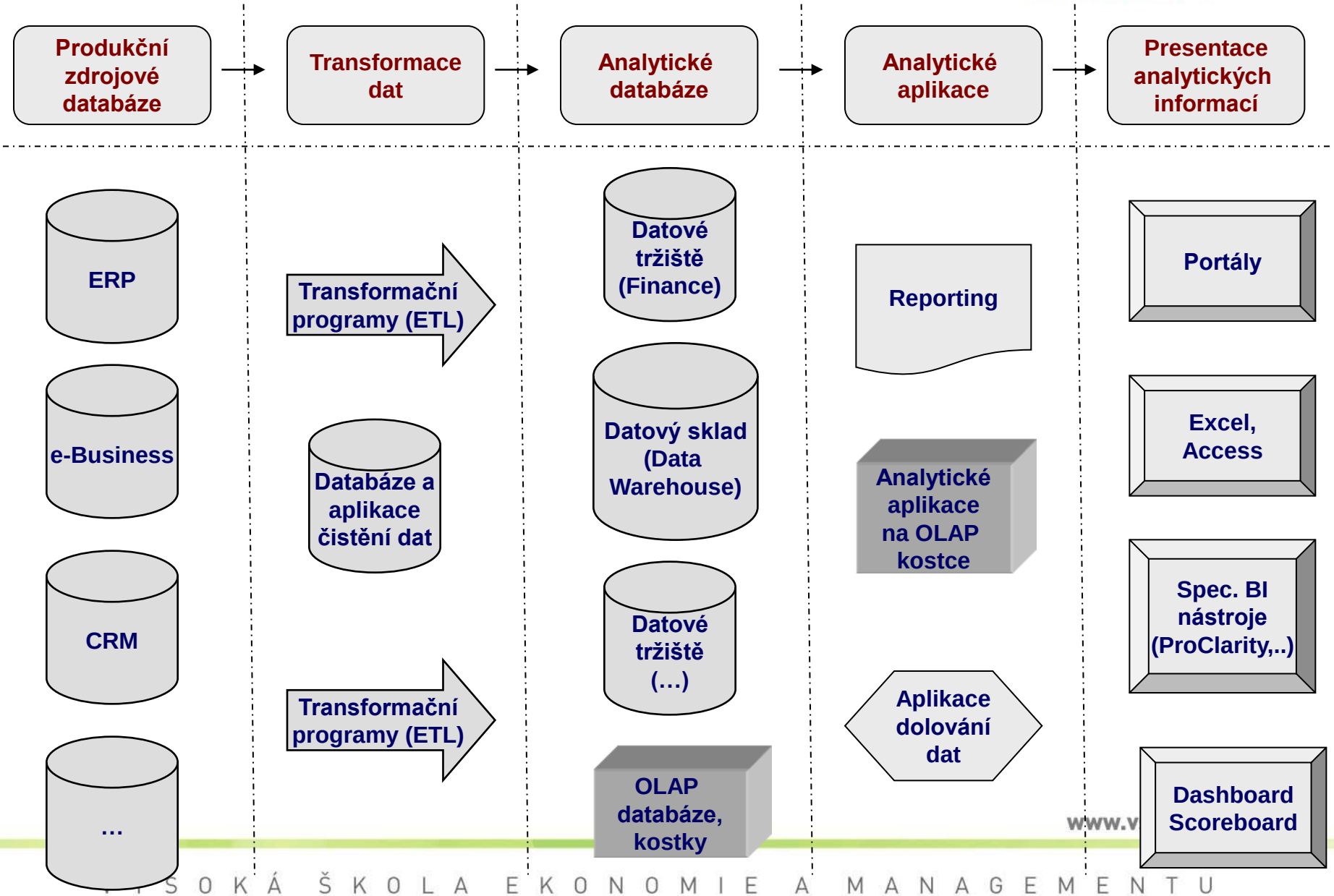
Zvolí se ten model, který vyhovuje konkrétním potřebám a zároveň finančním a časovým možnostem realizace.

Decentralizovaná (Kimball) i centralizovaná (Inmon) data v datových tržištích či datovém skladu mají jak výhody, tak i nevýhody.

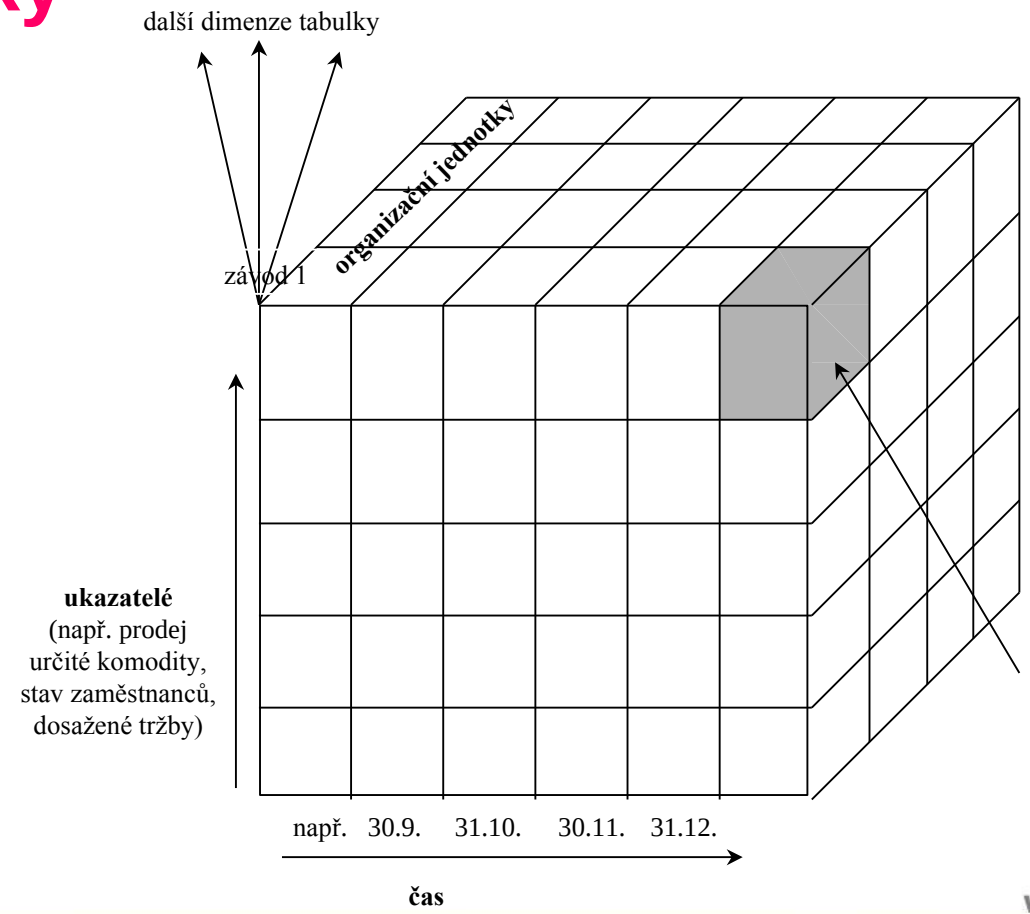
Transformace zdrojových dat



Základní vrstvy business intelligence



Architektura multidimenzionální databáze – OLAP kostky



Dimenzionální modelování dat ve skladu

Tabulky faktů

Fakta jsou údaje v numerickém či symbolickém (nenumerickém) vyjádření a sama o sobě jsou zcela neutrální.

Pozitivní či negativní význam faktů je dán pohledem na ně, tj. na tatáž data lze uplatňovat různá ***hlediska***.

Tato hlediska jsou zvána ***dimenze***.

Tabulky faktů obsahují obvykle velké množství řádků, které odpovídají pozorovaným a naměřeným faktům, a dále obsahuje externí připojení (tzv. cizí klíče neboli klíče

Dimenzionální modelování dat ve skladu

Tabulky faktů (pokračování)

Tabulky faktů obsahují obvykle velké množství *řádků*, které odpovídají pozorovaným a naměřeným faktům. Dále obsahují *externí připojení* (tzv. *cizí klíče* neboli klíče vytvořené z obsahu řádků – hodnot atributů – jiných tabulek).

Cizí klíče *připojují tabulky faktů k tabulkám dimenzí*. Tabulky faktů dále obsahují popisné atributy, potřebné pro provedení *rozhodovacích analýz a reportů*. Primárně tabulka faktů určuje, čím datový sklad podporuje rozhodovací analýzy.

Dimenzionální modelování dat ve skladu

Tabulky dimenzí

Tabulky dimenzí obklopují tabulky faktů, k nimž jsou připojeny cizími klíči.

Dimenzionální tabulky obsahují klasifikaci a agregační informaci o řádcích centrálních tabulek faktů.

Atributy dimenzionálních tabulek popisují data obsažená v tabulce faktů a určují, jak budou data analyzována a sumarizována.

Tabulky dimenzí mají vůči řádkům v tabulkách faktů vztah

Dimenzionální modelování dat ve skladu

Tabulky dimenzí (pokračování)

Tabulky dimenzí mají k řádkům v tabulkách faktů vztah 1:N.

Při dotazování hrají tabulky dimenzí roli typu *slice-and-dice*: ze všech hodnot v tabulce faktů vybírají hodnoty pro poskytnutí ad-hoc potřebné informace pro stanovené analýzy.

Dimenzionální modelování dat ve skladu

Propojení tabulek dimenzí s tabulkami faktů

Tabulky dimenzí jsou spojeny se svými tabulkami faktů cizími klíči, avšak *propojení může mít různou strukturu*.

Nejčastěji se jako prostředky implementace datového modelování v datových skladech používají schémata zvaná ***Star*** (hvězda) nebo ***Snowflake*** (sněhová vločka).

Tato schémata vyjadřují vztah mezi tabulkami faktů (vlastní data) a tabulkami dimenzí (pohledů na tato data).

Dimenzionální modelování dat ve skladu

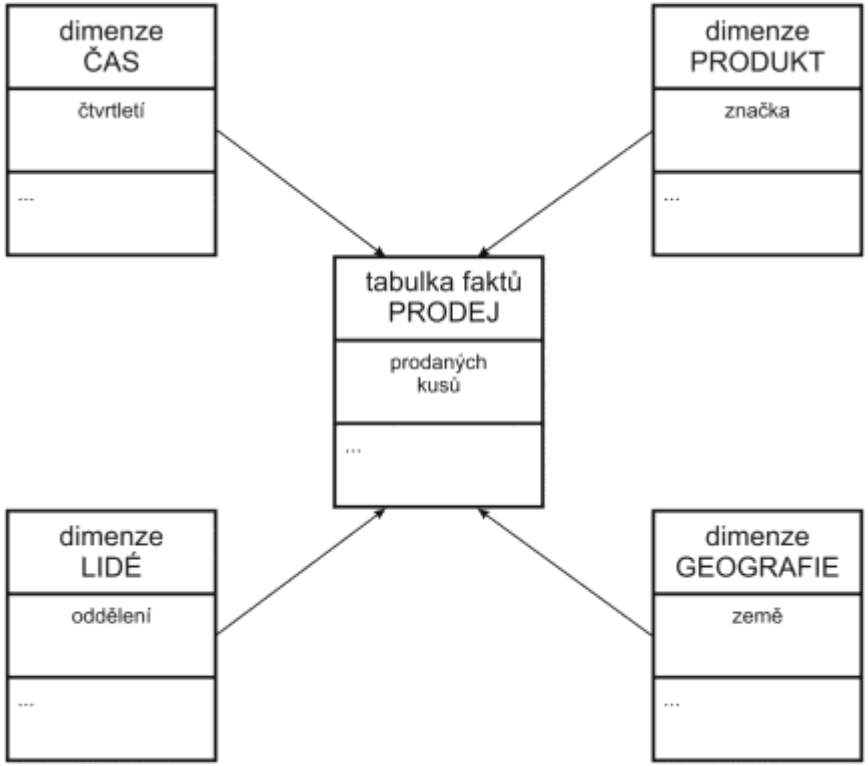
Schéma Star

Schéma jménem ***Star*** (tj. hvězda, proto někdy zvané jako hvězdicovité propojení dimenzí a faktů) je nejčastěji používané a nejjednodušší dimenzionální modelování.

Obsahuje centrální tabulku faktů obklopenou několika připojenými tabulkami dimenzí. Star poskytuje *časově rychlou odezvu na dotaz, je jednoduché, a snadno se udržuje v databázích určených pouze pro čtení*. Na schéma typu star lze hledět jako na speciální (jednodušší) případ propojení typu snowflake.

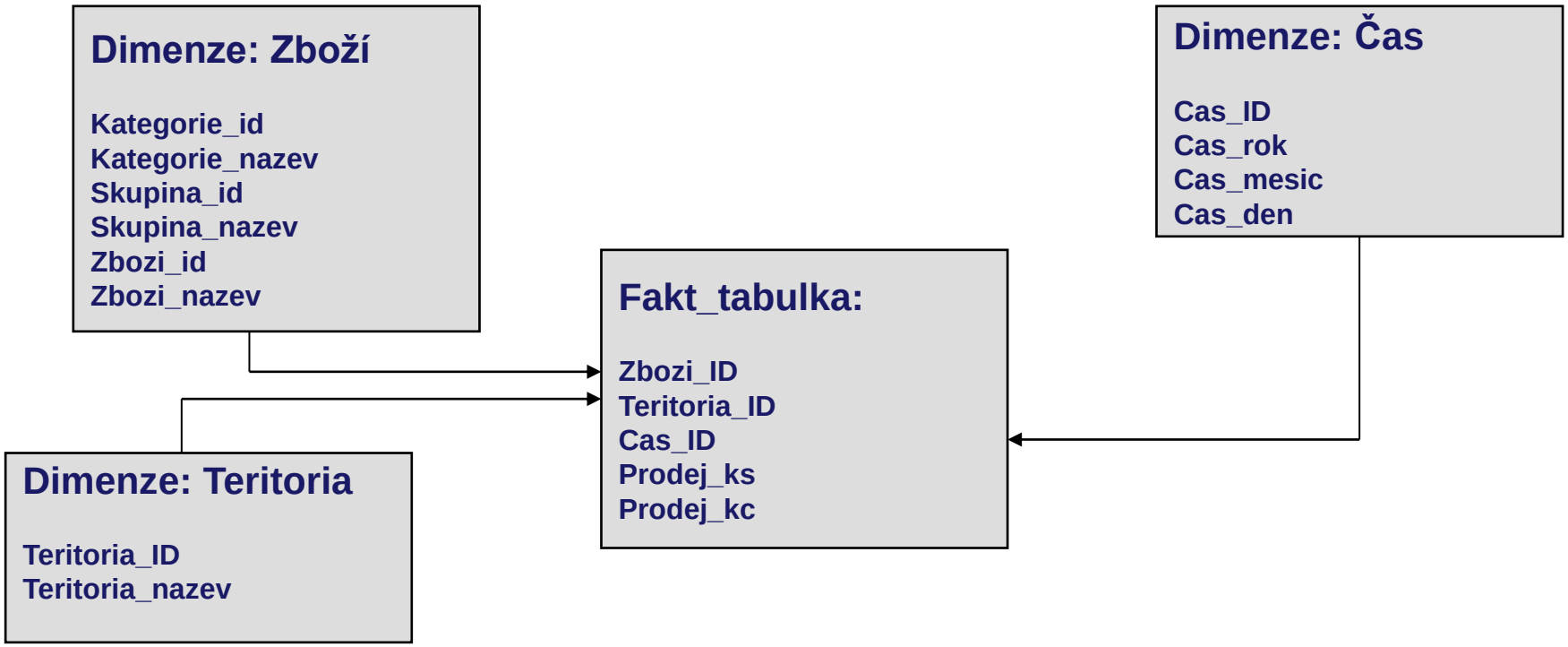
Dimenzionální modelování dat ve skladu

Schéma Star (pokračování)



Dimenzionální modelování dat ve skladu

Schéma Star (pokračování)



Dimenzionální modelování dat ve skladu

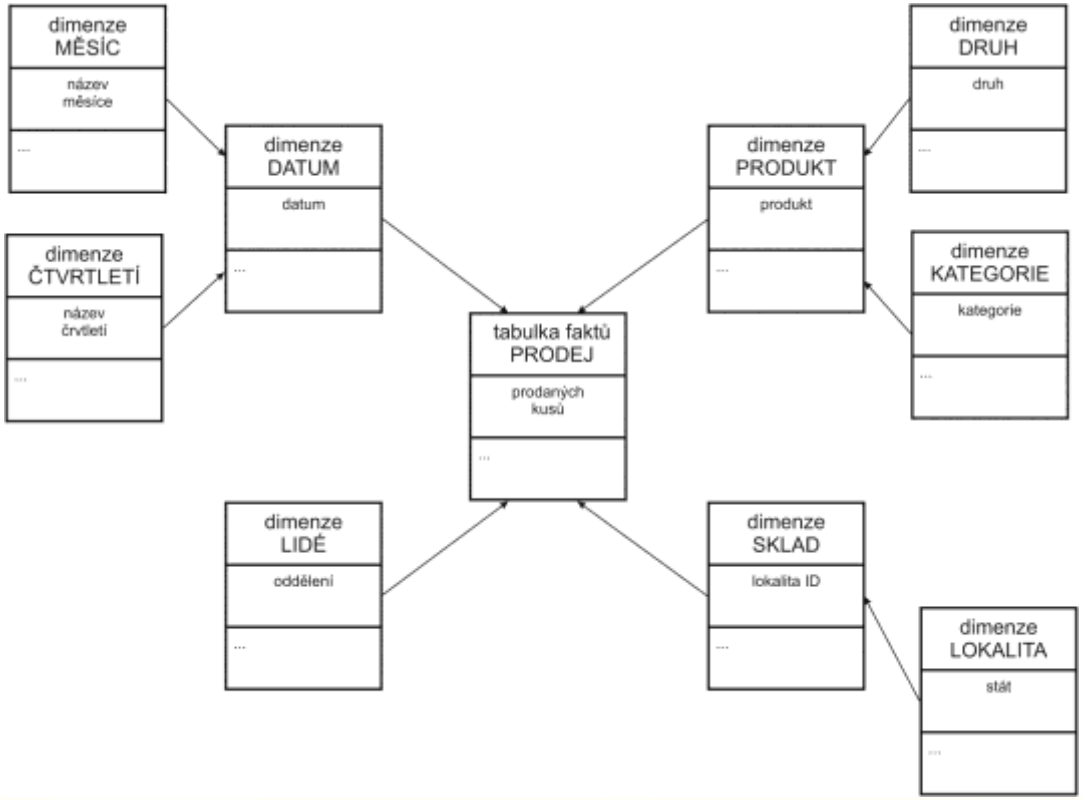
Schéma Snowflake

Schéma označované názvem ***Snowflake*** (tj. sněhová vločka) připomíná svým logickým uspořádáním tabulek v multidimenzionální databázi strukturu sněhové vločky.

Organizace tabulek pro snowflake se vyznačuje normalizací dimenzí na více mezi sebou propojených tabulkách – na rozdíl od schématu star, kde dimenze jsou denormalizovány, takže každá dimenze je představována individuální tabulkou.

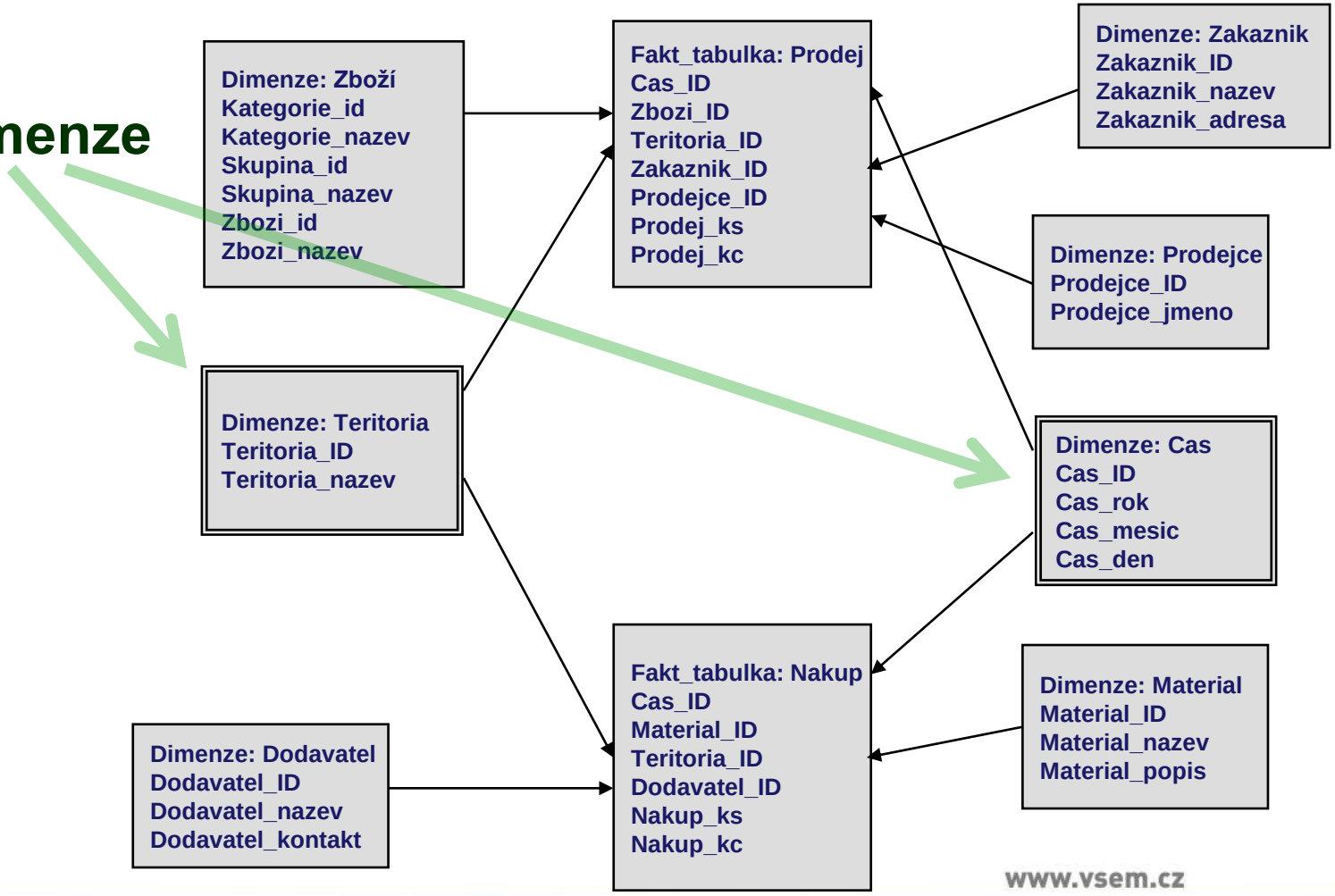
Dimenzionální modelování dat ve skladu

Schéma Snowflake (pokračování)

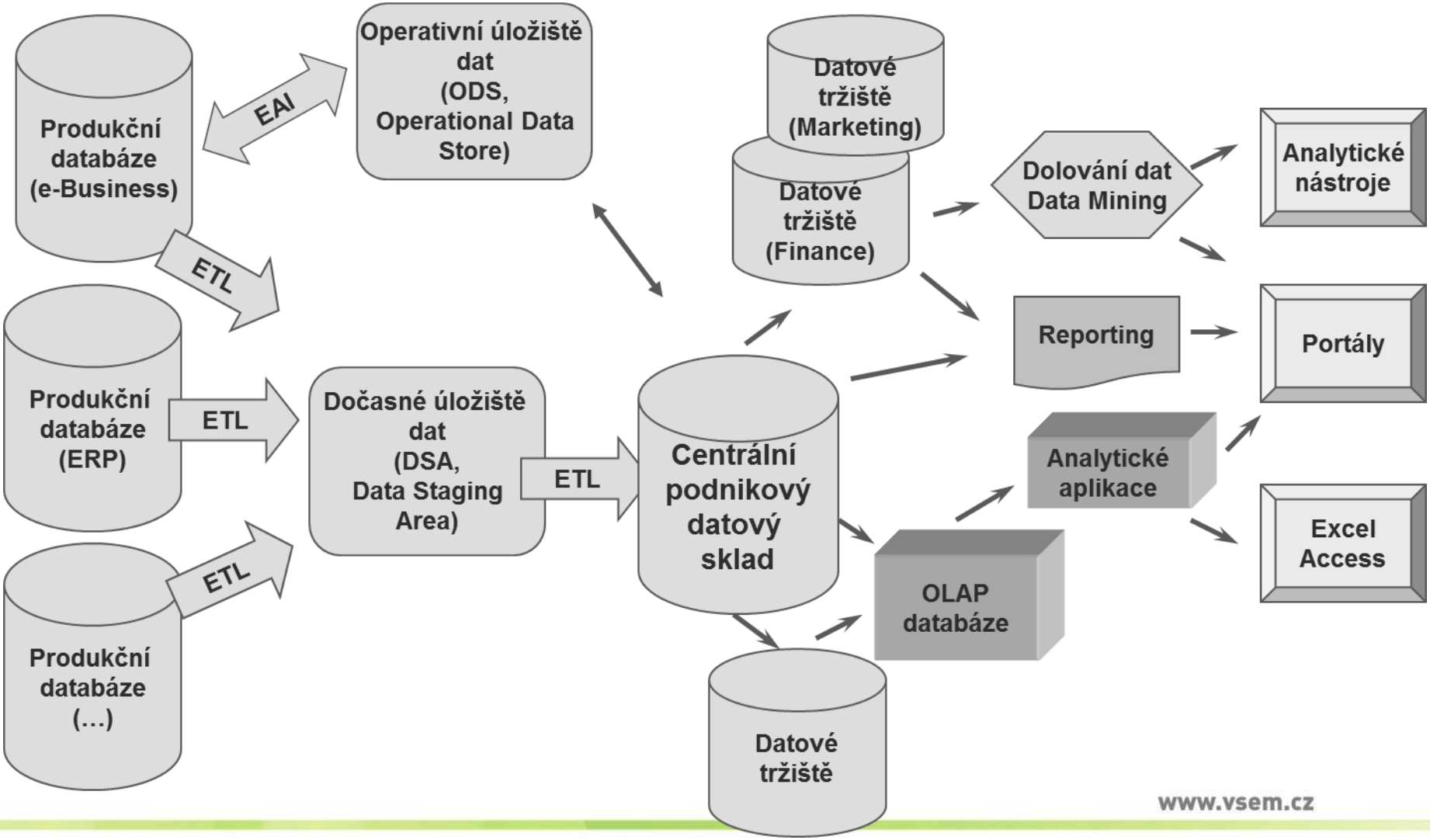


Dimenzionální modelování dat ve skladu

Sdílené dimenze



Komponenty BI řešení a jejich vazby



Datový sklad DWH a datová tržiště Dma

DWH - Datový sklad (**D**ata **W**are**h**ouse) – integrovaný, subjektově orientovaný. Stálá a časově rozlišená data, uspořádaná pro podporu potřeb managementu: Konsolidovaná, konsistentní, s časovou dimenzí, s orientací dle subjektů, normalizace/denormalizace.

DMa - Datové tržiště (**D**ata **Ma**rt) – „malý“ datový sklad určený pro omezený okruh uživatelů – oddělení, oblasti řízení. Obsahuje data stejného typu jako DWH. Snižuje dobu prvotní implementace a náklady; lepší orientace uživatelů v datech.

Požadavky na DWH a DMa

Informace z DWH musí být *jednoduše dostupné*:

DWH musí poskytovat informace jednoduše, obsah DWH musí být pochopitelný pro uživatele, musí nabízet informace v nejrůznějších kombinacích (slicing-and-dicing), zajišťovat co nejkratší dobu odezvy.

Informace z DWH musí být *presentovány konsistentně*:

Musí poskytovat věrohodné informace, tj. data musí být shromážděny z různých zdrojů, pečlivě kontrolovány a čištěny a poskytovány až tehdy, pokud jsou v pořádku.

Požadavky na DWH a DMa (pokračování)

Data musí být postavena na jasném vymezení *obsahu* a jasném *odlišení* pomocí jejich identifikátorů, resp. názvů.

DWH a DMa musí být *flexibilní* vzhledem ke změnám: Musí být *adaptabilní* ke změnám v uživatelských požadavcích, podnikovém prostředí, datových zdrojích, a technologiích.

DWH a DMa musí *zajišťovat bezpečnost dat*: DWH/DMa obsahuje kompletní a často citlivé informace, musí být zajištěno řízení přístupů k jednotlivým částem DWH/DMa.

Požadavky na DWH a DMa (pokračování)

DWH/DMa musí tvořit *základ pro zkvalitňování řídících a rozhodovacích procesů*: pro DWH/DMa musí být deklarovány efekty, které přinese (i když nemusí být vždy exaktně vyjádřeny a ve finančních ukazatelích).

Podniková komunita musí DWH/DMa *akceptovat* adekvátním způsobem: využití DWH/DMa (a BI) není, na rozdíl od transakčních systémů, nezbytné a je závislé na ochotě a invenci uživatelů. Proto je také nezbytné naplnit všechny předchozí požadavky.

Analýza dat v datovém skladu

Jakmile jsou data korektně uložena v datovém skladu, mohou být použita pro analýzy, jejichž výsledky podporují rozhodovací proces.

Nejpoužívanější technikou pro analýzy je **OLAP** (**O**n-**L**ine **A**nalytical **P**rocessing Systems).

OLAP umožňuje jednoduše a rychle odpovídat na ad-hoc dotazy tím, že vyšle analytické multidimenzionální dotazy směrem ke skladištím dat (centralizované EDW/DWH či decentralizované DMa).

Spolupráce systémů OLAP a OLTP

OLTP (**O**n-**L**ine **T**ransaction **P**rocessing systems) je transakční systém, který se primárně používá pro vyhledávání a ukládání dat souvisejících s běžnými podnikatelskými funkcemi.

OLTP systémy jsou efektivní nástroje pro stanovování kritických podnikových potřeb, automatizaci denních obchodních transakcí, vytváření zpráv v reálném čase a rutinních analýz.

Na druhé straně je třeba vidět, že OLTP systémy nebyly určeny pro ad-hoc analýzy a komplexní dotazování beroucí do úvahy množství datových položek.

Spolupráce systémů OLAP a OLTP (pokrač.)

Systém **OLAP** je oproti tomu konstruován tak, že se zaměřuje právě na řešení uvedených informačních potřeb poskytováním výsledků prostřednictvím ad-hoc analýz rozsáhlých dat dané organizace či podniku, a to daleko výkonněji a účinněji.

V každém případě OLAP i OLTP na sebe navzájem velmi silně spoléhají:

OLAP používá data získaná nástroji OLTP a na druhé straně OLTP automatizuje podnikové procesy řízené rozhodováním podporovaným nástroji OLAP.

Operace poskytované systémy OLAP

Hlavní operační struktura systému OLAP je založena na konceptu zvaném **kostka** (v originále **cube**).

Kostkou (nejen třírozměrnou, obecně vícerozměrnou) se v OLAP rozumí *multidimenzionální datová struktura*, která umožňuje rychlou analýzu dat.

Lze na ni hledět také jako na schopnost efektivní manipulace a analýzy dat z vícerozměrných perspektiv.

Operace poskytované systémy OLAP (pokrač.)

Uspořádání dat do kostek si klade za cíl překonat *omezení relačních databází*, které nejsou vhodně vybaveny pro *okamžité* (resp. téměř okamžité) *analýzy velkého množství dat naráz*.

Relační databáze jsou velice dobře vybaveny pro manipulace se záznamy (přidávání, rušení, aktualizace), které představují posloupnost transakcí. V relačních databázích existuje mnoho nástrojů pro vytváření reportů, ale tyto nástroje jsou pomalé, dojde-li na zpracování multidimenzionálního dotazu, který zahrnuje mnoho databázových tabulek.

Operace poskytované systémy OLAP (pokrač.)

Pomocí OLAP může analytik procházet databází a vyhledávat specifické podmnožiny dat včetně jejich časového vývoje změnou datové orientace a definováním analytických kalkulací.

Tyto typy uživatelsky spuštěných datových navigací pomocí specifikovaných datových výřezů (slice) a přístupů k více či méně detailním datovým položkám (drill down/drill up) prostřednictvím agregací a disagregací jsou někdy nazývány ***slice-and-dice***.

Hlavní operace OLAP

Slice: „rozkrájet na plátky“– podmnožina multidimenzionálního souboru (nejčastěji bývá v dvourozměrné representaci).

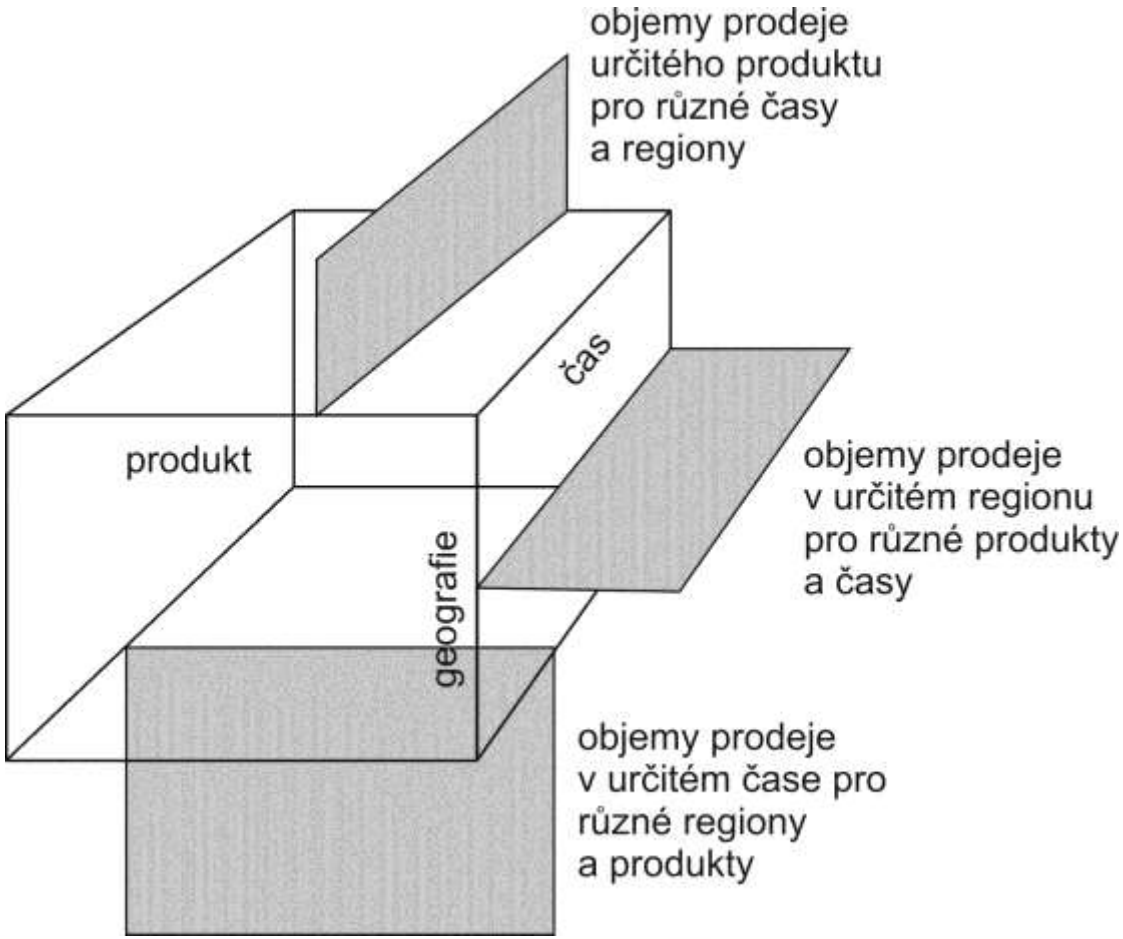
Dice: „nakrájet na kostky“ – operace slice na více než dvou dimenzích kostky.

Drill down: přístup hierarchicky níže k detailnějším položkám.

Drill up: přístup hierarchicky výše k obecnějším položkám.

Roll up: výpočty všech datových vztahů (souvislostí) pro jednu či více dimenzí.

Ilustrace operace OLAP zvané *slice*



Různé typy OLAP databází

ROLAP: relační OLAP, řeší multidimenzionalitu uložením dat v relační databázi; výhoda ve využití relačních SŘBD.

MOLAP: multidimenzionální OLAP, vyznačující se speciálním uložením dat v multidimenzionálních OLAP kostkách (výhodou je optimalizované uložení objemných dat).

HOLAP: hybridní OLAP, kombinace předchozích přístupů – detailní data v relační databázi a agregace v OLAP kostkách.

Různé typy OLAP databází

DOLAP: desktop OLAP; umožňuje připojit se k centrálnímu úložišti OLAP dat a stáhnout si podmnožinu kostky na lokální počítač. Analytické operace jsou prováděny nad lokální kostkou – výhodné pro mobilní aplikace a podporu mobilních uživatelů.

WOLAP: webový OLAP, kombinace OLAP a webových technologií umožňuje např. provádět analýzy připojením mobilního počítače k datovému skladu pomocí sítě (Internet).

Vytvoření datového skladu

Jedenáct kritických kroků pro vytvoření datového skladu:

1. Ustanovení požadavků na aktualizaci dat a dohod o úrovni servisu
2. Identifikace zdrojů dat a způsob jejich řízení
3. Naplánování kvality dat
4. Vytvoření datového modelu
5. Výběr vhodného ETL nástroje
6. Výběr vhodného software relační databáze a operačního systému
7. Přenos dat
8. Konverze dat
9. Uvedení všech částí do souladu
10. Naplánování vyčistění a archivace dat
11. Podpora koncových uživatelů (analytiků, vedení podniku)

Osvědčená vodítka výstavby datových skladů

Projekt musí vyhovovat *strategii* konkrétní společnosti a jejím podnikatelským cílům.

Musí být ze strany výkonných vedoucích pracovníků, vedení firmy a uživatelů dosaženo *kompletního a absolutního přijetí* navrhovaného a implementovaného projektu.

Je důležité úspěšně zvládnout *požadavky uživatelů* vztahující se k celkovému projektu.

Datový sklad musí být vytvářen *postupně*.

Osvědčená vodítka výstavby datových skladů (pokračování)

Přizpůsobivost a možnost *rozšiřování* je nutno vzít do úvahy hned od počátku implementace.

Realizace projektu musí být vedena jak pracovníky z oblasti *informačních technologií*, tak *profesionály v podnikání*.

Do datového skladu ukládat pouze *data*, která jsou *relevantní* pro podporu rozhodování a s tím spojených analýz, jsou *vyčistěna* a pocházejí z *důvěryhodných* a *známých* zdrojů interních a externích vůči konkrétní organizaci či podniku.

Osvědčená vodítka výstavby datových skladů (pokračování)

Nepodceňovat patřičná a *kvalitní školení*, protože zamýšlení uživatelé nemusí být vždy dostatečně informaticky vzdělaní.

Je zapotřebí dbát na výběr *osvědčených* softwarových a hardwarových nástrojů, které dobře zapadnou do již existující infrastruktury podniku.

Při realizaci datového skladu vzít na vědomí existující nebo případně možné *organizační tlaky*, *vnitřní politiku podniku*, a *lokální konflikty*.

Možná rizika při budování datových skladů

Zahájit realizaci projektu datového skladu bez zajištění dostatečného zdroje pokrytí nákladů.

Mít očekávání, která nelze splnit.

Praktikovat chybnou politiku.

Ukládat do skladu informace jen proto, že jsou k dispozici.

Domnívat se, že návrh databáze pro datový sklad je totéž jako návrh transakční databáze.

Možná rizika při budování datových skladů (pokračování)

Vybrat manažera datového skladu, který je zaměřen spíše technologicky než na koncové uživatele.

Zaměřit se na tradiční vnitropodniková záznamově orientovaná data a ignorovat hodnotu dat externích, textových, obrazových, zvukových a typu video.

Dodávat data s překrývajícími se a zmatenými definicemi.

Možná rizika při budování datových skladů (pokračování)

Věřit navržené výkonnosti, kapacitě a rozšiřitelnosti realizovaného datového skladu.

Domnívat se, že potíže zmizely v okamžiku, kdy byl datový sklad uveden do chodu.

Zaměřit se jen na ad-hoc dolování znalosti z dat a periodické zpravodajství (reporty) místo na včasné upozornění (varování) na neočekávanou změnu v dosavadní situaci.

Datové sklady v reálném čase

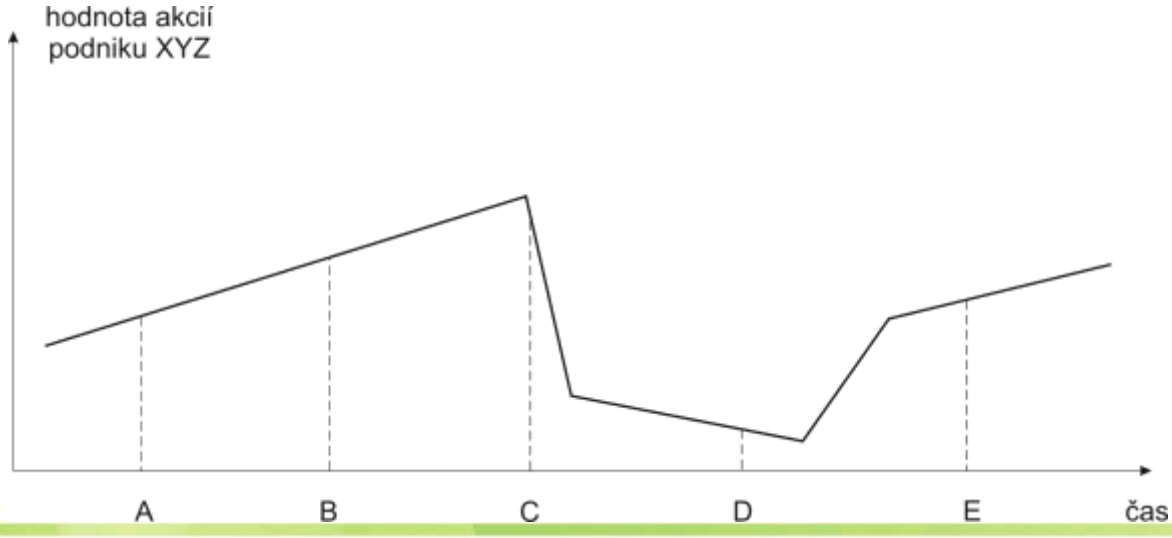
Proces **RDW** (**R**eal-time **D**ata **W**arehousing, skladování dat v reálném čase), známý i pod jménem **ADW** (**A**ctive **D**ata **W**arehousing, aktivní datové skladování).

Je to proces ukládání a poskytování dat tak, jak se právě stávají dostupnými.

Proces RDW/ADW se vyvinul z konceptu EDW/DWH. Aktivní vlastnosti RDW/ADW obohacují a doplňují tradiční funkce datových skladů a rozšiřují jejich použití do oblasti taktických rozhodovacích procesů vzhledem k reálnému času.

„Více pravd“ v souvislosti s reálným časem

Objevuje se možnost vzniku „několika pravd“ v závislosti na časovém okamžiku, kdy kdo dostane pro své analýzy jaká data – údaje se mohou měnit někdy velmi rychle, v rámci sekund (burzy), takže „okamžitá pravda“ může být i ***funkcí času*** poskytující rychle za sebou rozdílné hodnoty.



Bezpečnost a privátnost informací

Hlavní body bezpečnosti a privátnosti podnikových dat

Definice a vytvoření efektivní podnikové a bezpečnostní politiky a příslušných procedur. Efektivní bezpečnostní politika by měla začínat ve vrcholovém vedení.

Implementace logických bezpečnostních procedur a technik zamezujících neoprávněné přístupy k datům zahrnují spolehlivé prokázání toho, že uživatel je oprávněn s daty pracovat, jakým způsobem k datům přistupovat (číst, zapisovat, aktualizovat, rušit).

Bezpečnost a privátnost informací

Hlavní body bezpečnosti a privátnosti podnikových dat
(pokračování)

Důležitou součástí procedur jsou také šifrovací a dešifrovací technologie pro data.

Omezení fyzického přístupu do centra datového prostředí.

Zavedení účinného procesu ověřování a vnitřní kontroly s důrazem na bezpečnost a privátnost dat.

Dolování z dat pro business intelligence

Dolování z dat *(data mining)* představuje způsob *získávání informace a znalosti ze shromážděných dat*, která má podnik či organizace k dispozici a která jsou nějakým způsobem organizována a uložena.

Dolování z dat je založeno na do značné míry empirickém a experimentálním přístupu, protože řeší problematiku, na niž nelze použít tradiční (matematické či informatické) metody.

Induktivní získávání znalostí ze vzorků dat

Postupy dolování z dat jsou založeny na ***induktivním*** získávání znalosti *generalizací* nějakých typických společných vlastností jednotlivých konkrétních vzorků dat, které vícerozměrně (formou vektorů) popisují entity reálného světa.

Hodnoty vlastností popisujících individuální zaznamenané entity jsou získávány *měřením*, *pozorováním*, apod. Je-li k dispozici *dostatečný počet entit*, pak je lze např. rozdělit na skupiny, které mají něco společného, což je jedním z cílů dolování z dat a využívá to například marketing.

Znalost získaná *dolováním*

„Vydolovaná“ znalost může sloužit např. ve formě ***pravidel***, podle nichž se podnik bude řídit, nebude odpuzovat zákazníky nesmyslnými nabídkami a naopak je bude lákat nabídkami pro ně atraktivními. Pravidla tvoří jednu z mnoha forem znalosti, a to znalosti člověku srozumitelné.

Řadu pravidel lze odhalit i tradičními statistickými postupy, ale praxe ukázala, že na mnoha případech naopak statistika selhává, zejména tím, že v ní mizí individuálnost konkrétních případů a výsledkem je číselné vyjádření vlastností celku, resp. jednotlivých větších částí celku. Nebo nejsou splněny podmínky pro použití statistických postupů.

Termín *dolování znalosti*

Význam termínu *dolování z dat* spočívá v procesu, jehož úkolem je vyhledat v datech zřejmě existující, ale neznámé vzory (ve smyslu něčeho společného pro nějaké skupiny datových vzorků) – tzv. *pattern recognition* (rozpoznávání vzorů).

Z terminologického hlediska existuje celá řada názvů: *extrakce znalosti* (knowledge extraction), *odhalování znalosti* (knowledge discovery), *analýza vzorů* (pattern amalysis), *datová archeologie* (data archeology), *sklizeň informace* (information harvesting), *vyhledávání vzorů* (pattern searching), *prohrabávání se daty* (data dredging), a další.

Definice pojmu *dolování z dat*

Dolování z dat (*data mining*) je termín používaný k označení procesu vyhledávání znalosti skryté v rozsáhlých objemech dat z reálného světa popisujících velká množství jednotlivých konkrétních pozorovaných případů.

Hlavní charakteristiky, cíle a vlastnosti prostředí procesu dolování z dat

Data jsou často hluboce skryta v rozsáhlých databázích obsahujících údaje sbírané po mnoho let. V mnoha případech jsou data očištěna a konsolidována v datových skladech.

Prostředí pro dolování z dat má obvykle architekturu typu klient-server nebo informačního systému založeného na webu. paralelní zpracování.

Hlavní charakteristiky, cíle a vlastnosti prostředí procesu dolování z dat (pokračování)

Pokročilé nové nástroje, včetně účinných vizualizačních prostředků, umožnují odstranit informační „hlušinu“, která „zasypává“ užitečnou informaci uloženou v podnikových souborech dat nebo ve veřejných archívních záznamech. Nalezení a odstranění dat nerelevantních z hlediska konkrétního řešeného problému vyžaduje zpracování a synchronizaci dat tak, aby byly získány správné výsledky – tato část může zabrat většinu doby procesu dolování z dat. Odborníci na dolování z dat zhusta zkoumají i potenciální užitečnost tzv. *měkkých dat* (nestrukturované textové soubory, obsah podnikového intranetu, apod.).

Hlavní charakteristiky, cíle a vlastnosti prostředí procesu dolování z dat (pokračování)

„Dolovač“ z dat je často i koncový uživatel vyzbrojený nástroji pro vyhledání konkrétních detailů (*data drilling*) v obecné informaci a dalšími silnými dotazovacími prostředky umožňujícími klást ad-hoc dotazy a tak získávat odpovědi rychle, bez nutnosti psát počítačové programy na jednotlivé funkce.

Zaměření se na určité konkrétní věci často přináší neočekávané výsledky a na koncové uživatele klade značné nároky z hlediska kreativního myšlení během celého dolovacího procesu, nevyjímaje nápaditou interpretaci nalezených výsledků.

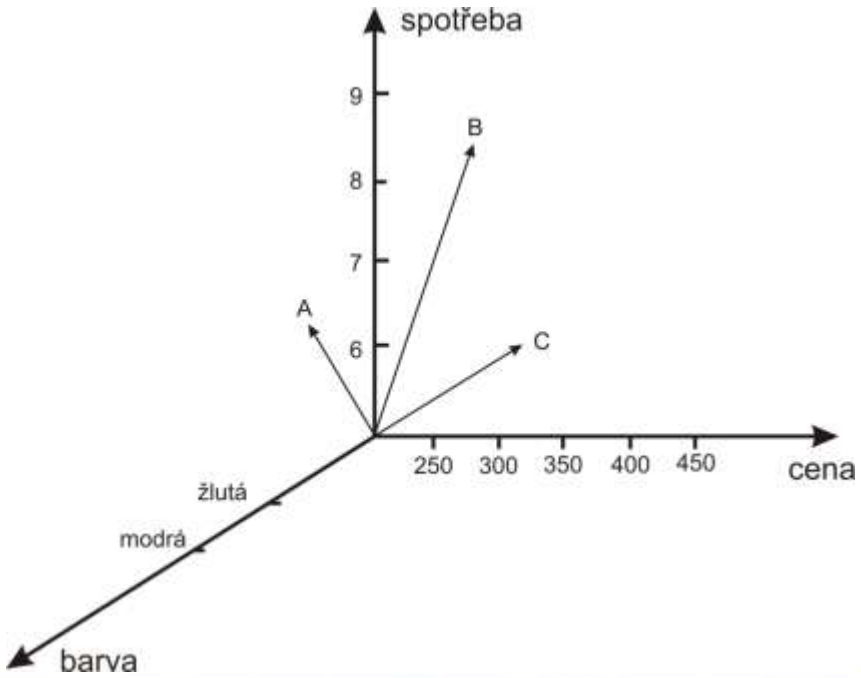
Hlavní charakteristiky, cíle a vlastnosti prostředí procesu dolování z dat (pokračování)

Nástroje dolování z dat jsou snadno kombinovatelné s tabulkovými procesory a dalším vhodným software. Tímto způsobem mohou být prozkoumávaná data rychle a snadno analyzována a rozvinuta do šíře.

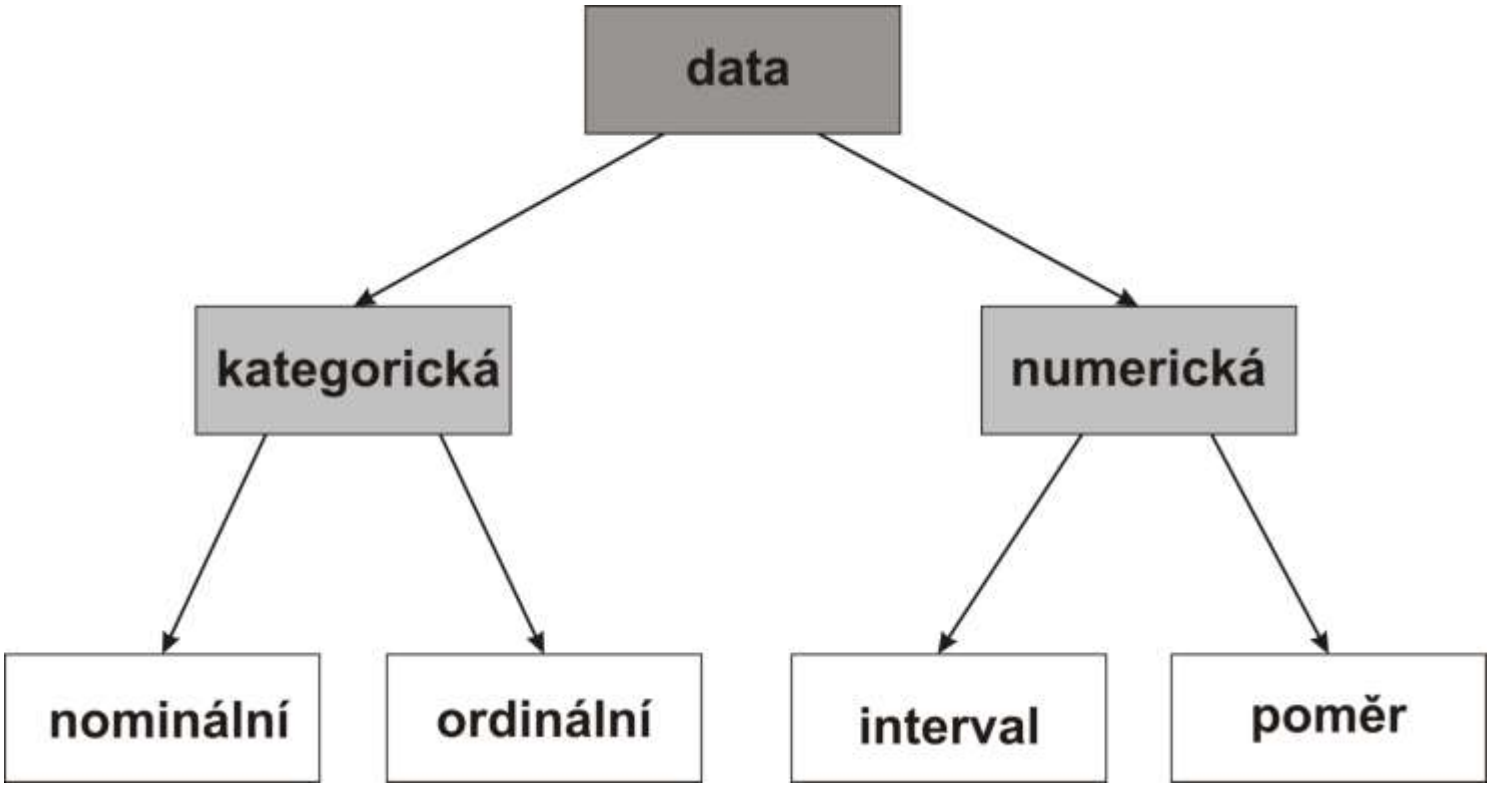
Kvůli velmi rozsáhlým objemům dat bývá někdy zapotřebí použít paralelní zpracování.

Vektorová representace vzorků dat

kategorie	barva	cena v 1000 Kč	spotřeba l/100 km	oblíben u zákazníků
automobil A	modrý	250	6.2	ano
automobil B	žlutý	300	8.5	ne
automobil C	modrý	450	7.3	ano



Taxonomie dat



Vyhledávání čtyř druhů vzorů

1. **Asociace** (sdružování) mohou najít, které věci se typicky vyskytují společně, v nákupních koších, např. pivo a dětské plenky. Závěry z takto nalezených vztahů musí pak udělat odpovědní pracovníci.

2. **Predikce** (předpovědi) jsou schopny sdělit, co se vyskytne v budoucnu na základě toho, k čemu došlo v minulosti, například kdy lze očekávat pokles zájmu o jeden druh zboží a nárůst zájmu o jiný druh.

3. **Shluky** (clusters) identifikují přirozené seskupování věcí na základě jejich známých vlastností, například rozdělení zákazníků do různých segmentů na základě demografie a nákupního chování v minulosti.

4. **Sekvenční vztahy** odhalují události uspořádané v čase, např. že zákazník banky, který má založený účet, si do roka otevře spořící účet následovaný účtem investičním.

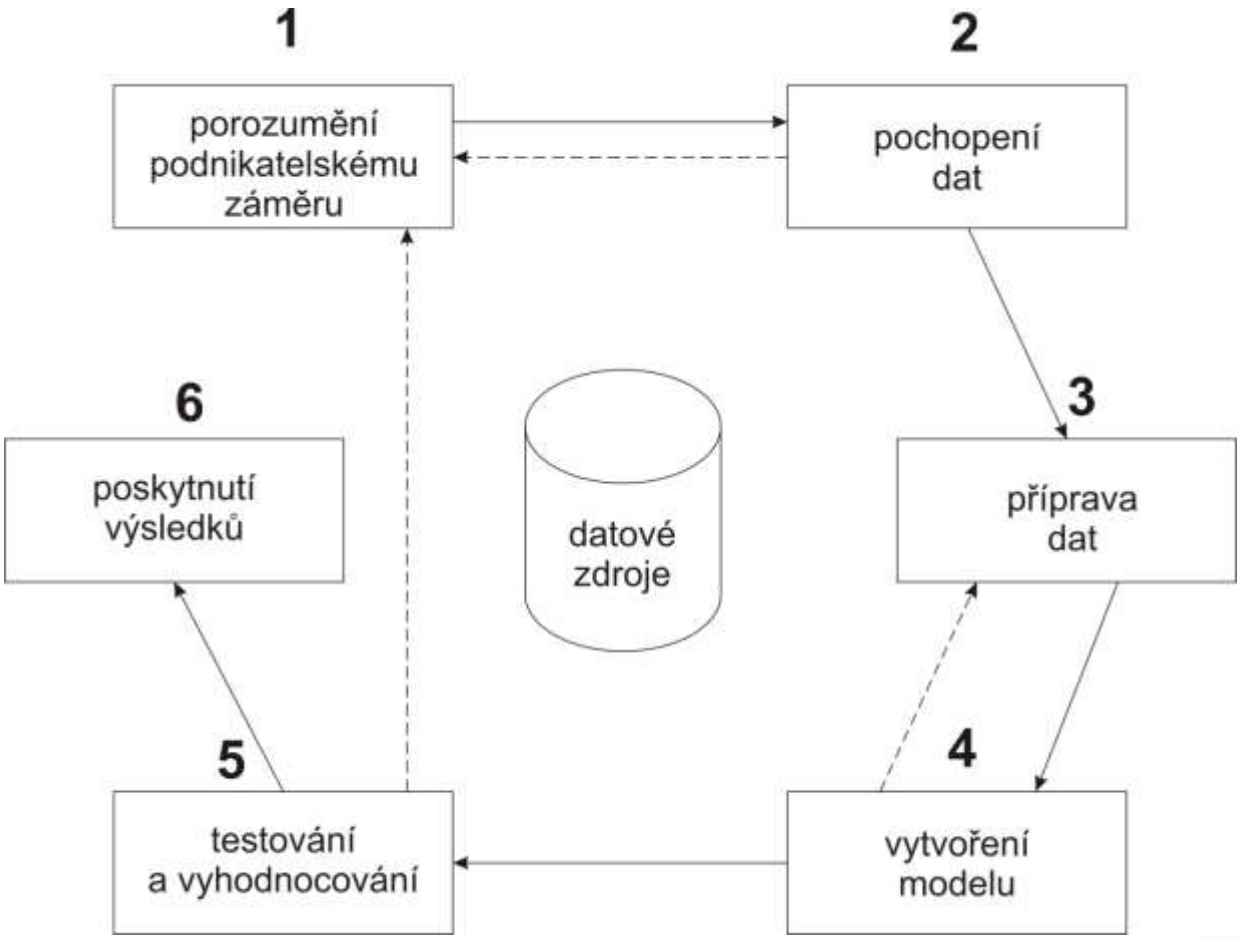
Tři typy trénování strojů pro dolování znalosti

1. **Řízené učení**, učení s učitelem (*supervised learning*): k dispozici jsou trénovací příklady, u kterých je známo, do které třídy patří. Role „učitele“ je v dohlížení na správnost zařazování, a k tomu učitel (součást trénovacího algoritmu) používá znalost o správné odpovědi, takže poskytuje „žákovi“ zpětnou vazbu

2. **Neřízené učení**, učení bez učitele (*unsupervised learning*): zde nejsou známy správné odpovědi, takže trénovanému algoritmu nelze sdělit, zda udělal chybu či ne; zpětná vazba chybí. „Žák“ může pouze hledat pomocí hodnot atributů, které instance k sobě patří, a tak vytvořit shluky (*clustering*).

3. **Částečně řízené učení** (*semisupervised learning*) je založeno na tom, že trénovacímu algoritmu se poskytne velmi omezený počet „ukázkových“ příkladů se známým zařazením do nějaké kategorie nebo třídy. Algoritmus pak postupně bere instance s neznámým zařazením a porovnává je s těmi známými. Neznámou instanci pak přiřadí ke vzoru, jemuž se nejvíc podobá svými vlastnostmi (atributy).

Dolování z dat procesem CRISP-DM



Metody dolování z dat: *Klasifikace*

Patří pravděpodobně k nejčastějším metodám, patří k tzv. řízenému učení či učení s učitelem. Cílem je správně zařazovat do tříd entity, u nichž je příslušnost do konkrétní třídy neznámá a stanovuje se na základě podobnosti se známými, „trénovacími“ entitami.

Pokud se klasifikuje do spojitých tříd, používají se metody regrese.

K *vyhodnocení kvality* klasifikační (či regresní) metody se používá řada parametrů, ke kterým patří následující nejvyužívanější:

Vyhodnocování kvality klasifikátoru

Prediktivní spolehlivost (*accuracy*). Jedná se o schopnost správně zařadit do příslušné třídy to, co do ní patří.

Rychlost. Výpočetní nároky na trénování a aplikaci algoritmu. Je-li více stejně spolehlivých algoritmů, vybere se nejrychlejší.

Robustnost. Schopnost modelu generovat rozumně spolehlivé výsledky i pro zašuměná data nebo data s chybějícími a uměle nahrazenými hodnotami.

Rozšiřitelnost. Schopnost efektivně vytvářet predikční model i pro velmi rozsáhlé objemy dat.

Interpretovatelnost. Míra možnosti porozumět modelu a pochopit jeho vnitřní principy (např. jak a proč dosahuje daný model své výsledky).

Odhad spolehlivosti klasifikačních modelů

Jednou z nejpoužívanějších metod je tzv. **matice zmatení** (*confusion matrix*). Ta přímo ukazuje, kolik případů z které třídy bylo chybně přiřazeno třídám jiným.

třída A	třída B	třída C	
50	0	0	třída A
9	36	5	třída B
2	7	41	třída C

Odhad spolehlivosti klasifikačních modelů

Pro měření spolehlivosti se používají také tyto parametry:

TP (*True Positive*, správně pozitivní)
TN (*True Negative*, správně negativní)
FP (*False Positive*, chybně positivní)
FN (*False Negative*, chybně negativní).

Jedná se o počty instancí, které byly do tříd zařazeny či nezařazeny, přičemž **TP** znamená, že byly zařazeny správně (protože tam opravdu patří), **TF** že nesprávně (protože tam opravdu nepatří), **FP**, že byly zařazeny, ačkoliv tam ve nepatří, a **FN**, že do třídy nebyly zařazeny, ačkoliv tam patří.

Odhad spolehlivosti klasifikačních modelů
(pokračování)

Pomocí těchto parametrů lze počítat další parametry, např.:

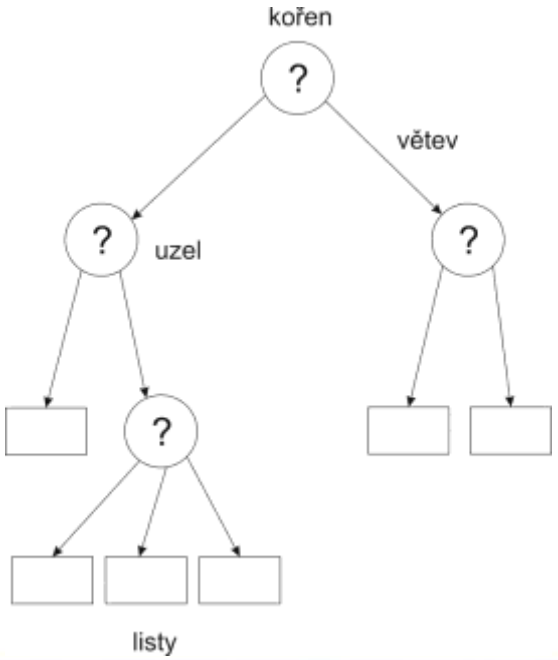
Spolehlivost (*accuracy*) = (TP+TN) / (TP+TN+FP+FN).

Přesnost (*precision*) = TP / (TP+FP).

Vybavení si (*recall*)= TP / (TP+FN).

Některé klasifikační techniky

Rozhodovací strom se skládá z tzv. uzlů, kde jsou dotazy na hodnotu vybrané proměnné, a dle odpovědi se větví buď do dalších dotazů, nebo je větev zakončena listem čili odpovědí.



Rozhodovací stromy

Jeden z nejznámějších algoritmů na generování stromů pochází od australského vědce Rosse Quinlana, který navrhl a implementoval metodu známou pod názvem c4.5 (ta je k dispozici bezplatně), resp. později založil firmu RuleQuest, kde lze zakoupit pokročilejší implementaci zvanou c5.

Algoritmy c4.5 a c5 jsou založeny na myšlence *minimalizace entropie*, neboli na snížení chaosu, který vznikl smícháním instancí patřících do různých tříd. Podaří-li se od sebe instance oddělit, entropie (chaos) se sníží. Dokonalé oddělení instancí (jejich zařazení do správných tříd) vede k nulové entropii (žádný chaos).

Rozhodovací stromy (pokračování)

Algoritmus c4.5 a c5 pracuje tak, že postupně zkouší, který dotaz na jaký atribut nejlépe rozdělí heterogenní množinu na homogennější podmnožiny.

Ten atribut, který to udělá nejlépe, je vybrán do dotazovacího uzlu. Pokud již nelze entropii nijak snížit, generování stromu končí. Jinak algoritmus zkouší další dotazy a další atributy jako v předchozím kroku. Ne vždy lze dosáhnout nulové entropie v listech, ale většinou se podaří původní množinu rozdělit na homogennější podmnožiny, které již obsahují výraznější převahu prvků jedné třídy.

Další základní klasifikační algoritmy

Mezi některé často používané techniky patří i následující:

k-NN (k-*nearest neighbors*, *k* nejblžších sousedů):

Do databáze se uloží vybrané známé případy a když se objeví případ neznámý, porovná se s uloženými vzory. Obdrží příslušnost do té třídy, do níž náleží nejpodobnější vzor soused, tj. ten, který je nejblíž. Případně lze vybrat více podobných sousedů a zvolit tu třídu, do níž patří většina z nejbližších sousedů. Parametr ***k*** se zadává a určuje, s kolika sousedy se má neznámý případ porovnávat.

Další základní klasifikační algoritmy

Umělá neuronová síť:

Snaha napodobit biologické učení centrální nervové soustavy. Uzly (neurony) představující přenosové funkce jsou propojeny do sítě. Síť má vstupní vrstvu, nejméně jednu vnitřní (tzv. skrytou), a výstupní.

Každý uzel vyšle výstupní signál, pokud je na vstupu dostatečně silný vstupní signál. Hodnoty výstupů se sečítají a ve vnitřních vrstvách tvoří vstup do dalších neuronů. Pokud na hodnoty atributů na vstupu odpoví výstup sítě správně, není nutno nic měnit. Pokud nesprávně, mění se iterativně hodnoty tzv. vah, což jsou čísla, která zesilují nebo zeslabují význam všech propojení mezi uzly. Cílem je najít takové hodnoty vah, kdy síť odpovídá správně. Umělé neuronové sítě se učí pomalu, ale jsou schopny najít téměř libovolné tvary hranic oddělujících hranice tříd.

Další základní klasifikační algoritmy

Naivní Bayes:

Metoda založená na inferenčním vzorci Thomase Bayese. Ve vzorci vystupují podmíněné a nepodmíněné pravděpodobnosti. Bayesův algoritmus neříká, kam instance patří a kam ne; místo toho poskytuje pravděpodobnosti, s jakými instance náleží do jednotlivých tříd, a obvykle se vybere takové náležení, které má pravděpodobnost nejvyšší.

Naivita spočívá v tom, že se předpokládá, že hodnoty jednotlivých atributů jsou na sobě nezávislé. Výsledky mohou pak být o něco horší, pokud v datech existuje nějaká vzájemná závislost atributů, avšak naivní přístup výrazně snižuje výpočetní složitost.

Používá se např. běžně pro klasifikaci, zda je e-mail spam či ne.

Shlukování (*clustering*)

Shlukovací algoritmy v datech hledají, které instance jsou si navzájem podobné. Pokud splňují danou míru podobnosti, vytvoří tzv. **shluk** neboli skupinu, která již představuje hledané zobecnění. Úloha je ztížena skutečností, že obvykle se neví, kolik by těch shluků mělo být – algoritmy jsou v principu schopny vytvořit libovolný počet shluků, omezený pouze počtem instancí (v krajním případě pak každý shluk obsahuje jedinou instanci, takže nebylo dosaženo zobecnění).

Podobnost se stanovuje na základě blízkosti či vzdálenosti bodů v abstraktním prostoru, nebo pomocí úhlu mezi vektory představujcmi jednotlivé instance.

Shlukovací algoritmus *k*-means

Nejznámější je algoritmus zvaný ***k*-means** (k-středů). Hodnota ***k*** se zadává a určuje, kolik shluků se má vytvořit. Pojem ***střed*** představuje střed každého shluku, něco jako těžiště dané výskytem (pozicemi v prostoru) instancí zařazených do daného shluku. ***k*-means** hledá tedy pozice středů shluků. Na počátku se neví, kde ty středy mají ležet, takže se vygeneruje náhodně ***k*** pozic. Pak se předkládají jednotlivé instance representované jako mnoharozměrné body. Ke každé instanci se zjistí, který střed je jí nejblíž, a ten se pak k ní posune. Takto se iterativně postupuje tak dlouho, dokud už k žádnému posuvu nedochází nebo pokud byl dosažen uživatelem stanovený limit počtu iterací. Nakonec jsou k dispozici nalezené středy shluků, které již představují nějaké třídy. Objeví-li se nová neznámá instance, najde si nejbližší střed a bude pak patřit do příslušného shluku (po skončení iteračního trénování už mají středy svá pevná neproměnná místa).

Asociační techniky

Asociační techniky se používají na objevování **asociačních pravidel**.

V praxi se velmi dobře uplatňují například při zkoumání nákupních košů – hledají se asociace mezi dvojicemi, trojicemi, atd., jednotlivých typů zboží v zákazníkově nákupním vozíku či koši.

Velmi efektivní, často používaná a principiálně jednoduchá je metoda zvaná algoritmus **Apriori**, která našla svou významnou pozici např. v marketingu.

Asociační metoda *Apriori*

Příklad použití metody **Apriori** – nákupní koš:

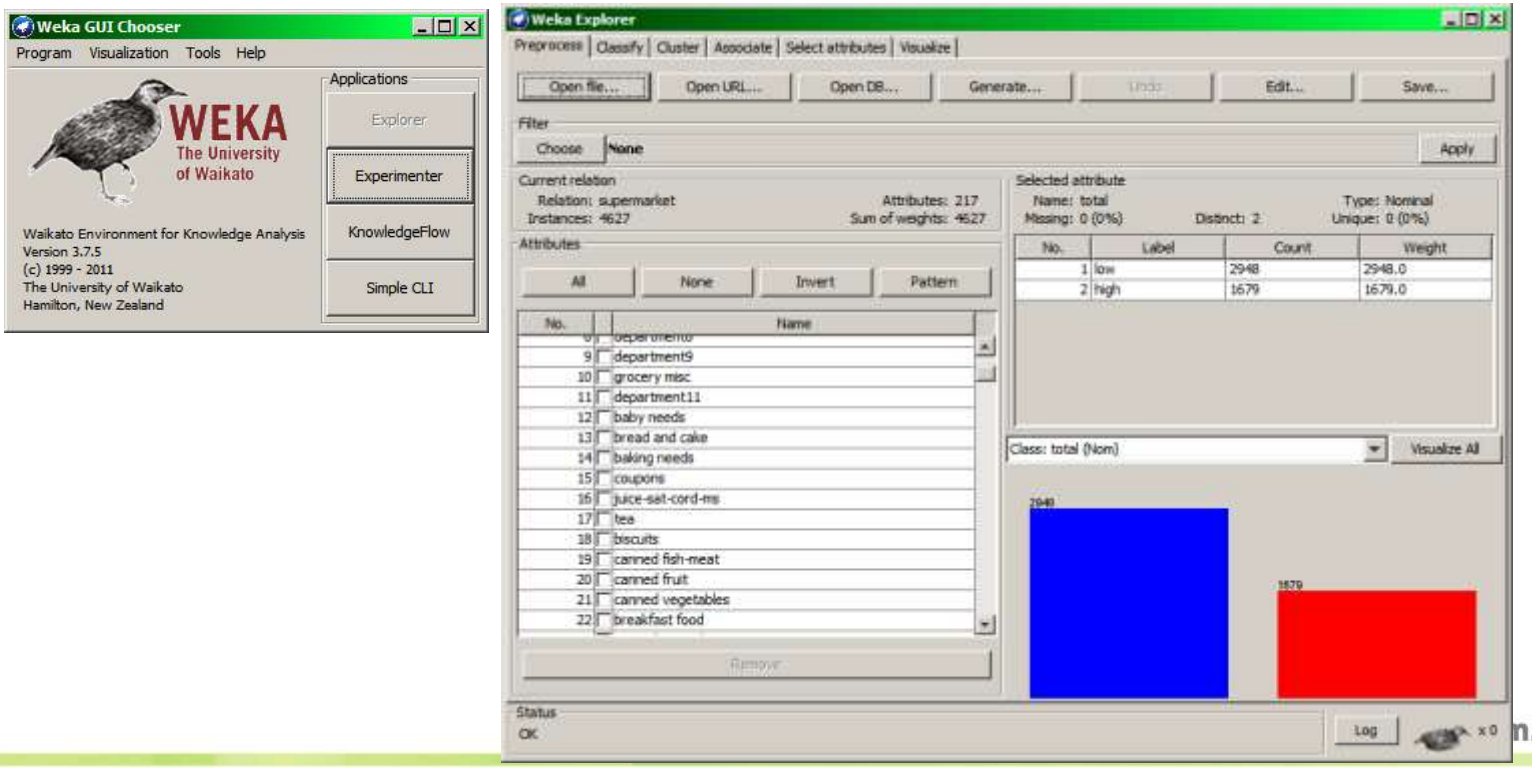
Obchod sleduje, které položky si zákazník koupil, a data o tom ukládá do datového skladu. V prvním kroku se spočítají četnosti jednotlivých položek koupených zákazníky. Musí být nastavena nějaká minimální hodnota počtu zakoupeného druhu zboží, aby se bral do úvahy.

Dále se hledají společné dvojice různých typů zboží. Libovolné dvojice přicházejí do úvahy, avšak z možných dvojic jsou vyřazovány ty, které nesplňují minimální četnost výskytu. Zbytek tvoří nalezené dvojice.

V dalším kroku se uvažují trojice. Vytvoří se všechny možné trojice a vyřadí ty, které nesplňují minimální limit výskytu. Může se tak pokračovat pro libovolné *n*-tice (jejich počet však velmi rychle klesá).

Softwarové nástroje pro dolování z dat

K dispozici jsou softwarové nástroje *open source*, např. **WEKA** implementovaná v jazyce Java:



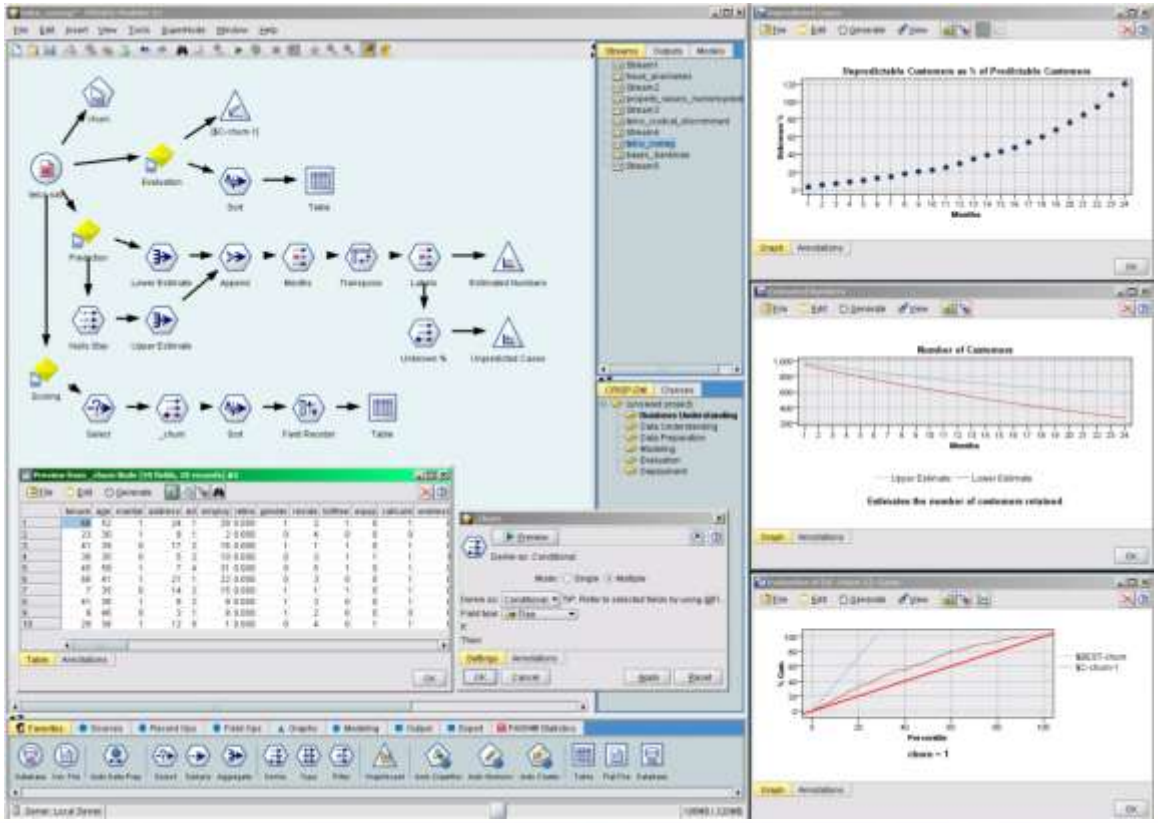
Softwarové nástroje pro dolování z dat

Další software, příbuzný s WEKA, je **RapidMiner.** Komerčních nástrojů pro dolování z dat existuje velká řada. K nejznámějším patří např. následující, které jsou velmi kvalitní:

SPSS (PASW Modeler; známý i pod názvem Clementine),
SAS (Enterprise Miner), StatSoft (Statistica Data Miner),
Salford (CART, MARS, TreeNet, RandomForest),
Angoss (KnowledgeSTUDIO, KnowledgeSeeker),
Microsoft SQL Server data Mining,
Oracle Data Mining,
Megaputer (PolyAnalyst).

Příklad komerčního nástroje na dolování z dat

PASW Modeler, známý i pod názvem Clementine:



VYSOKÁ
ŠKOLA
EKONOMIE
A MANAGEMENTU

Nárožní 2600/9a,158 00, PRAHA 5
tel. +420 841 133 166
info@vsem.cz

www.vsem.cz