

SLOVNÍČEK

TPS = Transaction Processing System, operativní řízení, krátkodobý horizont; **MIS** = Management Information System, taktické řízení, střednědobý; **EIS** = Executive Information System, strategické řízení a podpora vrcholového vedení, dlouhodobý

EDI = Electronic Data Interchange, el. komunikace s okolím podniku; **OIS** = Office Information System): část IS pro podporu kancelářských prací

ETL = Extract Transform Load (x ELT = Extract Load Transform)

DSS = Decision Support Systems

OLTP = Online Transaction Processing; **OLAP** = Online Analytical Processing

(ROLAP, MOLAP nebo HOLAP),

BPM = Business Performance Management; KPI = Key Performance Indicator; ERP = Enterprise Resource Planning

CIF, DIS model

SDA, DDA

DŮLEŽITÝ:

BI – podstata, proč využívat, 3V, ETL, DSA(?); architektura BI; kde je

ETL

Analytické systémy

Datová kostka – funkce, nákres, operace

Datový sklad, tržiště, rozdíly

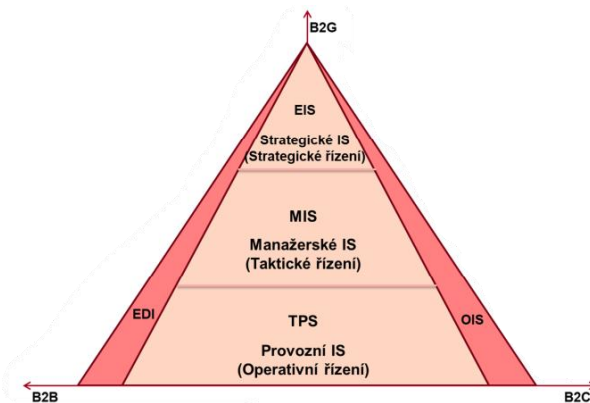
Tabulka dimenzí, fakt x atribut, význam

Hierarchie

Granularita

1 Úvod do problematiky manažerských informačních systémů

Typy IS podle úrovně řízení



- EDI (Electronic Data Interchange)
- OIS (Office Information System)

- **EIS** – data s vysokou agregací a strukturovaností, pořizována v TPS a MIS
 - Často prezentační vrstva pro BI
 - FCE: plánování, ekonomická analýza hospodaření firmy, výkaznictví, rozbor situace na trhu apod.
- **MIS** – evidenční a analytické práce
 - Pevné, pravidelně plánované reporty založené na datech z transakčních systémů
- **TPS** – podpora na operativní úrovni – provozní úroveň řízení, sledování transakcí (výrobních operací); OLTP – On-Line Transaction Processing

1.1 Agregace a sumarizace

Agregační operace

- Agregací funkce = každá funkce, která (na základě dané relace a jejího atributu) vrací **hodnotu**, která **závisí pouze na hodnotách** daného **atributu** v relaci
- Agregací operátory v SQL: AVG, COUNT, INTERSECT, MAX, MIN, OR, SUM, UNION, XOR, XUNION, ...
 - SELECT COUNT(*) FROM grades WHERE date='2022-01-30';

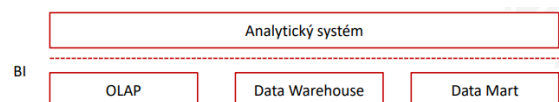
Agregace x Sumarizace

- Agregace – **jediná hodnota** stanovená ze všech hodnot atributu dané relace
- Sumarizace – **relace** obsahující hodnoty vypočtené z (částí) hodnot (některých) atributů v dané relaci.

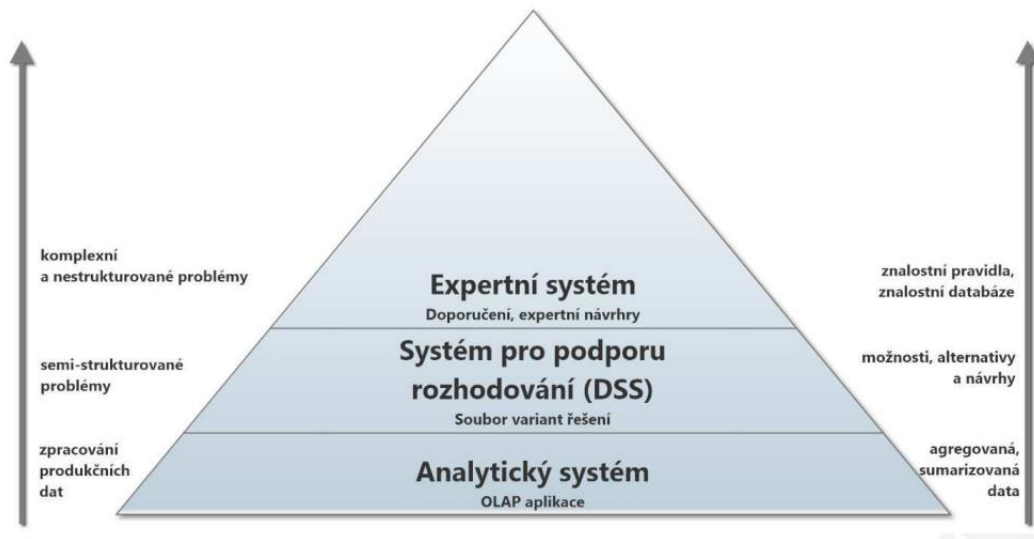
1.2 Analytické systémy

Definice

- Pro podporu analytického **zpracování dat**, **odhalování** skrytých **informací** (pochopení a předvídání **potřeb uživatelů**)
- Obecně **nejméně tři** vrstvy:
 - Vrstva pro **transformaci** dat (často ETL – Extraction-Transformation-Loading nástroje)
 - Vrstva pro **ukládání** dat (datové sklady, datová tržiště, operativní databáze)
 - Vrstva pro **analytické zpracování** dat
- Nejčastěji systémy založené na **konceptech Business Intelligence**, které pro svoji činnost využívají zejména **OLAP a datové sklady**



- **Koncept** analytického systému



1.3 Systémy pro podporu rozhodování

Definice

- DSS poskytují uživateli **varianty k rozhodování**
- **Interaktivní počítačový systém** skládající se ze tří hlavních komponent (UI, znalostní báze, databáze)
- Skryté vztahy v datech (*obr; nalezení funkce/predikce na základě předchozích hodnot*)

- Můžete navrhnout, jaká by měla být další hodnota y ?

x	y	x	y	$y = 2^x$
0	1	4	?	
1	2			
2	4			
3	8			

Strojové učení (Machine Learning)

- Schopnost systému **učit se** nebo **predikovat** stavy, které zajímají uživatele
- Využití dobře klasifikovaných dat a technik hlubokého učení (složitě umělé neuronové sítě)
- Metody dělíme na učení **s učitelem**, **bez učitele**, **posilující** učení
- Rozdíl mezi **strojovým učením** a **tradičním programováním**
 - Tradiční programování: Data + **Program** → (Počítač) → **Výstup**
 - Strojové učení: Data + **Výstup** → (Počítač) → **Program** (model)

- Kurz CZK/EUR

Dnes	Zítra	Dnes	Zítra
0,0408	0,0406	0,0411	?
0,0406	0,0407		
0,0407	0,0408		
0,0408	0,0407		

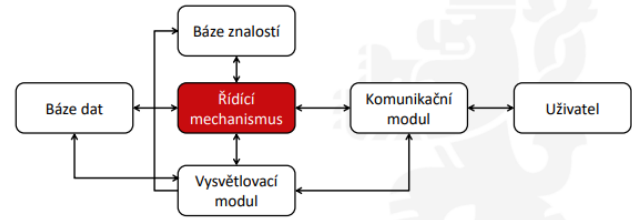
1.4 Expertní systémy

Definice

- Představují **znalostní báze kontrolních modelů**, které mohou být podmíněně rozděleny do dvou částí:
 - **Znamé informace** – z již existujících kontrolních modelů (subsystémů) = *předem vytvořený scénář* → *databáze výsledků predikcí situací*
 - **Expertní komponenta** využívá informace, modely a data expertního typu, založené na znalostech, zkušenostech a intuici odborníků

Schéma

- Části expertního systému:
 - Soubor **znalostí**, soubor **faktů** o aktuálním případě, inferenční (**řídící**) **mechanismus**, **vysvětlovací** modul, uživatelské **rozhraní**



Proces inference

- Postupně se hledá řešení:
 - IF Má peří AND Má zobák THEN Je to pták.
 - IF Pták AND IF Umí létat THEN Létající pták.
 - IF Létající pták AND IF Aktivní v noci THEN JE to sova

IBM Watson; virtuální asistenti, chatboti, ...

1.5 Business intelligence

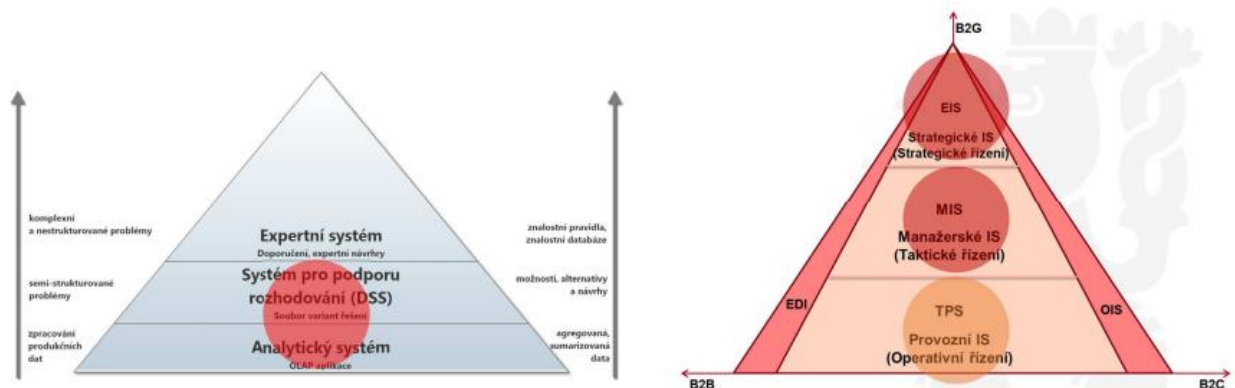
Podstata BI (proč BI)

- V současnosti se v organizacích nejčastěji uplatňují **analytické systémy** založené na **konceptech BI**.

→ **Podpora a zlepšování rozhodovacích procesů.**

- Nutnost se rozhodovat čím dál rychleji, pod větším tlakem, častěji a pracovat s větším množstvím faktů** = změny atributů rozhodování
- Změny **atributů rozhodování** a **tlaků na zlepšení efektivity** rozhodovacích procesů (dodání správných dat ve správný čas), vedou k **širokému uplatnění BI**, jako **nástroje pro zlepšení rozhodovacích procesů** v organizacích.
- Provozní (transakční) aplikace** jsou z hlediska analýzy a plánování **omezené** (nejsou dostatečně pružné, obtížný přístup k agregovaným datům, určeny primárně pro zaznamenávání; *mohou realizovat analytické operace, ale s omezenou rychlostí a flexibilitou*)
- Technické znalosti pro potřeby BI** – může být těžce implementovatelná (technické nároky na zabezpečení – náklady, technické znalosti pro efektivní využití)

Kde je Business Intelligence?



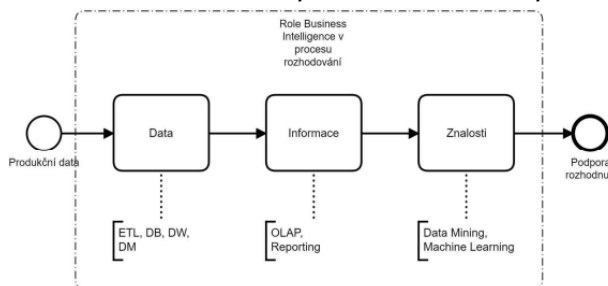
2 Principy Business Intelligence a jeho hlavní komponenty

Charakteristika Business Intelligence

- „**Sada konceptů a metod určených pro zkvalitnění rozhodnutí firmy**“. (H.J. Dresner, Gartner 1989)
- Business Intelligence je **sada procesů, aplikací a technologií**, jejichž **cílem** je účinně a účelně **podporovat rozhodovací procesy** ve firmě. Podporují **analytické a plánovací** činnosti podniků a organizací a jsou postaveny na **principech multidimenzionálních pohledů na data**.

Podstata BI (jak a proč BI)

- Nejčastějším výstupem BI jsou **reporty**
 - Kontingenční tabulky, grafy a **celá řada grafických a tabulkových výstupů** – srozumitelné pro uživatele, a na jejich výstupů (analýz) mohou **přijmout příslušná rozhodnutí**
 - (o změně marketingu, vyřazení z portfolia, změně segmentu zákazníků atd.)
 - Rozhodnutí z analytického zpracování **velkého množství** uložených faktů a historických dat
- **Hlavní důvod** využívání BI v organizacích = **podpora a zlepšování rozhodovacích procesů**
- Posláním BI **není** tvorba reportů nebo vlastní rozhodování, ale **podpora procesu rozhodování**.



Model role BI →

2.1 Architektura Business Intelligence

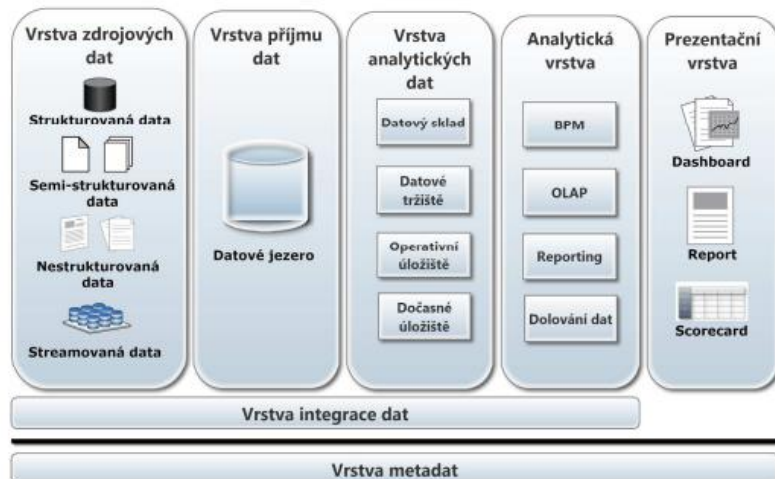
BI jako černá skříňka pro uživatele

- OLTP → BI → Reporty

Tradiční architektura

- Strukturovaná data → ETL → Datový sklad → OLAP → Reportovací nástroje

Hlavní komponenty BI a jejich vazby



1. Vrstva zdrojových dat

- **Strukturovaná** data = relační databáze (**referenční integrita**, **omezující podmínky**) z OLTP (ERP, CRM,)
- **Semi-strukturovaná** (polostrukturovaná) data = strukturovaná, **ale** neodpovídají formální struktuře relačních modelů dat; sémantický aparát **se může měnit**
 - *Webové stránky, soubory XML, JSON, CSV a další.*
- **Nestrukturovaná** data = **nemají** předdefinovaný **model** nebo nejsou organizována podle předem definovaného **schématu**; (rostou meziročně o 30–60 %)
- **Streamovaná** data = **mění se v čase**, vysoce **distribuovaná** a mají **volnou strukturu**
 - Datový proud = uspořádaná **kolekce datových položek**; položky jsou průběžně **generovány** a **odesílány** ke zpracování; **pořadí přijetí** nemůže být příjemci kontrolováno
 - *Data ze senzorů, z akciových trhů, monitorování stavu pacientů, satelitní data v dopravě, ...*

2. Vrstva příjmu dat

- **3V (Volume, Variety, Velocity)**
 - **Objem** – velké množství dat, která **vznikají** a neustále **rostou** v podnicích a organizacích
 - **Typ** – data přicházejí v **různých podobách** (strukturované DB až nestrukturované soubory)
 - **Rychlost** – obsah dat se **rychle mění** včetně rostoucí **potřeby jejich analýzy** v reálném čase
- **Datové jezero** (Data Lake) *Například Apache Hadoop*
 - Přizpůsobeno zejména pro **čtení dat**; sbírá a ukládá surová data z různých zdrojů, uložena v jednom repositáři
 - Data **uložena v původní struktuře** – definuje strukturu až když jsou využívána

3. Vrstva analytických dat

- **Datové sklady** (Data Warehouse) – **integrují a ukládají** podniková data;
 - navržena pro **optimalizaci výkonu** prováděných dotazů pro analytické úlohy
- **Datová tržiště** (Data Marts) – **podmnožina datového skladu** (jeho součást nebo nadstavba);
 - **subjektově orientovaná** analytická databáze (např. pouze data o prodeji v konkrétní zemi)
- **Operativní úložiště dat** (Operational Data Storage) – **podpůrná databáze** (očištěných, transformovaných) dat určených pro datový sklad
- **Dočasná úložiště dat** (Data Staging Areas) – pro dočasné uložení dat **před jejich vlastním zpracováním** do datových skladů nebo tržišť

4. Analytická vrstva

- **BPM** (Business Performance Management) – **řízení podnikové výkonnosti**; zaměření na sledování, měření a porovnávání **výkonnostních metrik** → **KPI** (Key Performance Indicator)
- **OLAP** (On-Line Analytical Processing) – **technologie** pro komplexní **dotazování** nad analytickou (OLAP) databází
- **Reporting** – komponenta **převádí data** (z OLAP) **na informace** prostřednictvím přehledů (reportů);
- **Dolování dat** (Data mining) – systémy zaměřené na sofistikovanou **analýzu velkého množství dat**

5. Prezentační vrstva

- **Zajišťuje prezentaci analytických pohledů na data pro koncové uživatele v BI**
 - samostatná BI platforma, nástroj integrovaný v EIS nebo systémech ERP, specializované nástroje, analytické aplikace

3 Analytické potřeby, analytický pohled na data a jejich příprava v ETL systémech

3.1 Vrstva integrace dat

Podstata integrace dat

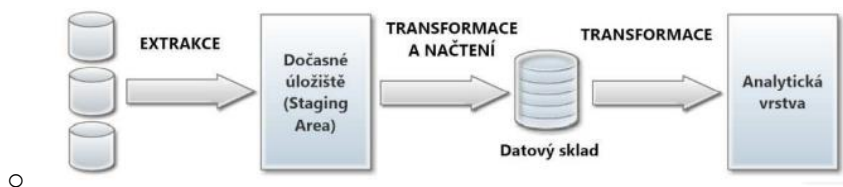
- Uživatelé BI potřebují **pracovat s historickými daty** (pro analýzu trendů), které **často v produkčních systémech chybí** (z důvodu výkonosti databázového systému jsou data často zálohována a odstraněna z provozní databáze).
- Velmi často jsou data **kódována pro potřeby produkčních systémů** – místo názvu země její číselný kód, místo atributu pohlaví použitý datový formát typu boolean – muž... 0, žena... 1,
 - Uživatelé BI **potřebují vidět srozumitelné hodnoty** („Hlavní město Praha“ a ne kód „19“)

Přístupy pro integraci dat v BI

- **ETL (Extract, Transform, Load)** – tradiční přístup
- **ELT (Extract, Load, Transform)** – typický pro integraci velkých dat

3.2 ETL procesy a jejich modelování

- ETL = **automatizovaný** proces pro **extrakci, transformaci a přenos** dat
- Periodicky **integruje data** ze strukturovaných zdrojů dat, která **transformují a reorganizují** do následujících komponent:
 - pomocných nebo dočasných **databází**,
 - datových **skladů**,
 - BI **modelů** (OLAP aplikací) s dimenzionální strukturou dat



Kroky návrhu procesu ETL

1. **Analýza požadavků** na datové pohledy
2. Vytvoření **referenčních dat**
3. **Extrakce** dat z různých zdrojů
4. **Validace** extrahovaných dat
5. **Transformace** dat
6. Uložení do **dočasného úložiště**
7. Výběr **cílového zdroje** a **mapování atributů**
8. Nahrání dat do **datového skladu**

(ad 8.) **Datová pumpa** = (semi)automatizované zpracování ETL procesu, „kopíruje“ data do dat. skladu; definuje přenos a transformaci dat ze zdrojového do cílového úložiště

Konceptuální model ETL

- Konceptuální modely procesů ETL jsou určeny pro **popis datových zdrojů a vztahů mezi nimi a datovým skladem**

- **Entity v modelu**

1. **Koncept** (Concepts),
2. **Atribut** (Attributes),
3. Vztah „**je součástí**“ (Part-of Relationships),
4. **Transformace** (Transformations),
5. Sekvence transformací (Serial Composition of Transformations),
6. Vztah „**poskytovatel**“ (Provider Relationships),
7. **Poznámky** (Notes),
8. **Omezení ETL** (ETL Constraints),
9. Kandidátní **vztahy** (Candidate Relationships)

Logický model ETL

- Pro dokumentaci podrobného **popisu pracovního postupu** a **činností** ETL
 - Popisují zejména **toky dat z datových zdrojů** směrem **k cíli**
- Návrh zejména těchto modelů:
 - **Workflow ETL** – definice sémantiky pracovního postupu ETL (sekvence toků dat mezi procesy – ze zdrojů k cíli)
 - **Procesy ETL** – návrh aktivit v ETL
 - **Scénář ETL** – plán provádění jednotlivých aktivit / zpracování a synchronizace činností / provádění scénáře jako celku / monitorování a protokolování stavu provedených činností

Komponenty ETL

- Nástroje pro **efektivní řízení workflow** = Konfigurace zasílání **chybových zpráv**, Trendové křivky o **spuštěných balíčcích** ETL procesů, Nastavení **pořadí spouštění** ETL procesů

4 Datový sklad a datové tržiště

4.1 Datové sklady

Definice

- = databázový systém, jehož účelem je **podporovat rozhodovací procesy** v organizaci prostřednictvím **integrovaného a konsolidovaného pohledu** na data pocházející z **různých datových zdrojů**.
- Vyvinuly se jako **součást BI řešení**, ve kterém datové sklady **integrují a transformují** podniková nebo organizační data **do centrálního úložiště**, které jsou **následně zpracovány** pro potřeby rozhodování.
- Představují **základní technologii**, která je v současnosti běžnou součástí **správy podnikových dat**.

Řešení datových skladů (DS)

- V současnosti **dva typy modelů dat** pro řešení DS:
 - **Relační** databáze a **Multidimenzionální** databáze
- Pro použití **relačních databází** jsou zejména následující **důvody**:
 - Jsou **velmi rozšířené** v podnikové a organizační praxi, **podporovány mnoha dodavateli** systémů, **snadno** se udržují a spravují, využívají **stejných dotazovacích prostředků**
- Pro použití **multidimenzionálních databází** jsou následující **důvody**:
 - Organizace uchovává svá data v **nerelačních databázích** (tzv. **NoSQL**)
 - klíčové-hodnotové, dokumentové, sloupcové, grafové a vícemodelové

Koncept datových skladů podle Inmona (1991)

- DS je **předmětově orientovaný**, **integrovaný**, **neměnný** a **časově rozlišený** souhrn dat pro podporu **rozhodování**. (= vlastnosti DS!)

Předmětově (subjektově) orientovaný DS

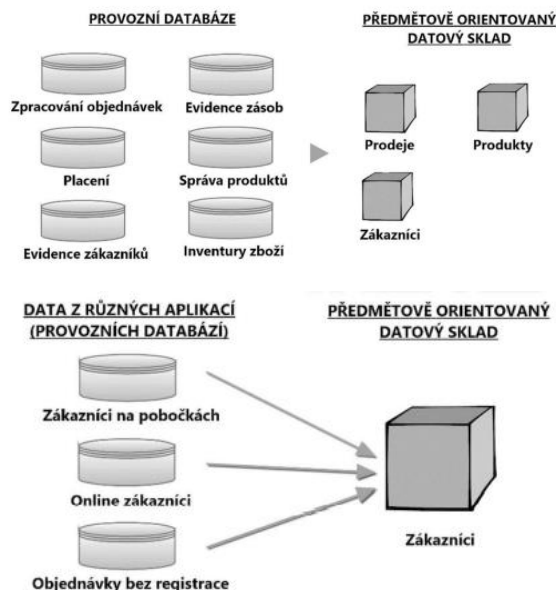
- Data organizována kolem **významných předmětů zájmu** organizace (nikoliv jako záznamy jednotlivých transakcí v TPS)
- Hlavní **předměty zájmu** např: **zákazníci**, **produkty**, **prodeje** nebo **prodejci**, ...

Integrovaný datový sklad

- Data v datovém skladu jsou integrována **ze všech provozních systémů, databází**, z některých **metadat** a **dalších** souvisejících externích dat.
- Centralizované úložiště **všech podnikových dat** se **společnou sémantikou a formátem**.
- Integrovaný datový sklad z nesourodných datových zdrojů je **základem pro efektivní rozhodování**

Neměnný datový sklad (nízká proměnlivost)

- Data v DS **nejsou** po nahrání (obvykle) **měněna** a **nejsou odstraněna**; **nejsou** určena pro **provozní úlohy**.
- **Jakákoliv změna** dat, je **zaznamenána v podobě snímku** (možnost sledovat historii dat)



Časově rozlišený datový sklad (historizace)

- Data **provozních systémů** obsahují **aktuální hodnoty**. X DS obsahují data za **několik desetiletí**.
 - Možnost **analyzovat trendy, vzory, korelace, pravidla a výjimky** z **historického hlediska**.
- Data v DS jsou určena pro **analytické zpracování**; pro analýzu nákupního modelu segmentu zákazníků jsou potřeba **nejen data o aktuálních nákupech, ale také o minulých** nákupech

Důvody fyzického oddělení DS od provozních databází:

OLTP databáze jsou optimalizovány pro dobu odezvy transakcí a propustnost dat.	DS jsou subjektově orientovány a optimalizovány pro analýzu dat (UNF).
OLTP systémy se zaměřují na správu aktuálních dat (transakce o aktuálním prodeji, ceny kurzu Kč apod.), bez odkazu na historické údaje.	DS jsou uzpůsobeny pro ukládání historických dat , pocházejících z různých míst v organizaci.
Přístupový vzor pro OLTP se skládá z krátkých atomických transakcí (SQL dotazu nebo procedurálního kódu – databázových triggerů).	DS jsou primárně navrženy pouze pro čtení , dotazy mohou být složitější.

4.2 Datová tržiště

Datové tržiště (Data Mart) = **datový sklad malé velikosti**, který obsahuje podmnožinu podnikového datového skladu nebo omezený objem agregovaných údajů **pro konkrétní analytické potřeby** podnikatelské **jednotky**, spíše než pro potřeby celého podniku.

- Datové tržiště (DT) je **podmnožinou datového skladu**; zdroj dat „malých“ datových rozměrů **zaměřený na vybraný předmět** (subjekt) zájmu, *například na prodej, finance nebo marketing*

Rozdíly mezi datovým skladem a datovým tržištěm

Cílem DS je zaměřit se na potřeby celého podniku (např. zákazníci, produkty, prodeje, nákup, zaměstnanci, pobočky).	Cílem DT je zaměřit se na potřeby podnikatelské jednotky (např. pouze na zákaznická data).
Granularita (míra podrobnosti) dat v DS je na transakční (atomické) úrovni.	DT je obvykle mírně agregované pro efektivní analýzu obchodní jednotky.
Pokrytí DS je plně historické – pro potřeby celého podniku.	DT je omezené na specifické potřeby oddělení .
Data do DS jsou pořizována z provozních sys. /z DT .	Data do DT jsou pořizována z DS/z provozních sys.

Kimball vs. Inmon

- Datový sklad jako **množina datových tržišť** (Kimball) – Návrh **zdola-nahoru**
- **Integrovaný** datový sklad (Inmon) – Návrh **shora-dolů**

Corporate Information Factory – CIF (Inmon)

- CIF = **Informační infrastruktura organizace** (Podniková informační továrna)
- Podnikový DS, který sleduje **architekturu datových toků** na nejvyšší úrovni.

5 Multidimenzionální databáze

Charakteristika

- **Multidimenzionální databáze** (MD) jsou často součástí **systémů datových skladů**, jejich účelem je ukládat data s **multidimenzionální strukturou**.
- MD ukládají **velké množství** dat **analytického typu** (v multidimenzionální struktuře), nad kterými jsou nejčastěji prováděny **analýzy a přehledy** prostřednictvím technologie **OLAP**.
- Pojem **multidimenzionální data** (multidimensional data) představuje **data souhrnných ukazatelů** vytvořených **různým seskupením relačních dat** a určených pro **on-line analytické zpracování** dat. (OLAP)

Charakteristické vlastnosti

- Reprezentují **multidimenzionální pohled** na data.
- Umožňují uživatelům klást **analytické dotazy** a provádět **náročné výpočty** nad daty.
- Ukládají **časovou dimenzi** pro tzv. **časové zpravodajství** (Time Intelligence).
- Vyznačují se navrhováním prostřednictvím **dimenzionálního modelování** (modelují se data dvou typů, tzv. dimenze a fakta) jiným **způsobem dotazování**.
- Svou strukturou databáze **podporují OLAP operace** (například operace roll-up a drill-down, které budou vysvětleny později)

Datová kostka

- **Data**, která jsou seskupena nebo kombinovaná v **multidimenzionálních maticích** označujeme jako **datová kostka** (data cube nebo pouze cube).
 - = **datová struktura** pro ukládání a analýzu **velkého množství vícerozměrných dat**.
 - = **základní konceptuální strukturu a techniku** k **popsání multidimenzionální databáze**, obdobně jako termín relace pro relační databáze.
- **Kolekce souvisejících datových kostek** se označuje jako **multidimenzionální databáze**
- Datová kostka se skládá **z dimenzí a měř**:
 - **Dimenze** (dimension) je hierarchicky uspořádaný soubor **rozměrových hodnot**, které poskytují **kategorické informace** charakterizující určitý **aspekt dat**.
 - **Míry** (neboli sledované ukazatele, measure) jsou především **kvantitativní údaje**, které se agregují **podle dimenzí**

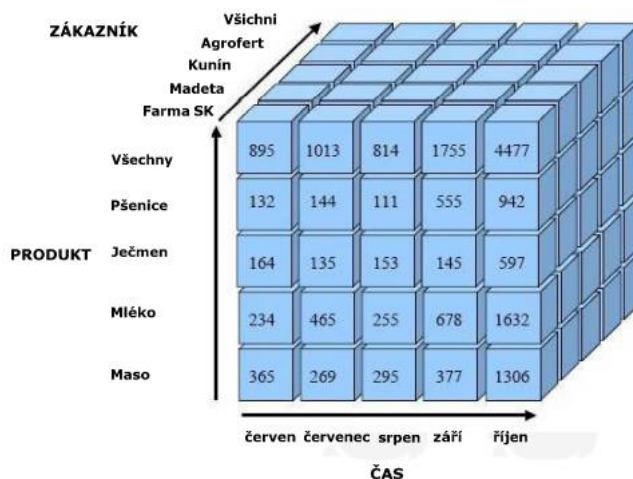
Multidimenzionální model dat

- Datová kostka je **základní konstrukt** MD a slouží jako **základní jednotka vstupu a výstupu** pro všechny operátory definované na MD.
- Datová kostka **je tvořena čtveřicí** $\langle D, M, A, f \rangle$, kde tyto složky indikují vlastnosti datové kostky následovně:
 - **D** = **množina dimenzí**; **M** = **množina měř** (*D a M jsou disjunktní*); **A** = **množina atributů**,
 - Zobrazení $f: D \rightarrow A$, existuje pro každou dimenzi a množinu atributů. Zobrazení je takové, že množiny atributů příslušející k dimenzi jsou v páru disjunktní,
- Pro **připisování hodnot** měř v rámci všech dimenzí je nutné definovat šestici $\langle D, M, A, f, V, g \rangle$
 - **V** představuje **množinu hodnot**, které byly použity k materializaci datové kostky
 - **g** reprezentuje **zobrazení** = ukazuje, **které hodnoty** jsou spojeny s **konkrétní „buňkou“**.

- Takové naplnění kostky se nazývá **instance kostky**
- Model datové kostky zahrnuje pojmy **datová kostka**, **dimenze**, **atributy dimenzí**, **atributy měř** a **hierarchie**.
- Multidimenzionální model se od relačního modelu liší tím, že používá **dva druhy atributů: dimenzionální atributy a míry**.
- Model datové kostky se zaměřuje na **symetrické zacházení** s **dimenzemi a mírami**.
 - To znamená, že **ostře rozlišuje** mezi dimenzemi a mírami.
 - Atribut označený jako míra **nelze použít** jako dimenzionální atribut a naopak.

Grafická ilustrace datové kostky

- **Dimenze** reprezentují **osy datové kostky**. Na obrázku má datová kostka dimenze PRODUKT, ZÁKAZNÍK a ČAS.
- Každá dimenze může mít **různou úroveň granularity**, například dimenze ČAS má granularitu den-měsíc-rok.
 - **Granularita** = PŘESNÝ POPIS ZAMĚŘENÍ MODELU; *míra podrobnosti, úroveň detailů, jemnost členění*
- Hodnoty, které jsou uvedeny v jednotlivých **buňkách** datové kostky, jsou **hodnoty vypočtených měř** (sledovaných ukazatelů).



6 Technologie OLAP

Charakteristika

- Výpočetní technologie pro **multidimenzionální analýzu dat** umožňující koncovým uživatelům **provádět složité výpočty** pro podporu plánování, prognózování, rozpočtování a finančního výkaznictví.
- OLAP umožňuje provádět **dotazy nad daty** a **analyzovat** je z **různých úhlů pohledu**.
- Ukládá data do **multidimenzionální databáze** (resp. do OLAP databáze), která nabízí snadné možnosti **prohlížení a navigace** dat, čímž umožňuje analytikům velmi **rychle vytvářet komplexní analytické pohledy**.
- OLAP je **stěžejní komponentou BI**
- Termín OLAP zavedli v roce 1993 Codd, E.F., Codd, S.B. & Salley, C.

Sestavili **12 pravidel pro definování OLAP**:

- Pravidlo 1: Multidimenzionalita
- Pravidlo 2: Transparentnost.
- Pravidlo 3: Dostupnost.
- Pravidlo 4: Konzistentní výkon reportování.
- Pravidlo 5: Architektura klient-server.
- Pravidlo 6: Generická dimenzionalita.
- Pravidlo 7: Dynamické řízení řádkových matic.
- Pravidlo 8: Podpora pro více uživatelů.
- Pravidlo 9: Neomezené operace napříč dimenzemi.
- Pravidlo 10: Intuitivní ovládání.
- Pravidlo 11: Flexibilní reporting.
- Pravidlo 12: Neomezený počet dimenzí a úrovně v hierarchiích.

Principy OLAP

- Pro usnadnění provádění dotazů OLAP a agregace dat byla v roce 1999 v SQL (v ISO/IEC 9075-2:1999) zavedena **tři rozšíření klausule GROUP BY**: operátor **CUBE**, **ROLLUP** a **GROUPING SETS**.

Příklad OLAP

- Pro názornost nebudeme nyní rozlišovat mezi tabulkou faktů a dimenzemi.

PRODUKT	KVARTAL	REGION	PRODEJE
A	Q1	CZ	20
A	Q1	SK	10
A	Q2	CZ	30
A	Q2	SK	60
A	Q3	SK	20
A	Q4	CZ	10
A	Q4	SK	40
B	Q1	CZ	30
B	Q1	SK	50
B	Q2	CZ	20
B	Q2	SK	60
B	Q3	SK	10
B	Q4	CZ	30
B	Q4	SK	40

```
SELECT KVARTAL,
REGION, SUM(PRODEJE)
FROM PRODEJE
GROUP BY CUBE
(KVARTAL, REGION)
```

KVARTAL	REGION	SUM (PRODEJE)
NULL	NULL	430
NULL	CZ	140
NULL	SK	290
Q1	NULL	110
Q1	CZ	50
Q1	SK	60
Q2	NULL	170
Q2	CZ	50
Q2	SK	120
Q3	NULL	30
Q3	SK	30
Q4	NULL	120
Q4	CZ	40
Q4	SK	80

•

6.1 OLAP operace

Charakteristika

- **Operace nad datovou kostkou** podobně jako v relačních databázích, kde bychom se pro práci s relacemi neobešli bez operací **projekce** a **selekce**.
- **Nejpoužívanější** OLAP operace:
 - **Roll-up** (konsolidace); **Drill-down**; **Slicing** and **dicing** (selekce a řez)

Hierarchie

- Hierarchie umožňují vidět data na různé **úrovni (levels) detailu**.
- **Logická stromová struktura**, která uspořádá členy dimenze tak, aby každý člen měl **jeden nadřazený člen** a **nula nebo více podřízených členů**.
- V hierarchii je možné data uspořádat do nižších a vyšších úrovní podrobností

Roll-up

- Operace roll-up **seskupí buňky** pro vybrané dimenze v datové kostce **na základě vzestupné agregace** v hierarchii dimenze, nebo prostřednictvím **redukce dimenzí**, aby bylo možné získat **hodnoty ukazatelů s méně podrobnou granularitou**.
- Např. původní hierarchie pro dimenzi PRODUKT je *Typ < Výrobce < Kategorie*.
 - Aby uživatel viděl širší perspektivu podle hierarchie dimenze, musí **seskupit sloupce a sjednotit hodnoty**; Roll-up zahrnuje **shrnutí informací** o dimenzi z *Typu* až po *Kategorii*.
- Při spuštění roll-up může být **odstraněna jedna nebo více dimenzí** z výsledné datové kostky.
- Pro roll-up operaci je potřeba **vypočítat ukazatele z nejnižší úrovně** dat.
 - = vliv na **výkon** dotazování; výpočty se provádí online nebo předem.

Drill-down

- Operace drill-down → **nižší agregační úroveň** (vyšší úroveň detailu) a je **opakem operace roll-up**.
- Operace **seskupí buňky** v instanci datové kostky prostřednictvím **sestupné agregace** v hierarchii příslušných dimenzí, nebo **přidáním nové dimenze**.
- Umožňuje koncovému uživateli zobrazit **vysoce detailní data** z méně detailního pohledu

Slicing

- Operace slice (neboli řez datovou kostkou) reprezentuje **dvojměrný pohled** na datovou kostku.
- Slice vybere jednoho konkrétního **člena** (member; položka v hierarchii) z dimenze z **instance** datové kostky a poskytne **náhradní podmnožinu** datové kostky na základě **stanovené podmínky** pro dimenzi.
- *Například produkt má několik kategorií (typy produktu), slice zobrazí data k vybranému konkrétnímu členovi (konkrétní typ)*

Dicing

- Operace dicing vybere **dvě nebo více dimenzí** z instance datové kostky a poskytne **novou datovou kostku** jako podmnožinu původní instance → umožní vybrat **konkrétní hodnoty více dimenzí**. = dicing
- Operace nad datovou kostkou, která **zahrnuje tři dimenze**, vytvoří novou podmnožinu datové kostky, která je dána následujícími kritérii:
- *Např. (kategorie, čas, kontinent) → nová kostka z (LCD, 2010, EU) → dále dělená na (výrobce, Q, země)*

7 Dimenzionální modelování

Úrovně v dimenzích

- **Datové kostky narozdíl od relací poskytují skutečnou multidimenzionalitu.**
- Rozšíříme příklad z předchozí přednášky o **úrovně v dimenzích**:
 - **Úroveň** = **pozice v hierarchii** a opačně, **hierarchie** = **způsob uspořádání dat podle úrovní**.
 - Každá **úroveň nad** nejnižší úrovní představuje souhrnný **součet úrovní pod ní**.
 - **Atributy** úrovně poskytují **doplňující informace** o členech dimenze na konkrétní úrovni hierarchie dimenzí: *např. v instanci dimenze ČAS na první úrovni tyden_id, můžeme mít atributy úrovně tyden_popis, tyden_datum-konec, tyden_rozpeti*).

Charakteristika

- Modelování pomocí **relačních modelů** je pro multidimenzionální návrh nevhodné – **složité struktury**.
- **Cílem** dimenzionálního modelování je **vytvořit datové** (multidimenzionální) **struktury**, které uživatel **snadno pochopí** a zároveň budou **kompatibilní s principy OLAP**
- Dimenzionální modelování je **proces modelování dat** pomocí **modelovacích konstrukcí** zajišťující **vícerozměrný** (multidimenzionální) model dat.
- MD modely **kategorizují data**, buď jako
 - **fakta** asociovaná s **numerickou mírou** (ukazatelem),
 - nebo jako **dimenze**, která charakterizují fakta a jsou **většinou textová**.
- **Dimenze** se používají pro dva účely: **výběr** dat a **seskupení** dat na požadované úrovni podrobností.
- Každá dimenze je **organizovaná hierarchicky** podle **úrovní**

Proces multidimenzionálního modelování (Kimball)

1. **Výběr podnikového procesu nebo procesů pro model.**
 - (hlavní, které by měl pokrývat)
2. **Výběr granularity podnikového procesu.**
 - **Granularita** modelu je **přesný popis zaměření dimenzionálního modelu; podrobnost**
 - např. **jednotlivá položka** prodeje v daný okamžik (minuty), nebo **mírně agregovaná položka** prodejů produktů s podrobností na (dny až roky)
 - závisí na disponibilních datech a potřebách analýz
3. **Výběr dimenzí.**
 - (se všemi **úrovněmi** a **atributy** pro danou úroveň **granularity**)
 - Dimenze **popisují fakta** a jsou vyjadřována podstatným jménem (produkt, čas, zákazník atd.)
4. **Výběr měr (ukazatelů).**
 - Určíme míry, které se mají **vypočítat** pro každou **kombinaci hodnot** v dimenzích.
 - Míra se skládá z:
 - **numerické vlastnosti** faktu (například jednotkové náklady, tržby, prodejní cena, ...)
 - a **vzorce**, který reprezentuje většinou nějaká agregační funkce (např. průměr, suma, minimum apod.)

Fáze návrhu analytické databáze

- **Analýza požadavků** – všechny uživatelské požadavky na analytickou databázi.
- **Koncepтуální návrh** – vybírají se **koncepty** a **vazby** mezi nimi důležité pro navrhovaný model.

- **Logický návrh** – identifikuje **souvislosti mezi atributy** a identifikuje **typy vazeb**. V této fázi se zohledňují aspekty a požadavky na uložení MD dat (ROLAP, MOLAP nebo HOLAP), řeší se použití PK/FK.
- **Fyzický návrh** – schéma popsané dotazovacím jazykem MDX, SQL nebo jiným dotazovacím prostředkem.

Analýza požadavků na OLAP

- Tvoří se **seznam uživatelských požadavků** s **atributy**:
 - **název** požadavku, **popis** požadavku, případně jeho **funkcionality**, **aktéři**, **složitost** požadavku, **priorita** požadavku a případná **návaznost** požadavku na ostatní požadavky.
- **Metody analýzy požadavků** dělíme na:
 - **Supply-driven přístup** (SDA): Známy jako **data-driven**, vychází z podrobné **analýzy datových zdrojů** pro stanovení MD konceptů v procesu reengineeringu.
 - **Demand-driven** (DDA): Známy jako **requirement-driven** nebo **goal-driven**, se zaměřuje na stanovení **uživatelských OLAP požadavků** a až pak je mapován do datových zdrojů.
 - **Hybridní přístupy**: Kombinuje **oba vzory** pro návrh OLAP databází. V rámci tohoto přístupu se provádí analýza **disponibilních datových zdrojů** v relaci s **analýzou požadavků** uživatelů OLAP

GrHyMM framework (Graph-Oriented Hybrid Multidimensional Model)

- Smyslem je vytvořit **schémata faktů**, přičemž současně zohlednit **potřeby uživatelů a zdrojů dat**.
- Vytváří se model **popisující vztahy** mezi hlavními **aktéry** organizace a jejich **vlastními zájmy**.
- V rámci i* frameworku jsou **požadavky uživatelů (obchodní cíle)** rozloženy do detailnější **hierarchie cílů**: 1. **strategické cíle**, 2. **rozhodovací cíle**, 3. **informační cíle**

Konceptuální multidimenzionální návrh

- Založené na konceptech **dimenzí, faktů, měř a hierarchie**.
- MD návrhy schémat **vyžadují speciální techniky**, které jsou založeny na:
 - **denormalizovaném** schématu **hvězda** (tzv. star schéma)
 - a částečně (nebo zcela) **normalizovaném schématu sněhové vločky** (tzv. snowflake schéma).
- Nejpoužívanější řešení na konceptuální = star schéma, na logické úrovni = snowflake schéma

Dimenzionální model faktů (Dimensional Fact Model (DFM))

- **Grafický konceptuální model**, prvky pro **snadnou srozumitelnost** modelu
- Modelovanými základními pojmy jsou **fakta, míry, dimenze a hierarchie**.
- **Fakta** jsou graficky znázorněna rámečkem se dvěma částmi: horní pro **název faktu**; a spodní pro **název míry**.
- Konečné uzly představují **atributy a otcovské uzly** dimenze

MultiDimER model

- Založený na **ER modelu** – upraveno pro generalizované a nestriktní hierarchie.
- Hierarchie v tomto modelu používají notaci, která umožňuje jednoznačně rozlišit každý typ hierarchie.

UML

- Umožňuje **extenze v modelech**, aniž by byla narušena sémantika UML.

- K tomu se používají modelovací prvky vyjádřené jako metatřídy v metamodelu: **tagovaná hodnota** (Tagged Value), **omezení** (constraint) a **stereotyp** (stereotype).
- Zohledňuje hlavní vlastnosti MD modelování na konceptuální úrovni, jako jsou:
 - **vztahy** mezi fakty a dimenzemi,
 - **hierarchie** s více nebo alternativními cestami a nepřesné a úplné hierarchie,
 - **degenerované dimenze** (dimenze, která nemá žádný obsah, kromě svého primárního klíče; je součástí tabulky faktů, bez vazby na tabulku dimenze).

Logický multidimenzionální návrh

- Vstupem do procesu logického návrhu může být kromě
 - a) **konceptuálního schéma**,
 - také b) **specifikace pracovní zátěže** databáze (workload) charakterizována mírami, například rychlost zpracování, index výkonu pro dobu odezvy nebo pro rychlost zpracování apod.,
 - nebo c) **plánovaný objem dat** (který záleží na počtu odlišných hodnot pro každý atribut, velikosti atributu a počtu událostí pro každý záznam faktu

Proces logického návrhu

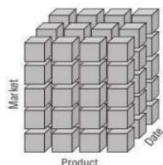
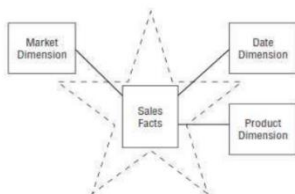
1. **Volba logického schématu** (schéma hvězda / sněhová vločka).
 2. **Převod konceptuálního schématu a mapování do logického modelu**. Rozlišujeme čtyři druhy mapování:
 - a) Mapování faktů a měr do datových kostek OLAP.
 - b) Mapování dimenzí a úrovní agregace.
 - c) Mapování agregačních hierarchií.
 - d) Mapování atributů dimenze k úrovním dimenze.
 3. **Volba materializovaných pohledů** – vytváří se pohledy bez zbytečných agregací, to znamená že se zahrnují pouze míry, které se agregují.
 4. **Optimalizace** – proces integrace schématu s databázovou technologií.
 - Ve fázi optimalizace se zohledňují požadavky na konkrétní typ databázové technologie (relační, multidimenzionální apod.), navrhují se tabulky a vztahy mezi nimi, specifikují se primární a cizí klíče.
- Pro účely tvorby logického modelu MD databáze se používá jeden ze tří základních typů návrhu schématu:
 - schéma **hvězda**,
 - schéma **sněhová vločka**
 - a schéma **konstelace**.
 - Princip modelování podle těchto modelovacích přístupů je založen na rozlišení dvou základních tabulek:
 - **buňky datové kostky**, tj. jednotlivé **ukazatele** a jejich **souřadnice** reprezentuje **tabulka faktů**;
 - **dimenze** a jejich **atributy** reprezentují **tabulky dimenzí**

8 Techniky dimenzionálního modelování

Logický multidimenzionální návrh

- Dimenzionální modely implementované v **ROLAP** se označují jako **hvězdicová schémata**.
- Dimenzionální modely implementované v **MOLAP** se označují jako **datová kostka**.
- Mají **společný logický design** s rozpoznatelnými dimenzemi; **fyzické provedení** se však liší.

Schéma hvězda versus OLAP kostka



- Přestože se možnosti OLAP neustále zlepšují, je **obecně doporučováno** (Kimball), aby byly **podrobné atomické informace načteny do hvězdicového schématu**; z něj jsou pak OLAP kostky **naplněny**
 - Většina technik dimenzionálního modelování je **formulována jako relační hvězdicové schéma**

8.1 Tabulka faktů

Charakteristika:

- Tabulka faktů (faktová tabulka) v dimenzionálním modelu **ukládá míry** (ukazatele výkonu) vyplývající z událostí podnikových procesů organizace.
 - Snažíme se ukládat **naměřená data** vyplývající z obch. procesu v jednorozměrném modelu.
 - Naměřená data jsou v drtivé většině **největší sadou dat**, **neměla by** být **replikována na více místech** pro více organizačních jednotek v celém podniku.
- Pojem **FAKT** představuje **obchodní míru** (ukazatel); každý řádek odpovídá **události měření**.
 - *Údaje na každém řádku jsou na specifické úrovni podrobností (detailu), označované jako zrna (grain), např. jeden řádek na produkt prodaný při prodejní transakci.*
- Tabulka faktů by měla zahrnovat **pouze skutečnou aktivitu** (pokud se neuskuteční, není vložen řádek)
- Málokdy **textové** (většinou diskrétní hodnoty), neměl by se vkládat nadbytečný text – **znemožnění analýzy**; pokud není text jedinečný pro každý řádek, patří do **tabulky dimenzí**
- **Aditivita** – zásadní vlastnost faktů; málokdy jeden řádek, spíše se sčítají (agregují); (+semi/neaditivní)
- **FK** – **dva nebo více**; připojení k **tabulkám dimenzí**
- **PK** – vlastní klíč **složený** z podmnožiny cizích (**každá** tabulka se složeným klíčem je tabulka faktů)
- Transakční, periodické snímkové, akumulované (stavové) snímkové

8.2 Tabulka dimenzí

Charakteristika

- Nedílnou **součástí tabulky faktů** – textový **kontext** spojený s událostí měření obchodního procesu.
- Často **mnoho atributů**; méně řádků, mnoho sloupců (*naopak od tabulky faktů*)
- **PK** – každá dimenze jeden klíč; **atributy** – zdroj omezení, seskupení, popisů
- **Hierarchie** – TD často představují hierarchické vztahy (*produkt – značka – kategorie; pro každý řádek v dimenzi produktu značka a popis kategorie*); hierarchické popisné informace uloženy redundantně; tabulky obvykle denormalizované

FAKT NEBO DIMENZE – otázka, zda je sloupec měřením, které nabývá mnoha hodnot a účastní se výpočtů (**skutečnost/fakt**), nebo jde o diskrétní popis, který je víceméně konstantní a podílí se na omezeních a označení řádků (dimenzionální **atribut**).

9 Dimenzionální tabulky

- Pohyb **v datech TF** podle **hierarchie** v dimenzi = pohyb po **různé úrovni detailu** dat (dril down/up)
 - **Alternativní struktury** (rok–měsíc–den, rok–kvartál–den apod.) = **více možností**, jak se pohybovat → Nejnižší úrovně (listy) – obsahují sloupce pro **klíče pro více nadřazených** struktur
- Umělé privátní klíče = slouží jako primární klíče v tabulkách dimenzí a jako cizí klíče v tabulkách faktů (složený PK)

Druhy dimenzí

- **Degenerované** = pouze primární klíč v tabulce faktů; nemusí existovat dimenzionální tabulka
- **Parent-child** = např. *zaměstnanec, výš manažer – odkazuje zpět na zaměstnance (manažera)*
- **Kauzální** = příčiny jevů; např. *promo akce = důvod zvýšení prodeje*
- Dimenze **času** = standardní součást, vždy definována; dimenze data/času dne;
- Dimenze **různých rolí** = když je potřeba jednu dimenzi použít ve vztahu k TF s různým kontextem

10 Implementace dimenzí

Změny v dimenzích

SCD (Slowly Changing Dimensions)

- **změny ve struktuře a prvcích** dimenzí (• přidání nových prvků (např. nové zboží, zákazníci apod.), • zrušení prvků (např. zrušení zboží), • změny hodnot atributů (názvy produktů, jména pracovníků, adresy zákazníků nebo dodavatelů apod.), • změny ve struktuře – zatřídění (zboží, pracovníků) do jiných skupin v rámci hierarchické struktury prvků dimenze)
- **Historie:**
 - 1) v rámci tabulky faktů = atribut se přesune z tabulky dimenze (číselníku) do tabulky faktů.
 - 2) historickými tabulkami dimenzí.

Typy změn

- 0 – **Ignorování** změn (nepromítají se do datového skladu)
- 1 – Přepis **hodnoty atributu** (Update Changes)
- 2 – Přidání **nové řádky** do dimenzionální tabulky (Keep Historical)
- 3 – Přidání **nového sloupce** do dimenzionální tabulky (Preserve Limited History);
 - = nový atribut pro starou hodnotu
- 4 – Vytvoření **nové historické dimenze** (pokud dochází k častým změnám)

Temporální (časové) tabulky

- Systémově **verzované** a **uchovávají historii** změn (vložení, odstranění, aktualizace)
- Vždy **dvě tabulky**: hlavní tabulka (aktuální záznamy) a tabulka historie (pouze předchozí stavy záznamů)