# Datový sklad

Tento článek pojednává o úložišti velkého množství počítačových dat. O prostoru pro skladování surovin a materiálu pojednává článek sklad.

**Datový sklad** (anglicky **Data Warehouse**, případně **DWH**) je zvláštní typ relační databáze, která umožňuje řešit úlohy zaměřené převážně na analytické dotazování nad rozsáhlými soubory dat.

## 1 Definice datového skladu

K definici rozdílu mezi „běžnou“ relační databází a datovým skladem se obvykle používá následujících charakteristik popsaných Williamem Inmonem:

**Orientace na subjekt** U běžné relační databáze je obvyklá snaha o co nejmenší redundanci uložení dat, které je dosahováno jejich normalizací do třetí normální formy a vnitřním provázáním jednotlivých logických funkčních celků. V datovém skladu je naproti tomu řešení vždy vedeno snahou o jasnou vnitřní separaci jednotlivých funkčních celků – výsledkem je struktura, která je čitelnější pro uživatele (manažera, business analytika) za cenu zvýšených nároků na paměťový prostor.

**Integrovanost** Běžná provozní aplikace (program) nad relační databází řeší určitý specifický okruh úloh nad „svými“ specifickými daty. V datovém skladu je třeba naproti tomu shromáždit informace z mnoha různých zdrojů a seskupit je nikoliv podle původu, ale podle logického významu (úzce souvisí s **orientací na subjekt** – všechna data týkající se určité funkční oblasti potřebuji mít „na jedné hromadě“ bez ohledu na to, odkud pocházejí).

**Nízká proměnlivost** Data jsou do datového skladu obvykle nahrávána ve větších dávkách (například v denních nebo týdenních intervalech) a pak již nejsou nijak modifikována.

**Historizace** Data jsou v datovém skladu obvykle udržována v historické podobě, nikoliv pouze v aktuálním stavu. To je dáno nutností provádění analýz zaměřených na vývoj v čase. V běžné relační databázi je z pohledu uživatelů obvykle zajímavý pouze aktuální stav datových objektů.

## 2 Technologické charakteristiky datového skladu

Z požadavků na **datový sklad** vyplývají jeho technologické charakteristiky:

- **Datový sklad** musí obsahovat nástroj pro nahrávání dat z různých datových zdrojů, tyto zdroje mohou mít různé datové formáty a různé fyzické umístění, nemusí se zdaleka jednat pouze o relační databáze.
- **Datový sklad** ukládá data nikoliv s ohledem na co nejlepší podmínky pro editaci, ale s ohledem na co nejlepší a nejrychlejší provádění složitých dotazů – proto je pro uložení dat používána často technologie OLAP.
- Nelze předem vědět, jaké dotazy a jaké úlohy budou chtít uživatelé nad **datovým skladem** v budoucnosti řešit. (V době budování **datového skladu** je obvykle známý pouze typ úloh, nikoliv všechny jednotlivé dotazy a úlohy.) Z toho vyplývá potřeba dostatečně flexibilních a přitom uživatelsky přívětivých analytických nástrojů.

## 3 Logická struktura datového skladu

Data v datovém skladu jsou z logického (uživatelského) pohledu členěna do **schémat** – každé schéma odpovídá jedné analyzované funkční oblasti.

Jádro každého schématu tvoří jedna nebo několik **faktových tabulek**. V nich jsou uložena vlastní analyzovaná data – číselné a finanční hodnoty, které jsou použity k analytickým výpočtům – agregacím, třídění apod. Většinu paměťového místa v datovém skladu zabírají faktové tabulky, které obsahují detailní údaje ze všech zdrojů – tedy řádově více údajů než ostatní tabulky.

Faktové tabulky jsou pomocí cizích klíčů spojeny s **dimenzemi**. Dimenze jsou tabulky, které obsahují seznamy hodnot sloužících ke kategorizaci a třídění dat ve faktových tabulkách.

### 3.1 Příklad

V datovém skladu je třeba uložit informace o všech prodejích z pokladen hypermarketů, data budou dále ana-

lyzována na základě doby prodeje, prodejny, typu zboží, dodavatele, probíhajících marketingových akcí a způsobu platby (kartou, hotově).

Schéma *Prodej* bude obsahovat faktovou tabulku *Položky prodeje*, kde bude pro každou prodanou položku uložen údaj o typu prodaného zboží, ceně a počtu kusů (případně prodané hmotnosti).

Kromě této faktové tabulky bude schéma obsahovat také dimenze pro třídění položek prodeje: časové dimenze *Datum* a *Hodina* (v rámci dne), dimenzi *Prodejna*, dimenzi *Typ zboží*, kde bude jeden řádek pro každou jednotlivou položku (například „Choceňský jogurt borůvkový 250ml“), dimenzi *Kategorie zboží* (obsahující řádky jako například „Jogurt“), dimenzi *Oddělení* (obsahující řádky jako například „Mléčné výrobky“), dimenzi *Dodavatel* (obsahující řádky jako například „Choceňská mlékárna a.s.“) a tak dále.

Faktová tabulka musí být pomocí cizího klíče přímo nebo nepřímo propojena s každou z těchto dimenzí.

## 3.2 Hvězda a vločka

V příkladu je vidět, že některé dimenze jsou jednoduché – například pro dimenzi *Prodejny* stačí navrhnout tabulku pro uložení údajů o prodejnách, na kterou bude napojena faktová tabulka.

Horší je situace v případě *Typ zboží* – ta se uvnitř sebe sama dále rozpadá – dimenze *Kategorie zboží* by vůbec nemusela být napojena na faktovou tabulku, ale na dimenzi *Typ zboží*, dimenze *Oddělení* na *Kategorii zboží*, dimenze *Dodavatel* na *Typ zboží*. V tomto případě mluvíme o **hierarchické dimenzi**.

V uložení hierarchických dimenzí mám v zásadě dvě možnosti:

1. Z celé hierarchie vytvořím jednu dimenzní tabulku, ve které budou údaje pro vyšší stupně hierarchie uloženy redundantně. Vznikne schema, kde je každá dimenzní tabulka vázána přímo na faktovou tabulku – podle tvaru svého diagramu se takové schéma nazývá **hvězda (Star schema)**.
2. Na hierarchickou dimenzi budu aplikovat normalizační doporučení 3NF, takže pouze dimenze na nejnižším stupni hierarchie bude vázána přímo na faktovou tabulku, ostatní pak na některou z nižších dimenzí v hierarchické struktuře – podle tvaru svého diagramu se takové schéma nazývá **vločka (Snowflake schema)**.

# 4 Zdroje textu a obrázků, přispěvatelé a licence

## 4.1 Text

- **Datový sklad** *Zdroj:* http://cs.wikipedia.org/wiki/Datov%C3%BD_sklad?oldid=12605205 *Přispěvatelé:* ToOb, Dinybot, Chrupoš, TXiKiBoT, MiroslavJosef, SieBot, Hubipe, ArthurBot, Luckas-bot, Xqbot, DixonDBot, EmausBot, FoxBot, Manubot, Addbot a Anonymové: 3

## 4.2 Obrázky

## 4.3 Licence obsahu