PV005 – Služby počítačových sítí: Data Warehouses

Jaroslav Bayer[1]

Fakulta informatiky Masarykova univerzita

27. 11. 2013

[1] CVT FI MU, B310, email: xbayer@fi.muni.cz

Obsah přednášky

Normální formy

Relační datový model – E. F. Codd, 1969

- 1NF
  - atributy obsahují pouze atomické hodnoty
    (nevyskytují se opakující se skupiny atributů)
- 2NF
  - 1NF + žádný neklíčový atribut není závislý na vlastní podmnožině nějakého KK
    (všechny neklíčové atributy jsou závislé na každém celém KK)
- 3NF
  - 2NF + všechny neklíčové atributy přímo (netranzitivně) závisí na každém KK
    (každý atribut tranzitivně závisející na klíči je klíčový atribut)
    (každý atribut je funkčně závislý na klíči a pouze na klíči)

Normální formy

- BCNF[1]
  - pro každou závislost $X \rightarrow Y$ platí, že buď $Y \subseteq X$ nebo X je SK
  - (každá netriviální závislost $X \rightarrow Y \Rightarrow X$ je nadmnožinou nějakého klíče nebo klíč)
  - BCNF $\Rightarrow$ 3NF (obráceně nikoli!)
- 4NF
  - 3NF + odstraněny podmíněné funkční závislosti (nevyskytují se entity, které nemají přiřazeny hodnoty některých atributů)
- 5NF
  - project-join normal form, relace nelze již bezeztrátově rozložit
- 6 NF
  - nesplňuje žádnou netriviální „join dependency“
- EKNF[2], DKNF[3], . . .

[1] Boyce-Codd Normal Form
[2] Elementary Key Normal Form
[3] Domain-key Normal Form

On-Line Transaction Processing (OLTP)

- silně normalizované databáze
- hlavním cílem je snížení redundance dat
- optimalizováno na
  - velké množství malých transakcí
  - transakce přenášející DB z konzistentního stavu do konzistentního stavu
  - kombinace čtení/zápis
  - snadné modifikace ve víceuživatelských prostředích
  - snižování redundance v datech
  - zajištění datové integrity
- prakticky nejrozšířenější přístup v relačním modelu
- tzv. *operační/produkční databáze*

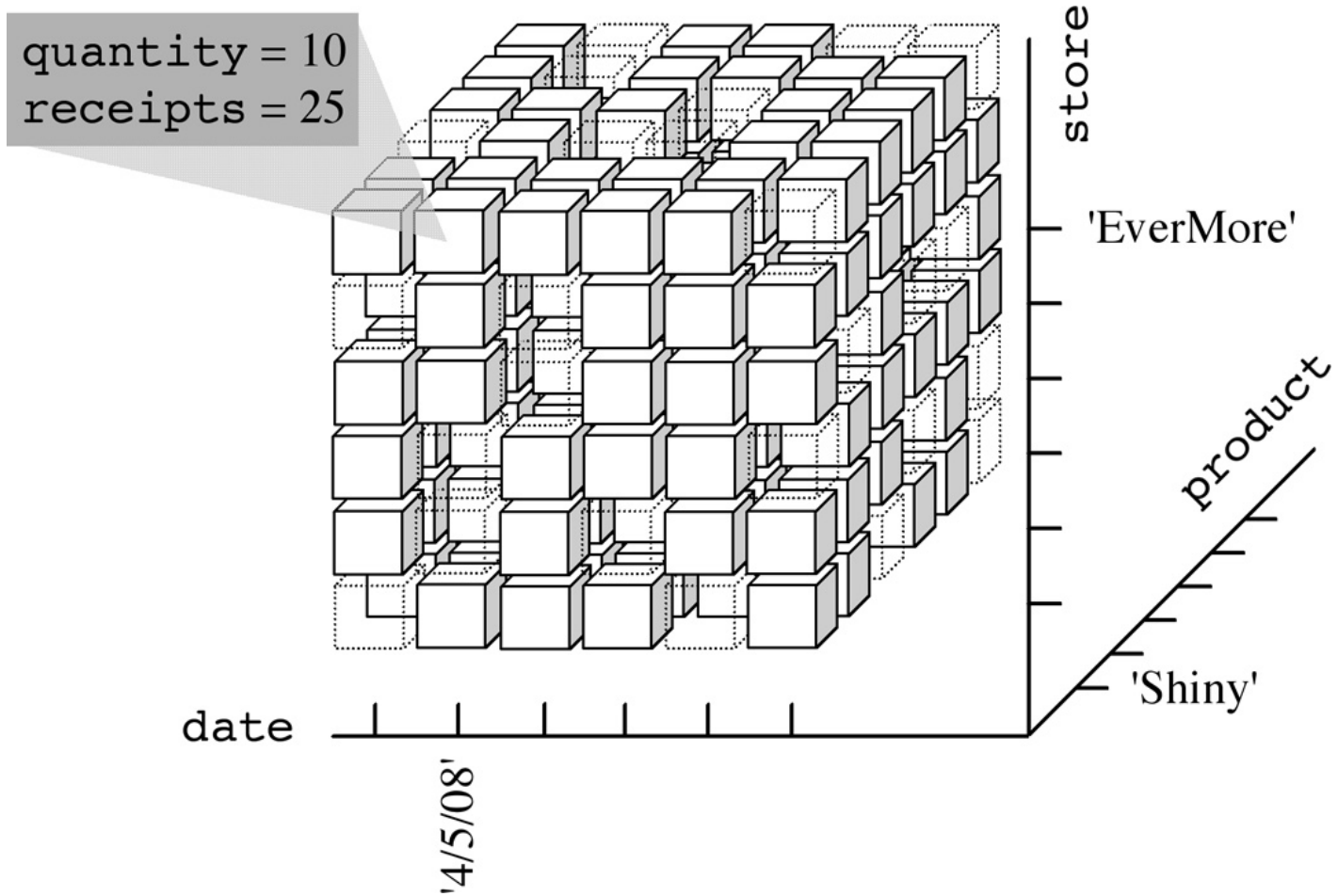
Denormalizace

- normalizované DB nevhodné pro analytické zpracování dat
  - dotazy často vyžadují přístup do velkého množství tabulek
  - zbytečně časově náročné *join* operace
- denormalizace
  - doplnění *redundantních dat*
  - předpočítání *agregovaných*, *seskupených* či *sumarizovaných* dat
  - odlišná DB schémata
  - *optimalizace pro čtení*
- materializovaný pohled (materialized view)
- schéma hvězdy (star) nebo vločky (snowflake)
- OLAP kostka
- ...

On-Line Analytical Processing (OLAP)

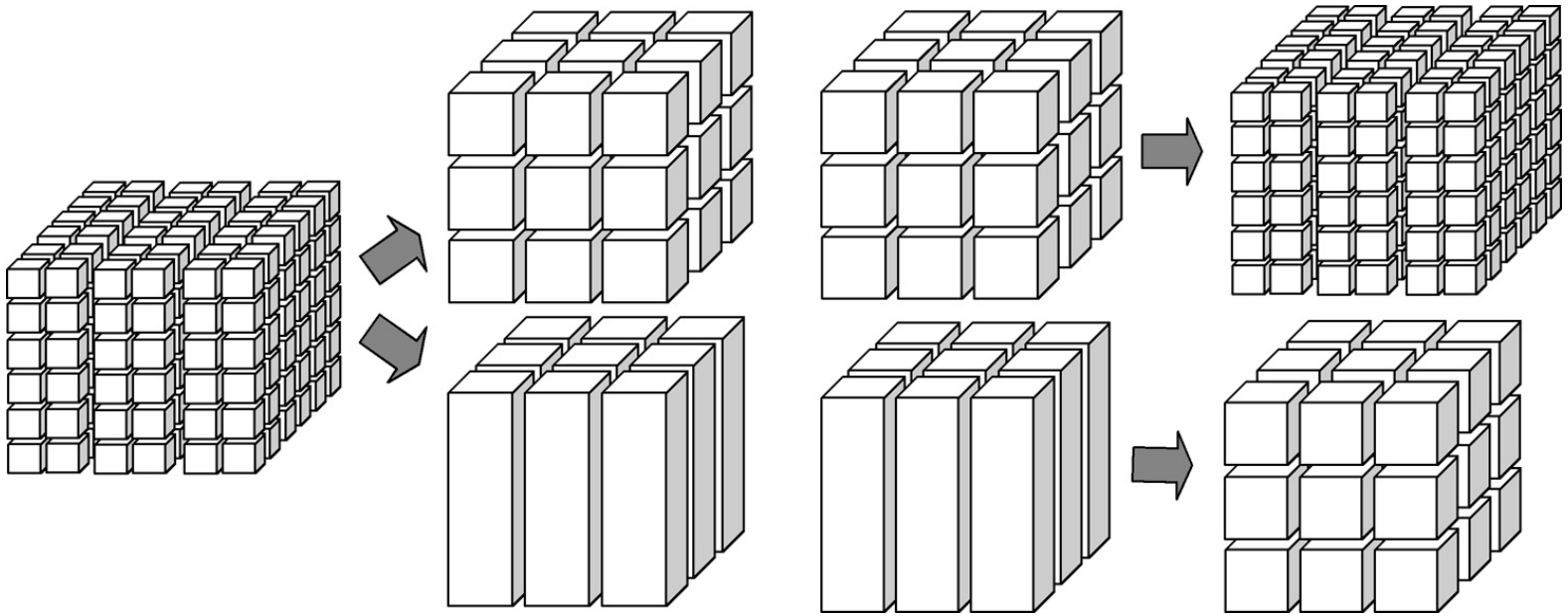
- mj. technologie ukládání dat v DB
- zaměřeno na
  - ukládání velkých objemů dat pro budoucí zpracování
  - podporu analytického zpracování dat
  - efektivní zpracování multi-dimenzionálních dotazů
  - čtení (read-mostly DB)
  - ukládání dat ve snadno pochopitelném formátu
  - ukládání historie dat
    - vedení, analytici, specialisté mimo IT oblasti apod.
- data většinou nahrávána periodicky
- málo uživatelů
- orientováno na subjekt
- pouze operace *insert* a *select*

OLAP kostka



Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

OLAP kostka, operace



Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

Data Warehouse: základní charakteristika

Data Warehouse: definice

- Data Warehouse (datový sklad) je:
  - kolekce dat pro podporu rozhodování s následujícími vlastnostmi:
    - orientovaný na subjekt,
    - integrovaný,
    - časově proměnný,
    - avšak stálý (konzistentní).
- definice dle Williama H. Inmona[4]
- data Warehousing je
  - kolekce metod, technik, nástrojů a přístupů k zajištění podpory pro *knowledge workers* při analýzách dat, které dopomohou k lepším rozhodnutím a zkvalitnění informačních zdrojů.

[4] The father of the data warehouse

DW: orientace na subjekt

- orientace na subjekty, kterými se podnik/organizace zabývá
  - zákazník, dodavatel, produkt
  - student, učitel, předmět
- zaměřuje se zejména na data vhodná pro strategická rozhodnutí
- jasné a čitelné oddělení funkčních celků
- vyšší paměťová náročnost

- DB pro OLTP se oproti tomu orientuje na transakce
  - faktura, vklad, půjčka, prodej
  - zápis, hodnocení, změna kreditace

DW: orientace na subjekt

- orientace na subjekty, kterými se podnik/organizace zabývá
  - zákazník, dodavatel, produkt
  - student, učitel, předmět
- zaměřuje se zejména na data vhodná pro strategická rozhodnutí
- jasné a čitelné oddělení funkčních celků
- vyšší paměťová náročnost

- DB pro OLTP se oproti tomu orientuje na transakce
  - faktura, vklad, půjčka, prodej
  - zápis, hodnocení, změna kreditace

DW: integrovanost

- integrace a sjednocení dat
  - více zdrojů dat (produkčních systémů[5])
  - sjednocení názvů, měřítek, jednotek, kódování, . . .
  - integrace dat do jednotné logické podoby

[5] též *operačních* či *transakčních systémů*

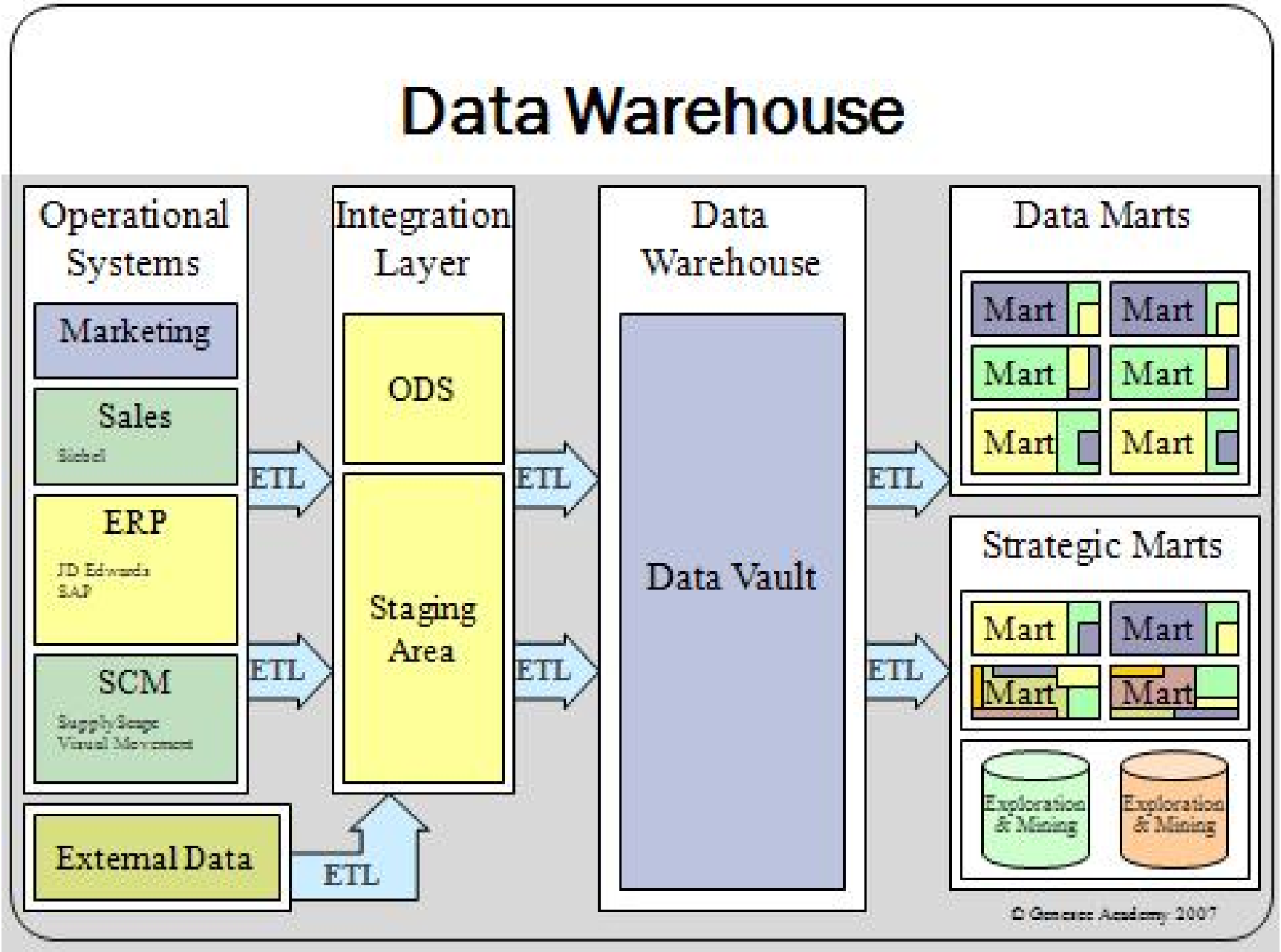
DW: časová proměnlivost

- data většinou nahrávána periodicky po větších dávkách
  - avšak existují i on-line aktualizované datové sklady
- data po vložení zafixována jako časový snímek produkční DB
- součástí datových záznamů jsou časové známky
- historie dat

DW: stálost (konzistence)

- uživatelé data zásadně nemění
  - pokládají zejména dotazy (select)
- data se po vložení prakticky nemění
  - až na výjimky v podobě chyb v datech či HW poruch
  - po exspiraci mohou být data vymazána

Data Warehouse: návrh



Zdroj: `http://upload.wikimedia.org/wikipedia/commons/4/46/Data_warehouse_overview.JPG`

DW: Staging Area a ODS

- (Data) Staging Area
  - mezilehlé datové úložiště
  - časově nestálé (data se po nahrání do DW mohou smazat)
  - sběr dat z více zdrojů
  - hledání rozdílů mezi aktuálními daty a daty v DW
  - předvýpočty agregovaných hodnot
  - čištění dat (data cleansing)
    - detekce a oprava porušených či nesprávných záznamů
    - nezaměňovat s pouhou validací dat
- Operational Data Store (ODS)
  - DB navržená pro integraci dat z různých zdrojů
  - data uložena s nejvyšší granularitou (atomická data)
  - data dostupná produkčnímu systému i DW
  - data omezena na aktuální stav (nebo stav jemu blízký)

DW: Staging Area a ODS

- (Data) Staging Area
  - mezilehlé datové úložiště
  - časově nestálé (data se po nahrání do DW mohou smazat)
  - sběr dat z více zdrojů
  - hledání rozdílů mezi aktuálními daty a daty v DW
  - předvýpočty agregovaných hodnot
  - čištění dat (data cleansing)
    - detekce a oprava porušených či nesprávných záznamů
    - nezaměňovat s pouhou validací dat
- Operational Data Store (ODS)
  - DB navržená pro integraci dat z různých zdrojů
  - data uložena s nejvyšší granularitou (atomická data)
  - data dostupná produkčnímu systému i DW
  - data omezena na aktuální stav (nebo stav jemu blízký)

DW: ETL

- Extract, Transform, Load (ETL)
  - Extract
    - získání dat z různých (a často nekompatibilních) zdrojů
    - analýza dat, kontrola souladu se vzory dat, . . .
  - Transform
    - transformace dat ze struktury zdroje do struktury cíle
    - výběr sloupců, změna kódování, spojení tabulek, agregace, disagregace, pivoting, validace dat, . . .
  - Load
    - nahrání dat do cíle, např. DW
    - triggery a ověření konzistence dat přes integritní omezení

DW: Data Marts

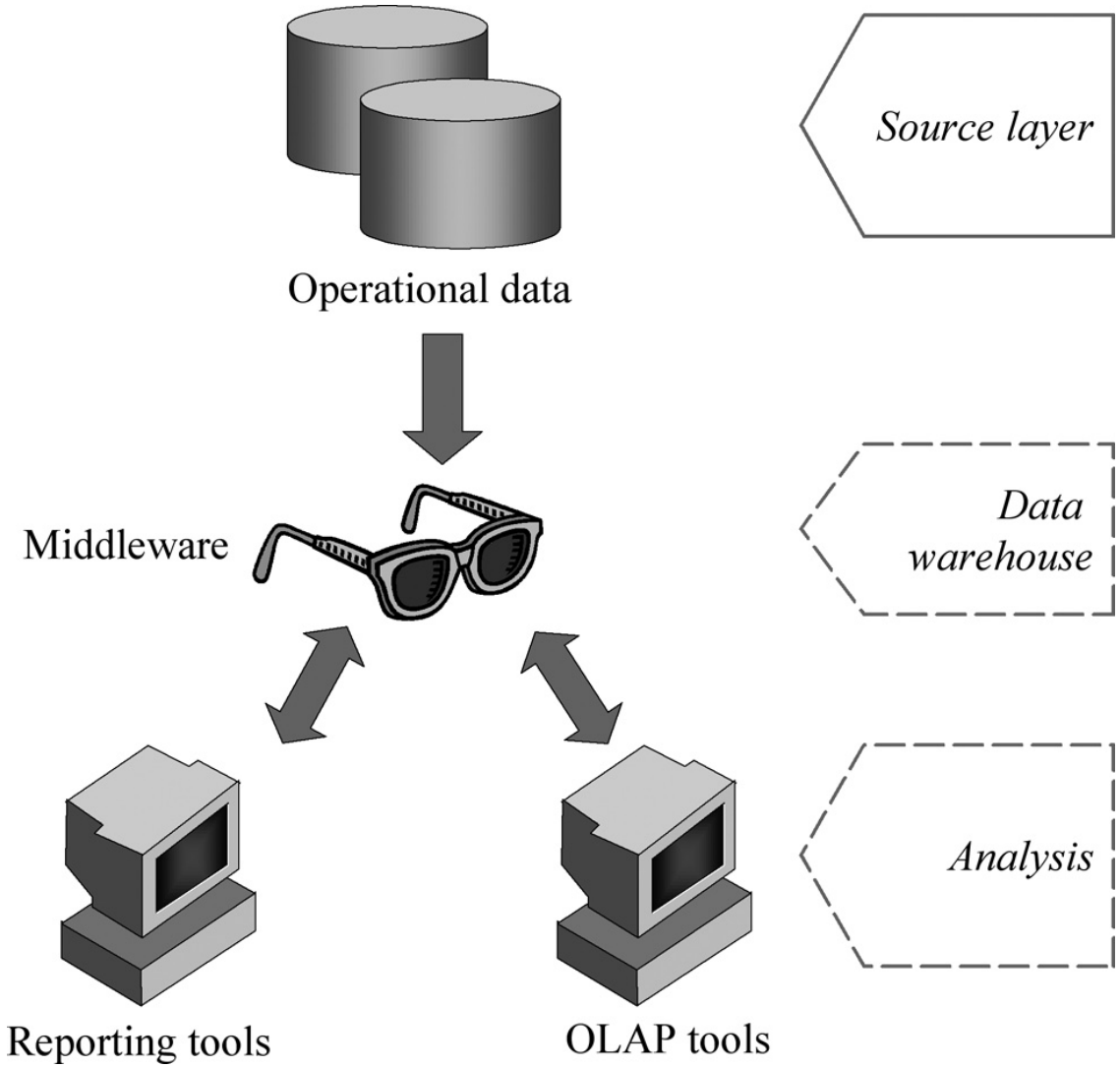
- Data Mart (DM, datová tržiště)
  - logická podčást DW
  - obsahuje podmnožinu dat z DW
  - zaměřen na konkrétní uživatele
  - přístupová vrstva pro získávání dat z DW
  - DM může mít vlastní HW, SW i data a DB
    - snížení doby přítupu
  - lepší definice uživatelů, bezpečnost

DW: Data Vault

- speciálně navržená DB
- Data Vault Modelling
  - databázová modelovací metoda
  - vyhovuje potřebám integrace i ukládání historie dat
  - podporuje sledování původu dat (data tracking)
  - zkracuje čas potřebný pro naplnění (loading time)
  - reaguje dobře na změny
  - splňuje požadavek 100 % dat po 100 % času
- může nahradit ODS
- (detaily později)

Data Warehouse: architektura

DW: jednovrstvá architektura

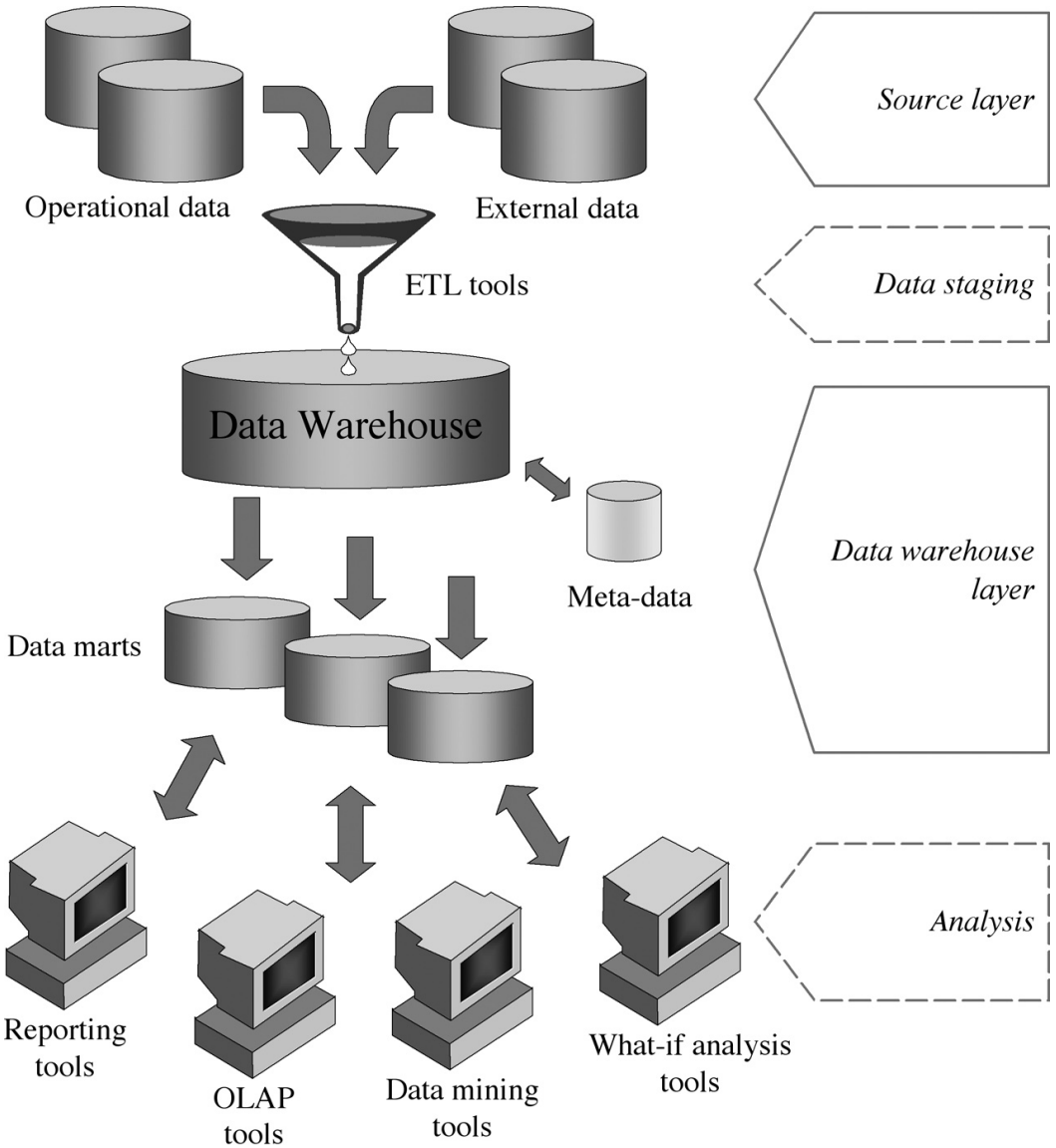


Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

DW: jednovrstvá architektura

- minimalizuje množství uložených dat
- vlastní DW je *virtuální*
- DW implementován jako multidimenzionální pohledy (views) do operační DB
- neodděluje analytické a transakční zpracování dat
  - analytické dotazy zatěžují operační DB
  - potenciální nedostatek výkonu
- neudržuje více dat než zdroj
- nejjednodušší, málo nasazovaný přístup

DW: dvouvrstvá architektura

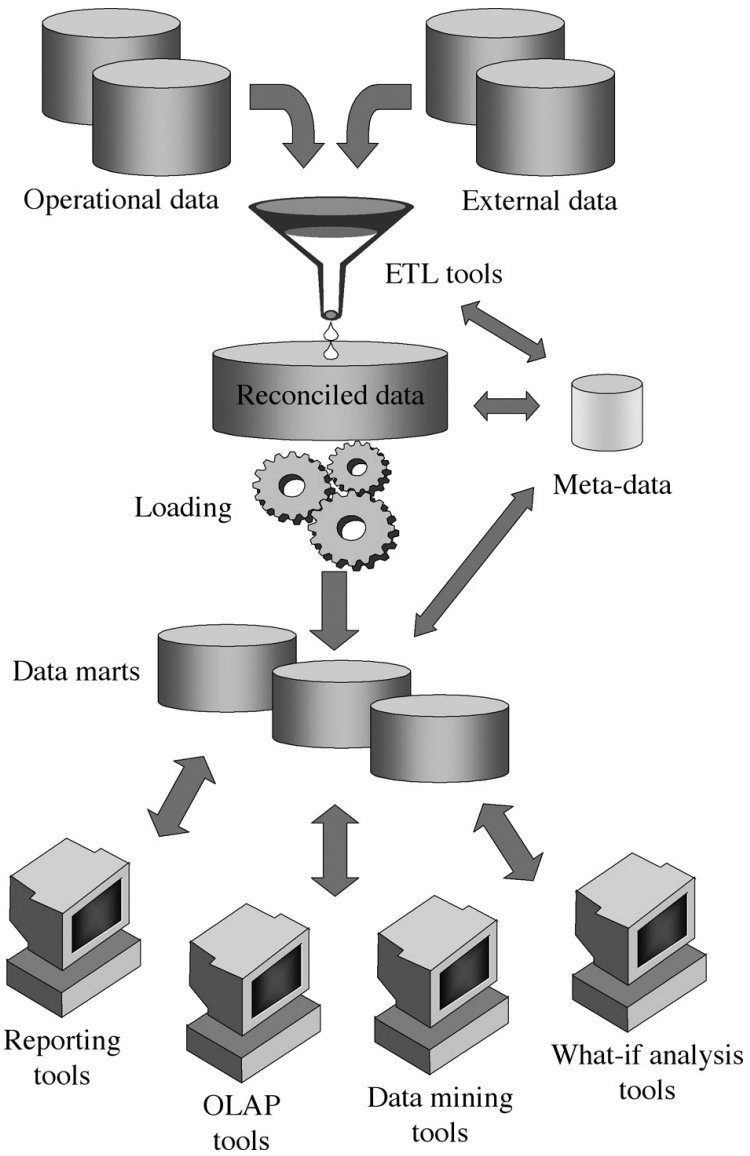


Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

DW: dvouvrstvá architektura

- odděluje analytické a transakční zpracování dat
- podpora integrace dat z více zdrojů, ETL
- DW existuje fyzicky
  - alternativní modelovací metody
- rozdělení na DM
- meta-data
- ukládání historie
- ...

DW: třívrstvá architektura



Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

DW: třívrstvá architektura

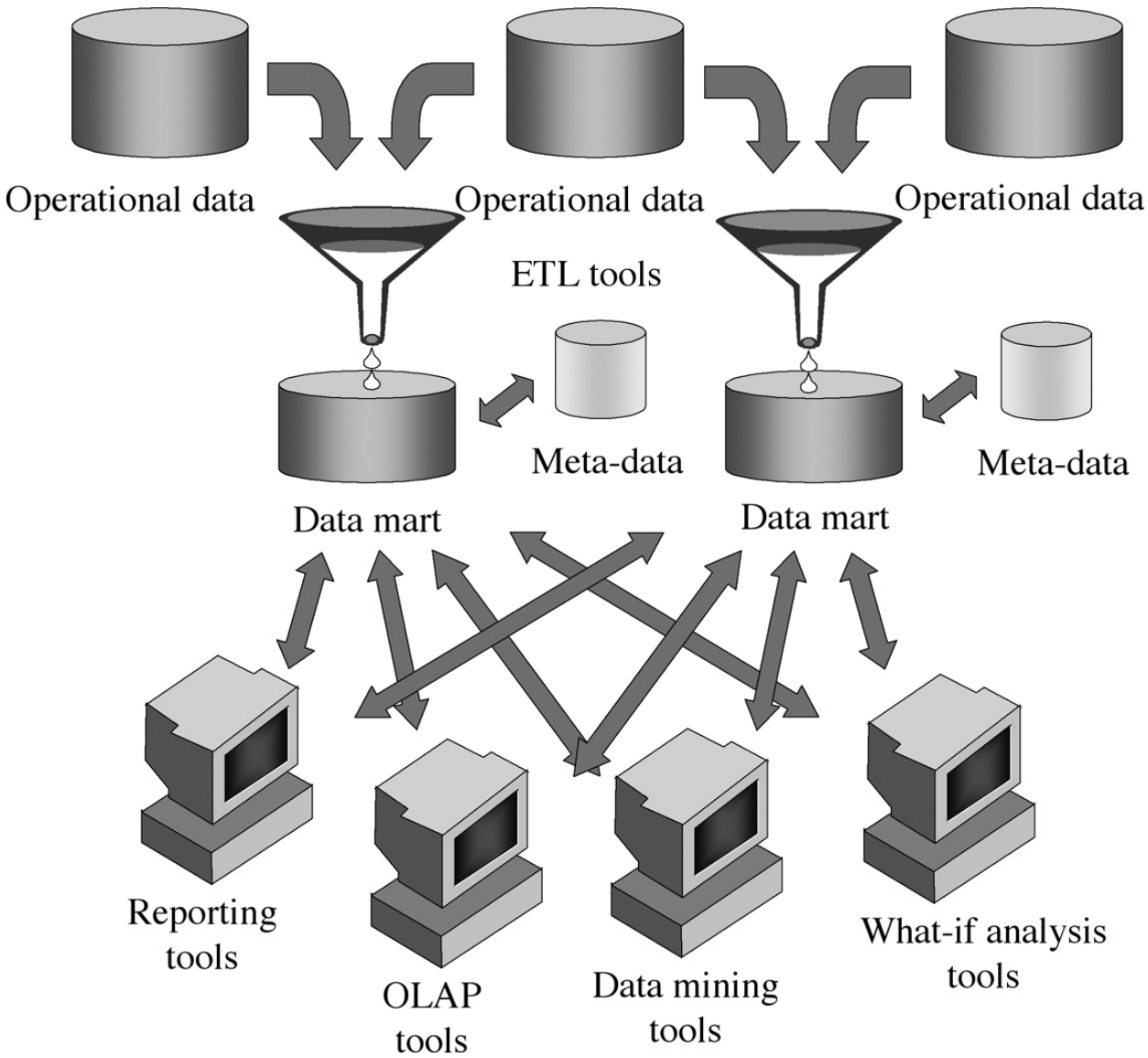
- dvouvrstvá architektura doplněna o tzv. Reconciled Data Layer (RDL, vrstva pro sladění dat) nebo ODS
- DW pak není plněn přímo ze zdrojů, ale z RDL/ODS
- odděluje problémy extrakce a integrace dat od plnění DW
- nová vrstva přidává další datovou redundanci do systému

DW: hybridní přístup

- na pomezí jedno a vícevrstvé architektury
- agregovaná či sumarizovaná data uložena fyzicky v DW
  - vhodné pro multidimenzionální dotazy
- detailní data uložena pouze ve zdrojové DB
  - v případě potřeby dostupné DW
- snižuje datovou redundanci a nároky na úložiště v DW

Design Methodologies

DW: nezávislá datová tržiště



Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

DW: nezávislá datová tržiště

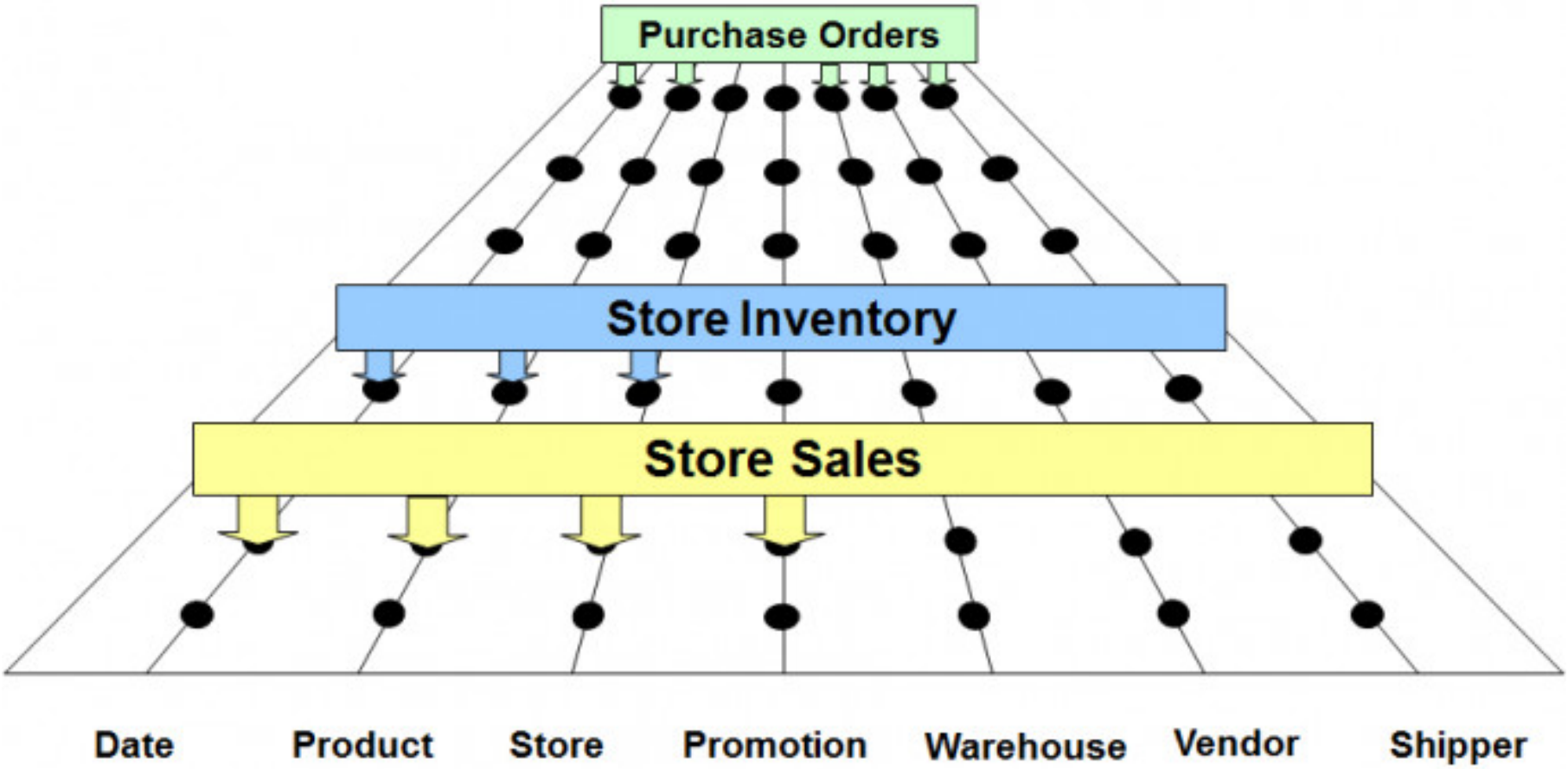
- Independent Data Marts
- datová tržiště vznikají nezávisle
- analytické nástroje je používají dle potřeby
- komplikuje integraci dat
- vhodné pouze v případě nedostatku zdrojů
  - limitující funkcionalita

DW: architektura sběrnice

- Bus Architecture
- návrh zespodu nahoru (bottom-up design)
- algoritmus pro detekci tzv. Conformed Dimensions v DM
- sestavení sběrnice z těchto dimenzí
- nezávislé, avšak homogenní DM tak vytvoří koherentní DW
- výhody
  - použitelné s prvním DM
  - iterativní přístup
- nevýhodou jsou problémy s granularitou při rozšiřování
- propagátorem metody je Ralph Kimball[6]

[6] http://www.kimballgroup.com

DW: architektura sběrnice



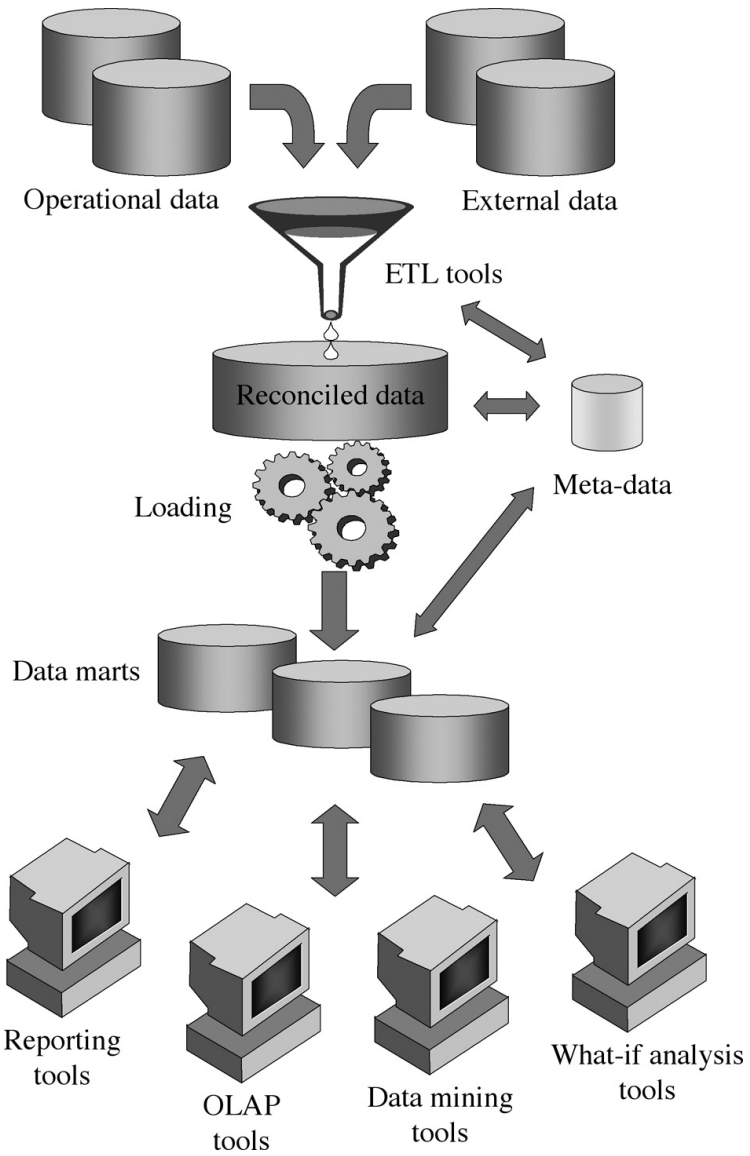
Zdroj: Kimball Group, Enterprise Data Warehouse Bus Architecture
http://www.kimballgroup.com/wp-content/uploads/2013/08/Data-Warehouse-Bus-Architecture-e1376687624708.png

DW: návrh shora dolů

- top-down design
- centralizovaný přístup
- detailní data v DW uložena normalizovaně (do jisté míry)
- DM v multidimenzionální formě jsou plněny z centrálního repozitáře
- výhody
  - produkuje vysoce konzistentní DM
  - po dokončení odolné vůči změnám v business procesech
- nevýhodou je značná časová náročnost do dokončení
  - DW není příliš využitelný před dokončením
- propagátorem metody je William H. Inmon[7]

[7] http://www.inmoncif.com/

DW: Hub-and-Spoke Architecture

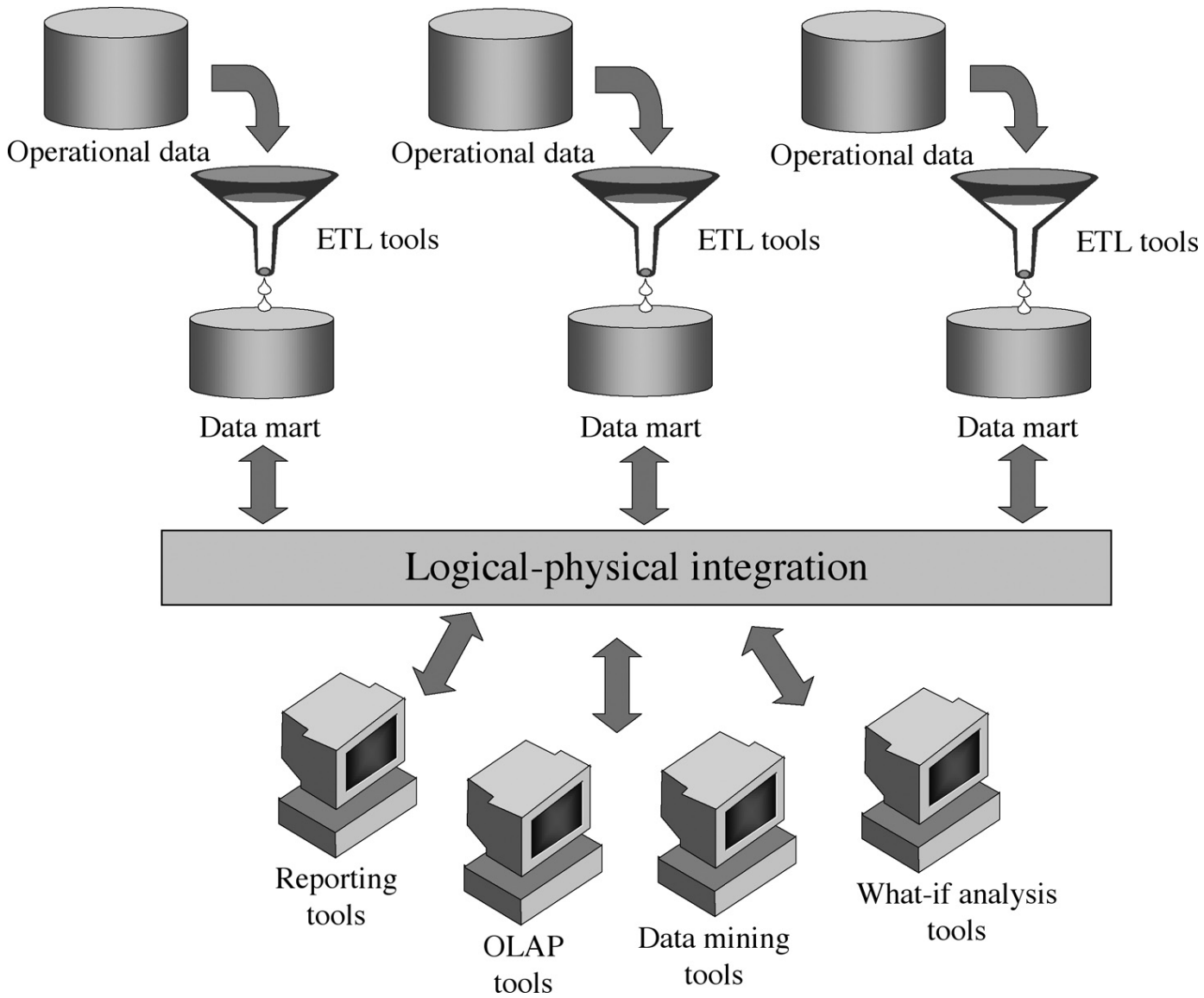


Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

DW: Hub-and-Spoke Architecture

- atomická data ukládána normalizovaně v Reconciled Data Layer (RDL)
- agregovaná a sumarizovaná data ukládána do DM v multidimenzionální formě
- podobné předchozímu návrhu
  - avšak detailní a agregovaná data nemusí být fyzicky uložena v jednom repozitáři
  - uživatelé většinou pracují s DM
  - k RDL přistupují pouze výjimečně pro detailní data
- Data Vault Modelling odpovídá této architektuře

DW: sjednocující architektura



Zdroj: Data Warehouse Design: Modern Principles and Methodologies, ISBN: 9780070677524

DW: sjednocující architektura

- Federated Architecture
- integruje již existující DW či DM
  - vytvoření jednotného rozhraní pro přístup ke všem datům

Data Warehouse: architektura

Data Loading Approaches (přístupy nahrávání dat)

- žádná data
  - použitelné pouze u jednovrstvého DW
  - zcela aktuální data
- off-line
  - data aktualizována v pravidelných intervalech, např. hodiny, dny, týdny, . . .
  - DW nemá vždy aktuální data
- on-line
  - DW aktualizován s každou transakcí
  - zcela aktuální data
- on-line z více zdrojů
  - jako předchozí bod, ale z více zdrojů

Data Back-End

Databázové systémy

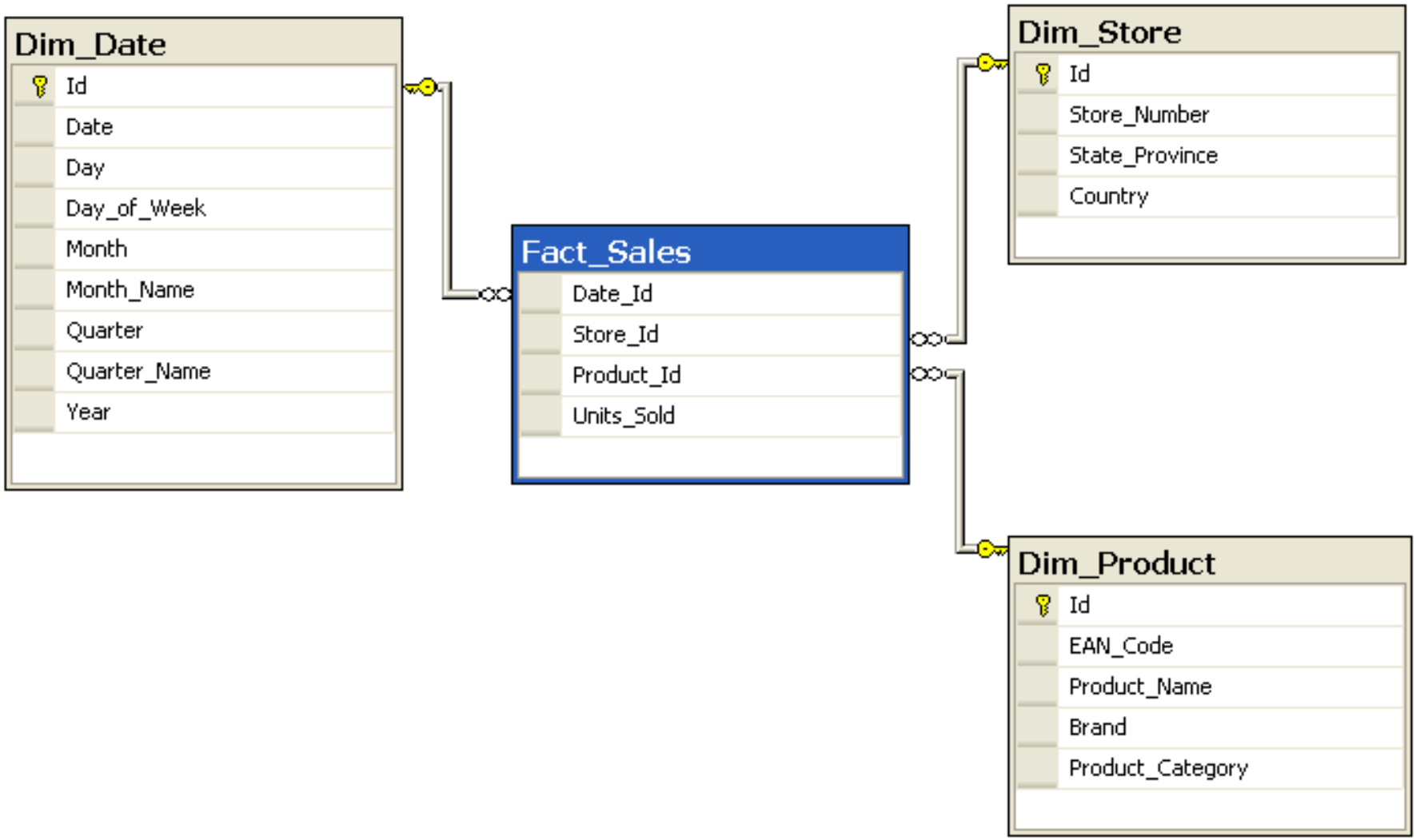
- relační model
  - data uložena v relacích (tabulkách) v podobě n-tic (řádků/záznamů)
  - tabulky definovány schématem relace
  - nejrozšířenější DBMS
  - ROLAP
  - implementace multidimenzionální funkcionality na relační DB
- multidimenzionální databáze
  - data ukládána v hyperkostkách
  - protiklad užívání tabulek v relačních db
  - dříve nepříjemná omezení
  - MOLAP

Databázové systémy (2)

- grafové databáze
  - data definována i uložena pomocí grafové struktury
  - založeno na teorii grafů
  - oproti relačním DB někdy rychlejší, lépe škálují
  - vhodné pro dotazy podobné grafovým operacím, např. hledání cesty
- síťový model
  - flexibilní způsob ukládání vztahů mezi objekty
  - uzel může mít více rodičů i potomků – zobecněný graf
- hierarchické databáze
  - data organizována ve stromových strukturách
  - registry MS Windows
- triplestore
  - vhodné pro ukládání trojic
- ...

Data Back-End

Schéma hvězdy

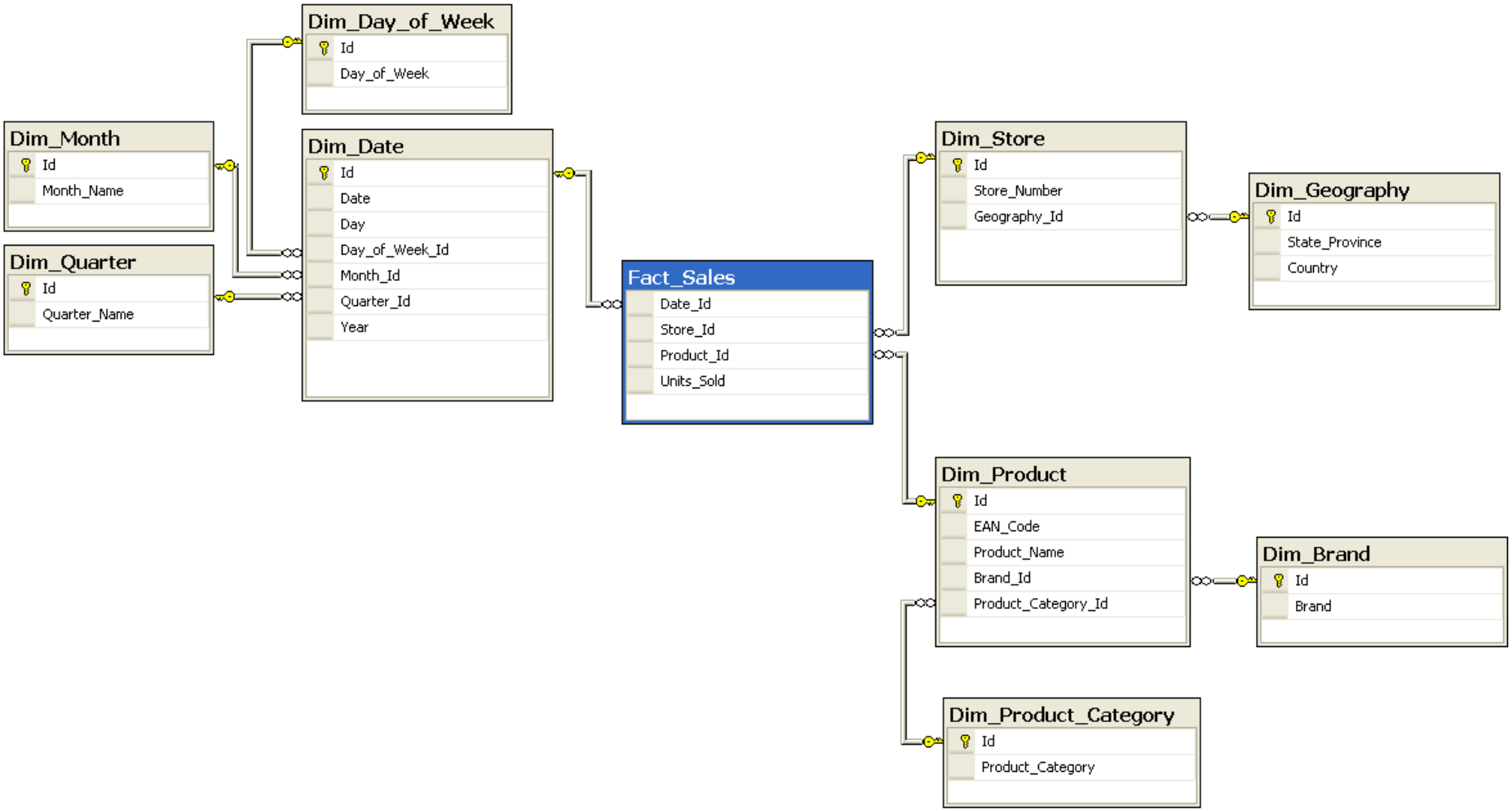


Zdroj: `http://upload.wikimedia.org/wikipedia/en/f/fe/Star-schema-example.png`

Schéma hvězdy

- Star Schema
- dovoluje relační DB simulovat multidimenzionální DB
- nejjednodušší schéma DW
- *faktová tabulka* uprostřed spojuje *dimenzní tabulky* okolo
- faktové tabulky
  - nesou (zejména) číselné údaje – fakta
  - zabírají nejvíce místa
- dimenzní tabulky
  - nesou atributy faktů – jejich popis, kontext
  - související atributy v jedné tabulce
  - minimalizace počtu dimenzních tabulek
  - mají velké množství sloupců (atributů)

Schéma sněhové vločky



Zdroj: `http://upload.wikimedia.org/wikipedia/commons/7/73/Snowflake-schema-example.png`

Schéma sněhové vločky

- Snowflake schema
- opět v centru faktové tabulky a okolo dimenzní
- avšak dimenze jsou normalizované (do určité míry)
- dimenzní tabulky rozloženy procesem normalizace do několika propojených tabulek
- efektivní zejména
  - pro díravé (sparse) dimenze
  - má-li dimenze velké množství atributů

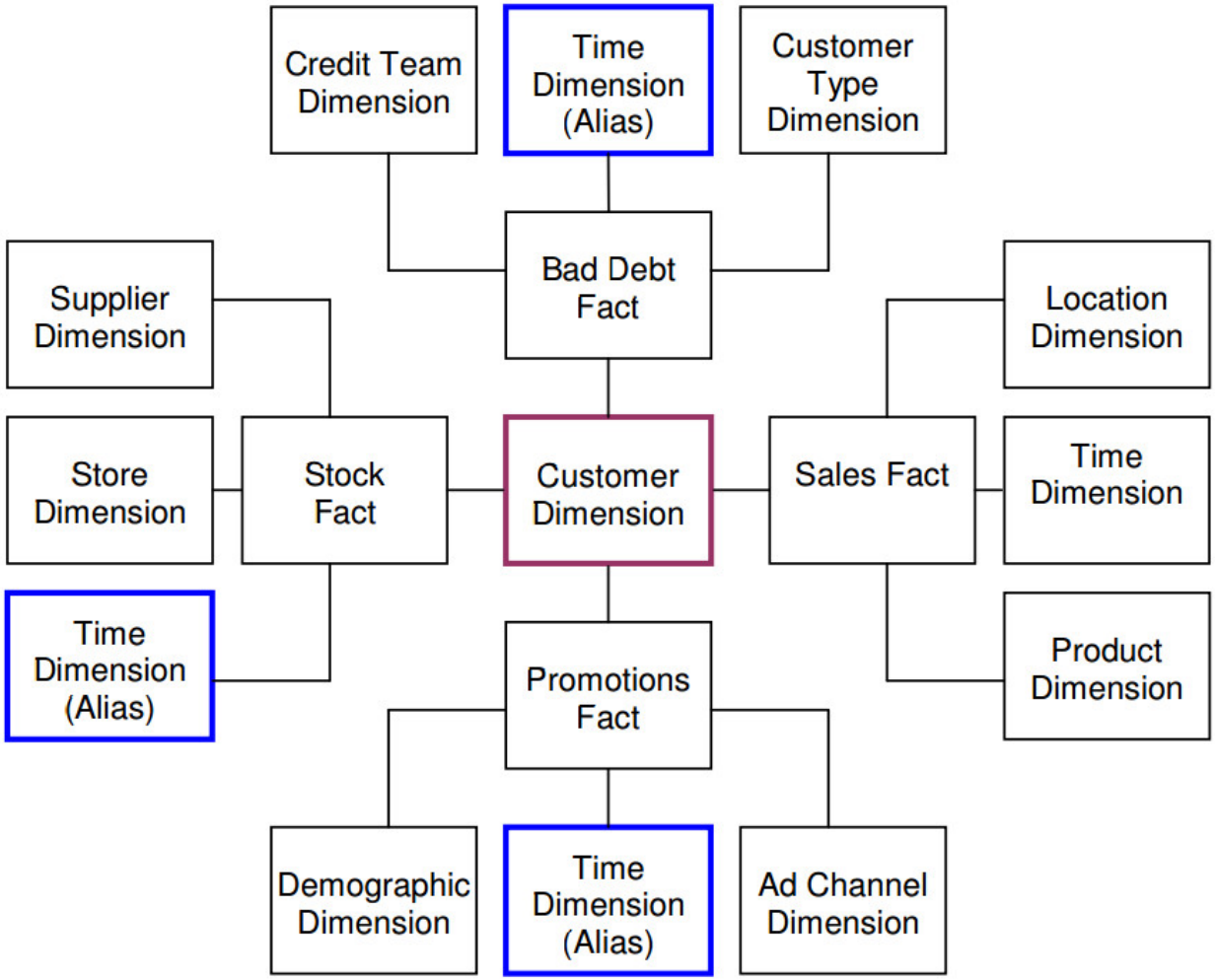
Hvězda vs. vločka

- obě optimalizují čas/rychlost získání dat (operace čtení)
- hvězda vhodná pro nástroje, které odhalují DB schema uživatelům
  - snadno pochopitelná, přirozená spojení
- vločka vhodná pro sofistikované nástroje, které oddělují data od uživatelů

Reverzní hvězda

- Reverse Star Schema
- optimalizace na získání velkého množství popisných dat
- návrh převrací některá pravidla hvězdy naruby
  - k centrální tabulce se připojují faktové tabulky
  - více centrálních tabulek
  - rozdílné kardinality
  - . . .

Reverzní hvězda



Zdroj: `http://www.pcthompson.co.uk/documents/The_Reverse_Star_Schema_v2.1.pdf`

Data Back-End

DV: přehled

- modelovací metoda
- ukládání historie dat
- integrace dat z více zdrojů
- datový audit, původ dat (data tracking)
- uchovávání chybných hodnot
- paralelní nahrávání dat
- 100 % dat 100 % času
- může nahradit ODS
- fakta vs. pravda
- navržen Danem Linstedtem[8]

[8] http://danlinstedt.com/

DV: komponenty

- Hubs
  - primární klíče (business klíče)
  - tyto se téměř nemění, např. učo
- Links
  - integrace transakcí a vztahů mezi Hubs
- Satellites
  - kontext Hubs a Links
- Point-in-Time Tabules, Bridge, pomocné tabulky, . . .

Shrnutí

Shrnutí

- Operační/Transakční DB
  - ukládá data s ohledem na bezpečné a efektivní zpracování transakcí v konkurečním mnohouživatelském prostředí
  - zaměřuje se zejména na aktuální data
  - čtení a zápis
  - velké množství uživatelů
- Data Warehouse
  - ukládá data s ohledem na efektivitu zpracování složitých dotazů
  - zejména operace *select* a *insert*
  - využívá multidimenzionální funkcionality
  - usnadňuje udržování historie data
  - snaha o čištění dat
  - analytické zpracování nezatěžuje transakční DB
  - podpora pro analytické nástroje
    - OLAP, reportování, DM nástroje, . . .
  - data tracking
  - správa chybných dat