evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

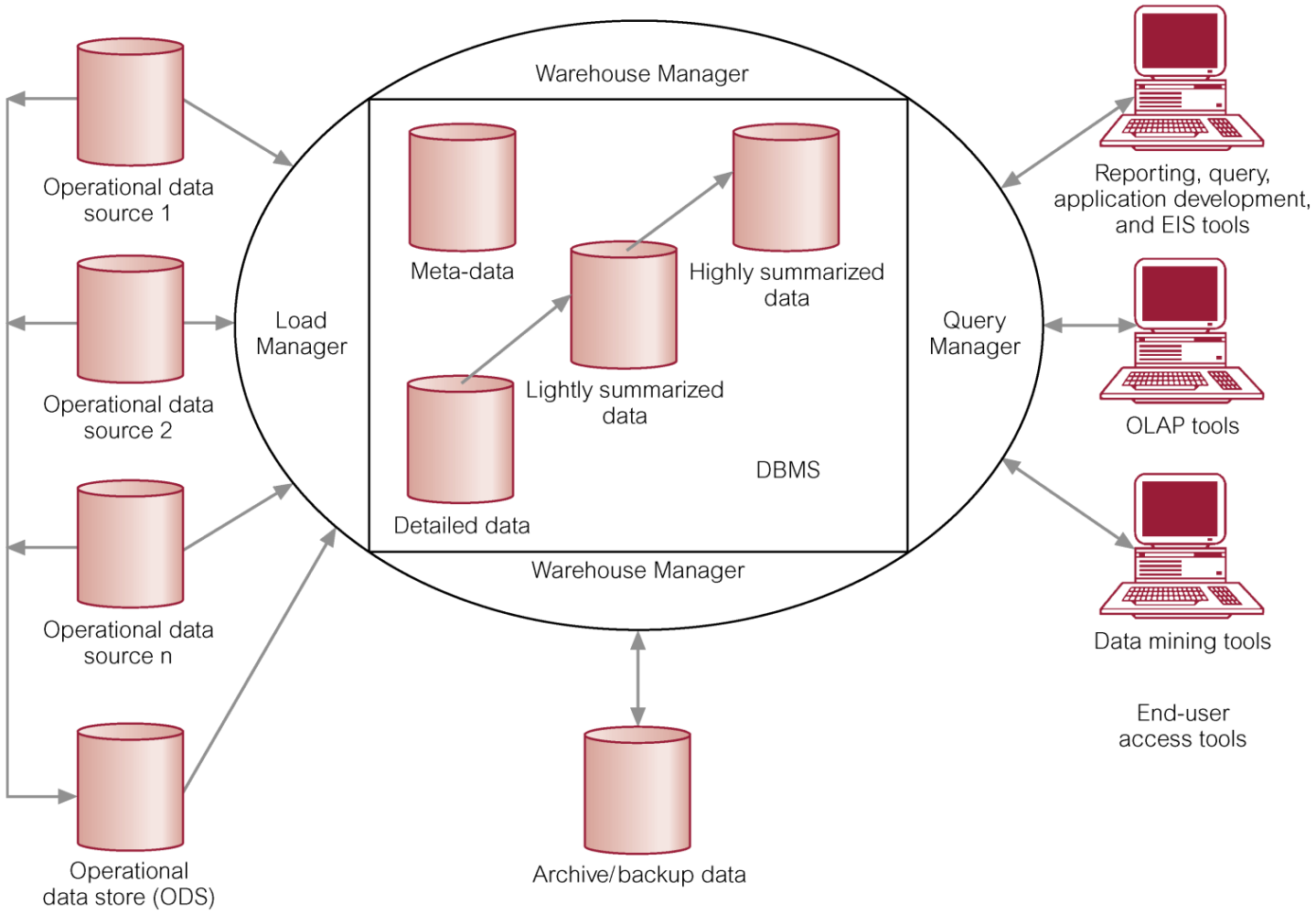
METODOLOGIE NÁVRHU DATOVÉHO SKLADU

Zdroj: Connolly T., Begg C.: Database systems, 2010, Chap. 33

Návrh datového skladu

- Datový sklad se používá pro podporu rozhodovaní a obsahuje data získaná z databází operativních dat, která byla transformována pro potřeby analytických dotazů, sjednocena a zkontrolována před jejich uložením do datového skladu.
- Na začátku vytváření datového skladu je vhodné si položit následující otázky:
  - Které požadavky uživatelů jsou nejdůležitější?
  - Která data máme zpracovat jako první?
  - Je vhodným řešením rozdělit požadavky na datový sklad na více jednodušších částí, pokrývajících požadavky pouze vybraných skupin uživatelů a budovat nejdříve datová tržiště (data marts) s tím, že cílem zůstává kompletní datový sklad?
- Málokterý tvůrce datového skladu je ochoten budovat komplexní datový sklad v jednom kroku.

Typická architektura datového skladu



Zdroj: Connolly T., Begg C.: Database systems, 2010

Zdroje dat pro datový sklad

- Operativní data z různých podnikových databází.
- Data z dostupných externích systémů (například z komerčních databází, databází dodavatelů a odběratelů).
- Sklady operativních dat – předzpracovaná data pro datový sklad, obsahující kromě operativních dat také jednoduché agregace.
- ETL (Extraction,Transformation, Loadig) manager
  - extrahuje data z prvotních zdrojů,
  - transformuje je do tvaru vhodného pro analytické zpracování a kompatibilního se strukturou datového skladu,
  - vloží upravená data do datového skladu.

Zpracování dat v datovém skladu

- Warehouse manager
  - zajišťuje konzistenci dat v datovém skladu,
  - připojí data k existujícím datům,
  - vytvoří indexy,
  - vytvoří agregace,
  - zálohuje data.
- Query manager zajistí zpracování dotazů.
- Management metadat je velmi složitý, neboť metadata obsahují informace nejen o struktuře datového skladu, ale také o
  - o původních strukturách a jejich transformaci,
  - o agregacích,
  - dalších strukturách souvisejících s analytickými dotazy.

Metodologie návrhu datového skladu

Existují dvě hlavní metodologie:

- **Corporate Informtion Factory (autor W. H. Inmon, 2001)**
  - Začíná vytvořením datového modelu pro celý podnik, z něhož se potom odvíjí vytváření datových tržišť - datových skladů pro potřeby jednotlivých oddělení.
  - Využívá tradiční databázové metody a techniky; předpokládá 3NF.
- **Business Dimensional Lifecycle (autor R. Kimball, 2008)**
  - Začíná identifikací požadavků na informace (analytických témat) a příslušných business procesů.
  - Vybere první skupinu uživatelů, pro kterou vytvoří datové tržiště; jednotlivá tržiště se potom integrují do datového skladu.
  - Pro návrh datového modelu pro každé tržiště používá novou techniku nazývanou dimenzionální modelování.

Dimenzionální modelování

Dimenzionální model má hvězdicové schéma, které má uprostřed jednu tabulku (tabulka faktů) se složeným primárním klíčem a z množiny menších denormalizovaných tabulek (tabulky dimenzí). Každá tabulka dimenzí má jednoduchý umělý primární klíč, který koresponduje s právě jednou složkou primárního klíče tabulky faktů. Dimenze obvykle tvoří hierarchie, jejichž členy jsou vhodnými kandidáty pro agregaci dat (faktů). Většina dimenzí se mění pouze pomalu; některé mají podobné vlastnosti jako číselníky. Obecnější přístup připouští množinu hvězdicových schémat (souhvězdí), kde jedna dimenze může být sdílena několika tabulkami faktů.

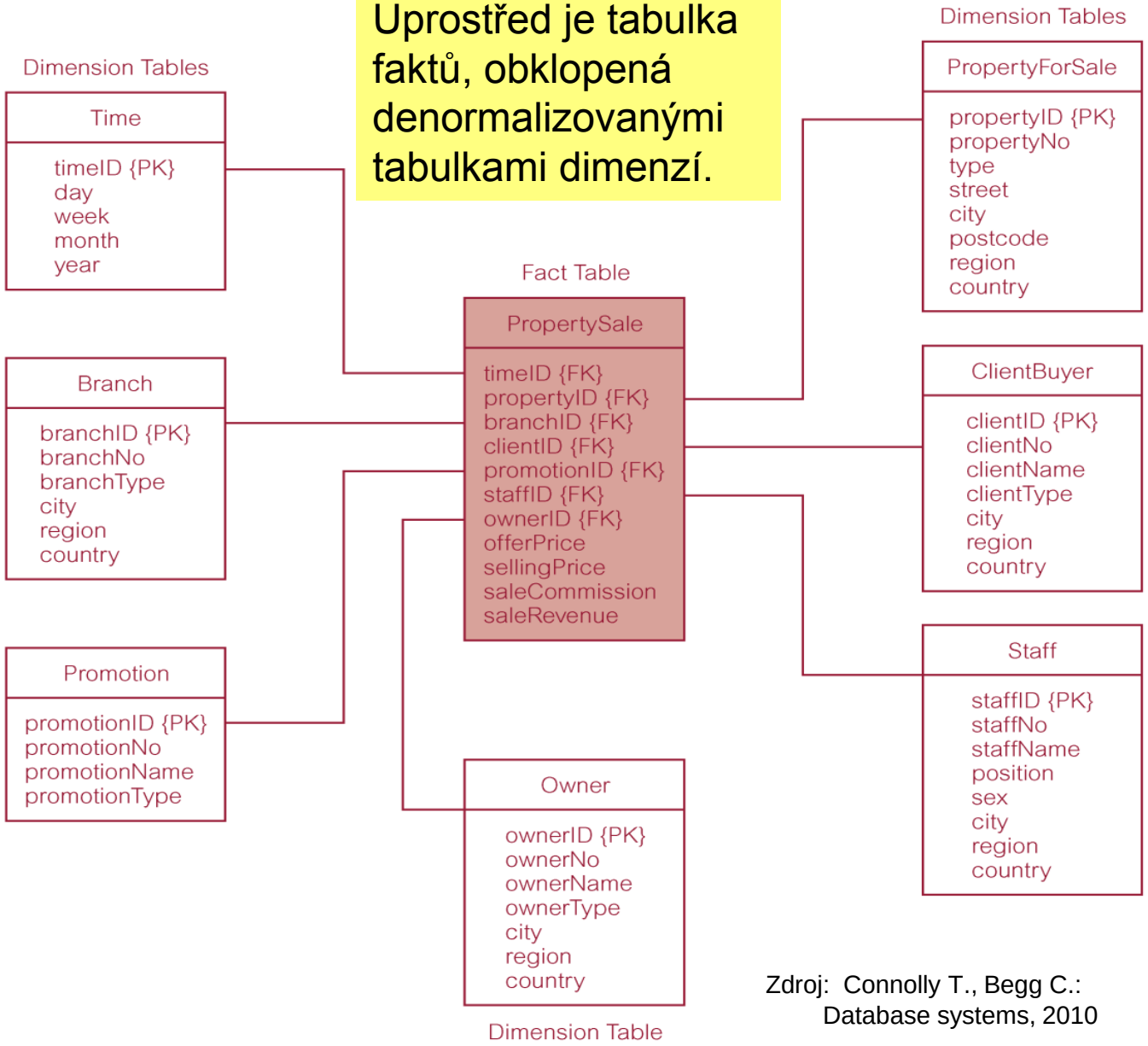
- Hlavním předmětem zájmu jsou fakta, která jsou modelována neklíčovými atributy tabulky faktů. Tyto atributy jsou obvykle numerické a aditivní, představují jisté míry (cena, množství, …).

Příklad: Hvězdicové schéma pro prodej nemovitostí v realitní kanceláři, která má několik poboček

Fakta (neklíčové položky v tabulce faktů) představují nabídkovou a prodejní cenu, provizi a celkový příjem z každého jednotlivého prodeje – a jsou to numerické a aditivní hodnoty.

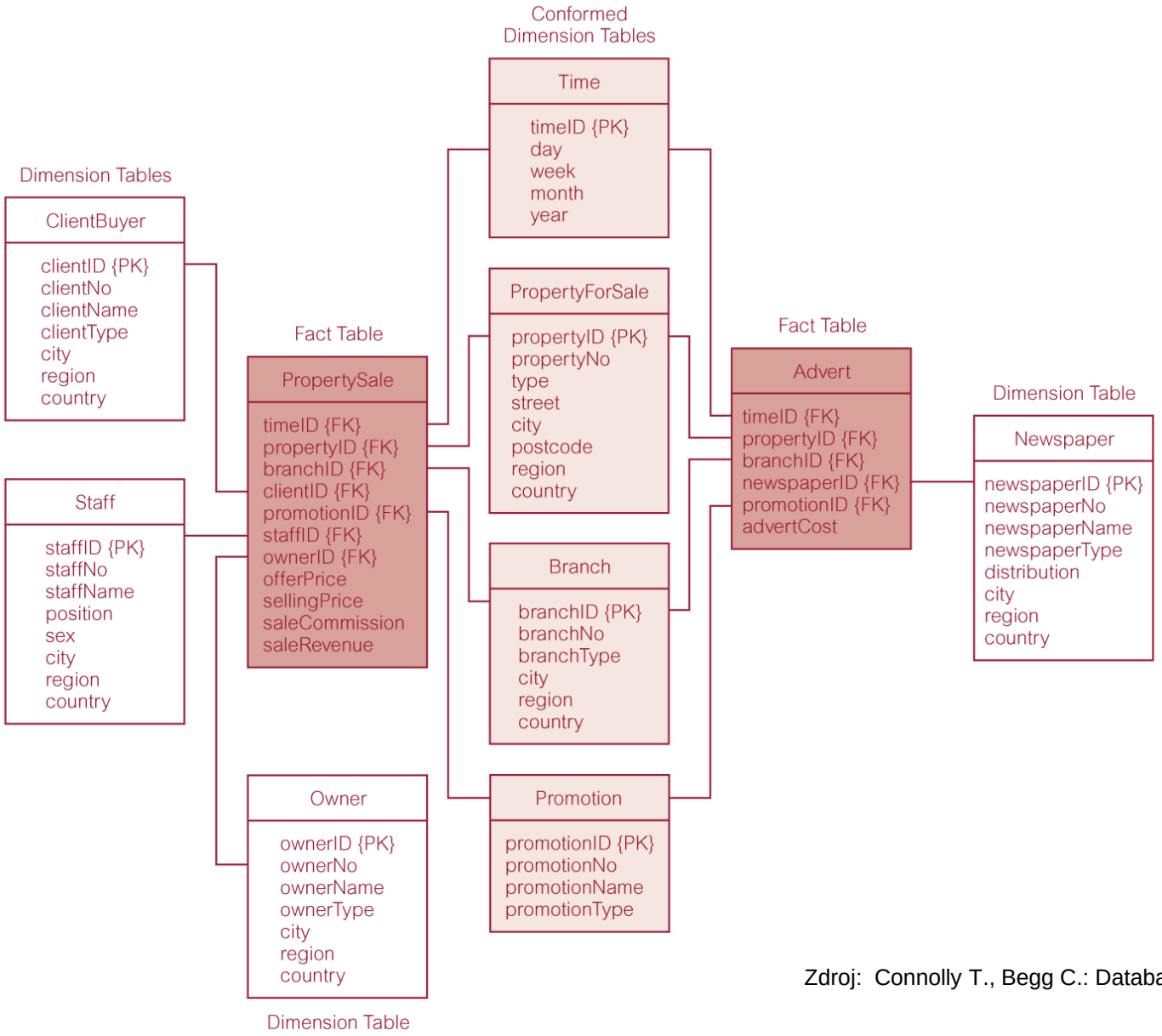
Tabulka faktů může být extrémně velká v porovnání s tabulkami dimenzí. Tabulku lze zmenšit použitím agregovaných hodnot.

Uprostřed je tabulka faktů, obklopená denormalizovanými tabulkami dimenzí.



Zdroj: Connolly T., Begg C.: Database systems, 2010

Prodej a inzerce – společné dimenze

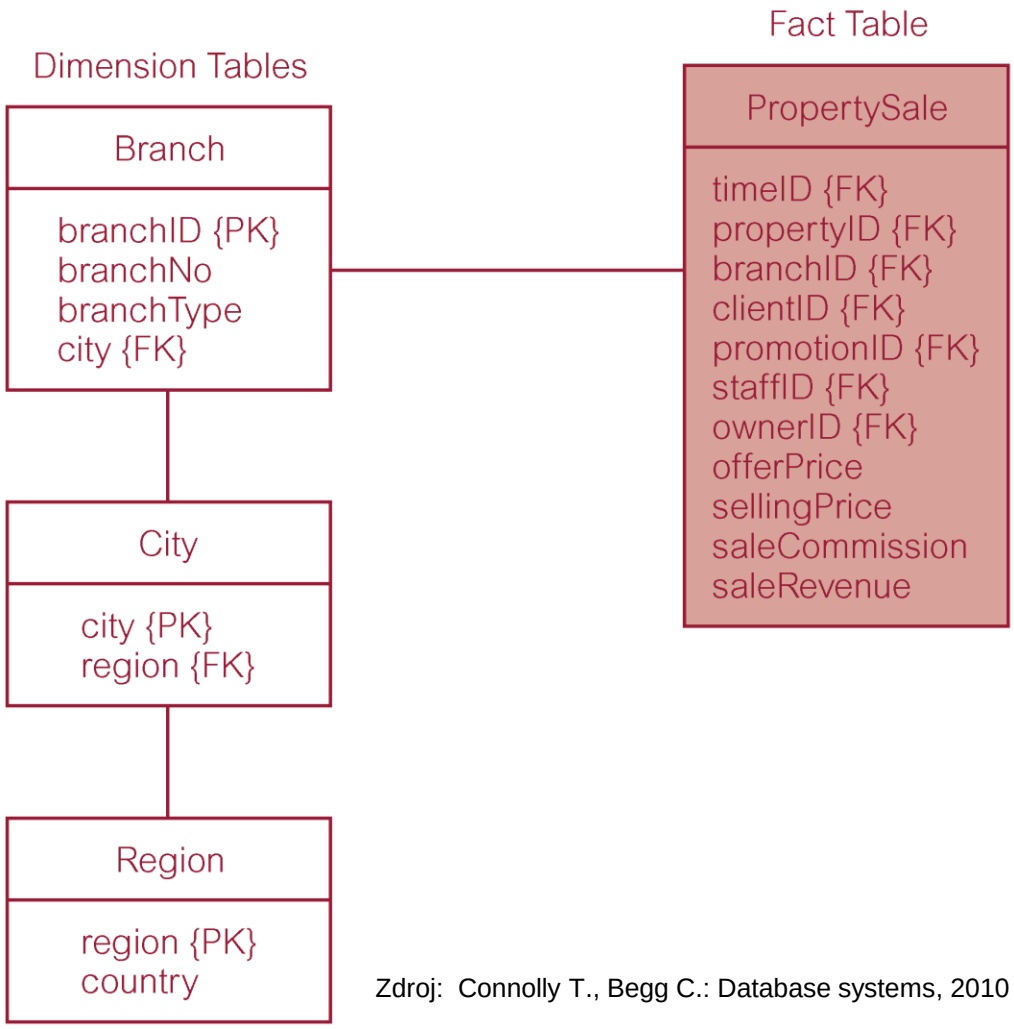


Zdroj: Connolly T., Begg C.: Database systems, 2010

Příklad: Část schématu sněhové vločky pro prodej nemovitostí

Schéma sněhové vločky:

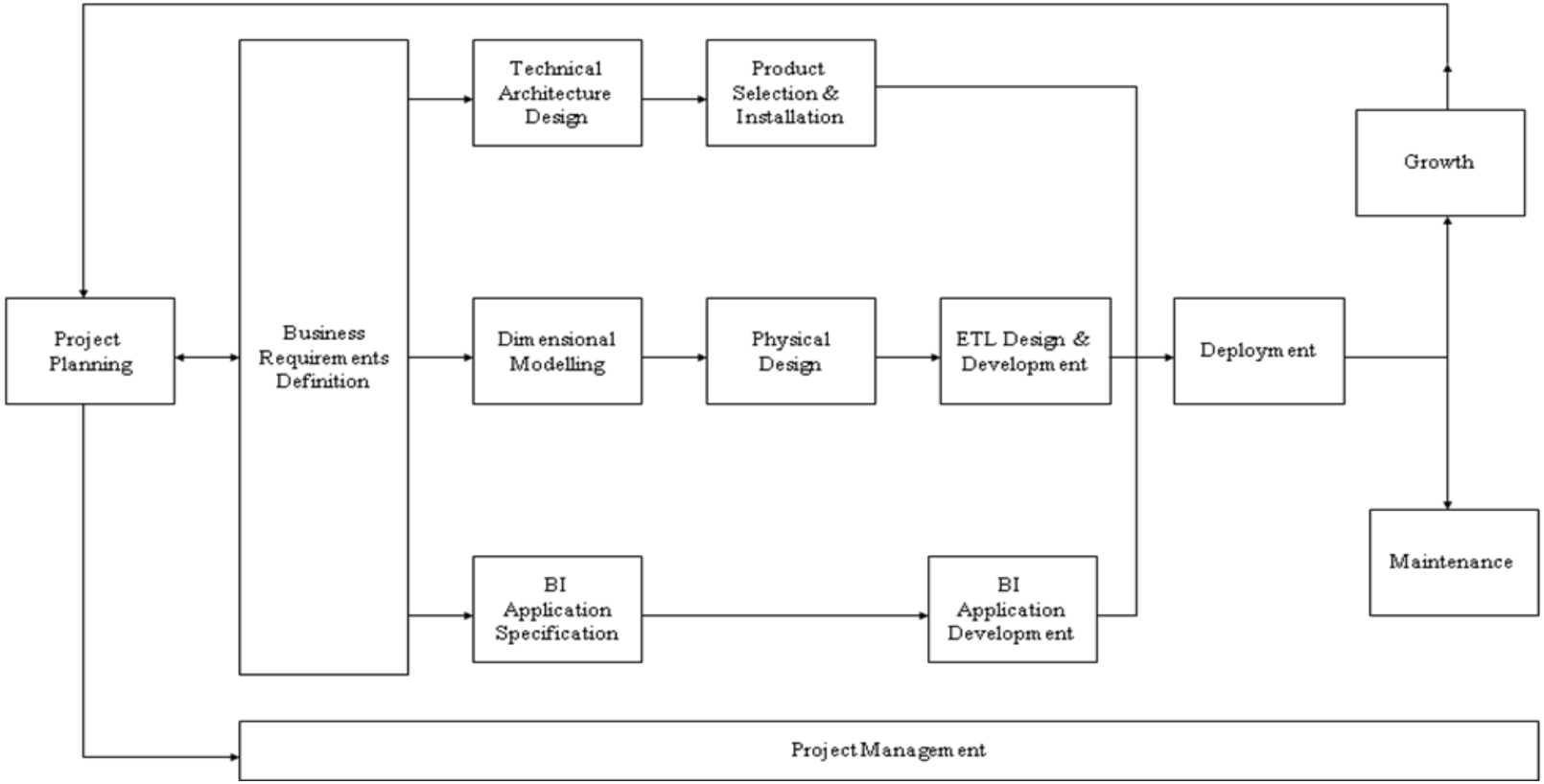
Dimenzionální model s tabulkou faktů uprostřed a normalizovanými tabulkami dimenzí.



Zdroj: Connolly T., Begg C.: Database systems, 2010

Kimballova metodologie

http://www.kimballgroup.com/



Zdroj: Connolly T., Begg C.: Database systems, 2010

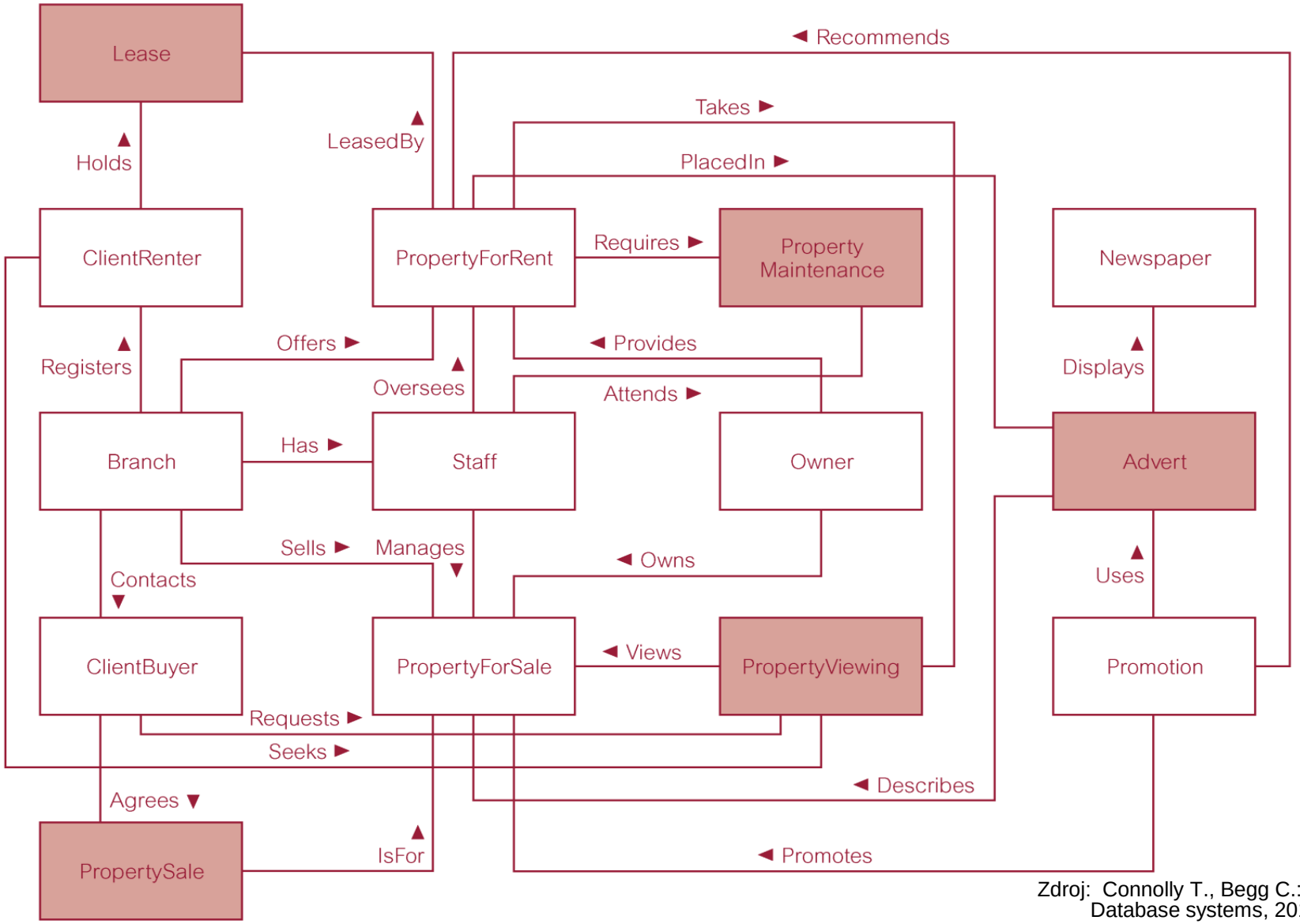
Základní kroky dimenzionálního modelování

1. Vyber business proces
   - první by měl být ten, který bude z ekonomického hlediska zajímavý a u kterého je předpoklad, že bude dokončen včas
2. Vyber granularitu
   - granularita určuje co budou reprezentovat hodnoty v tabulce faktů - zda jednotlivé či agregované hodnoty (například za určité období)
3. Vyber dimenze
   - dimenze, které budou reprezentované ve více než jednom dimenzionálním modelu musí být shodné, nebo jedna má být podmnožinou druhé. Společné dimenze hrají důležitou roli při integraci jednotlivých datových tržišť.
4. Identifikuj fakta
   - Fakta by měla být numerická a aditivní. Všechna fakta musí být vyjádřena ve stejné (vybrané) granularitě. V předchozím příkladu je úrovní granularity jednotlivý záznam, tj. jednotlivý prodej.
5. Identifikuj atributy dimenzí
   - Dimenze se určují v kontextu požadovaných dotazů na fakta. Obvykle obsahují popisné informace, které se využívají při formulaci dotazů.

Typické dotazy na datový sklad realitní kanceláře

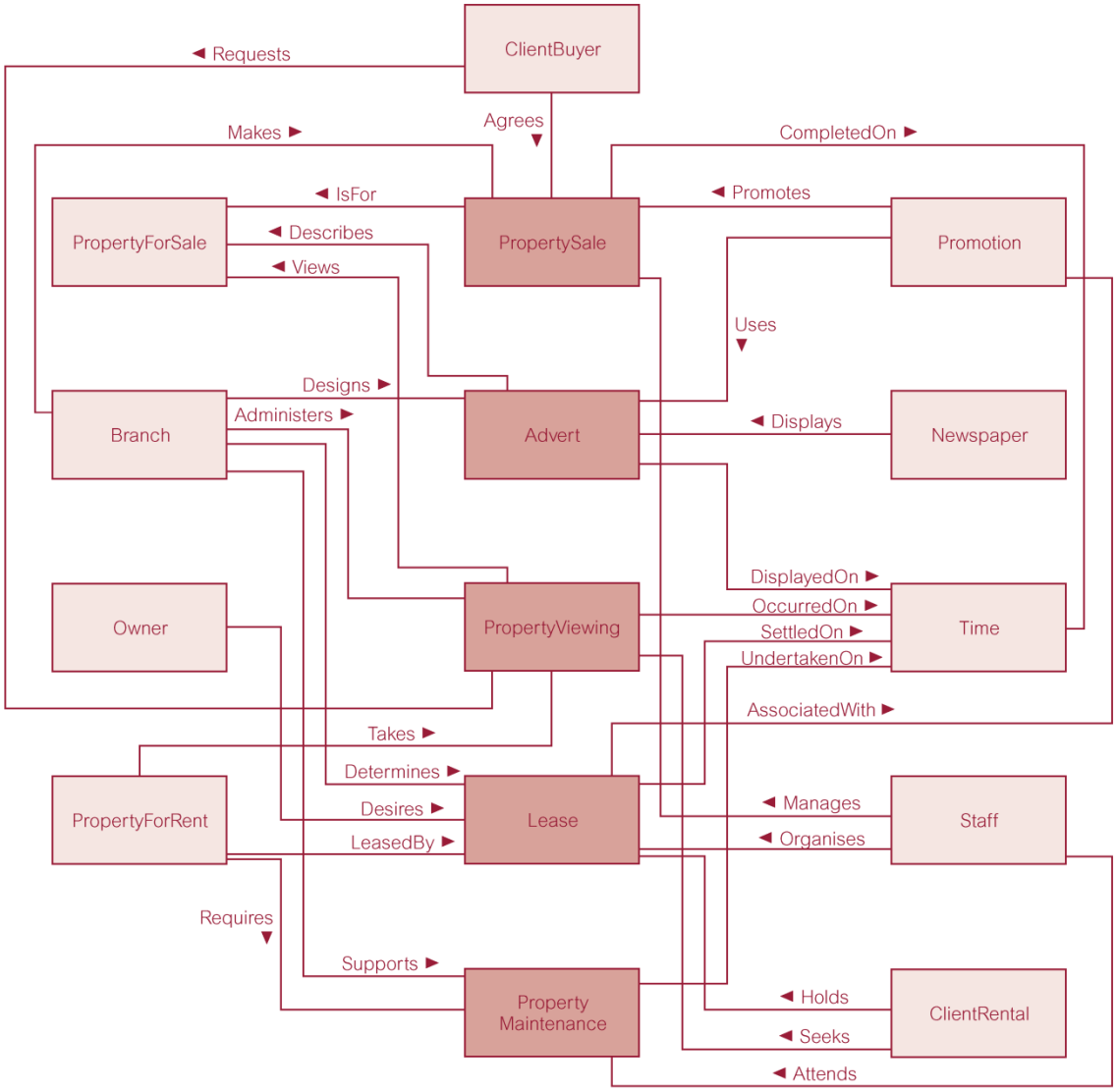
- Jaký byl celkový příjem v pobočkách ve Skotsku v roce 2012?
- Jaký byl celkový příjem ve firmě za rok 2012?
- Jaký byl výnos z prodeje jednotlivých typů nemovitostí v roce 2012?
- Která byla nejžádanější oblast v jednotlivých městech při pronájmu nemovitostí v roce 2012 a jak se změnily priority v porovnání s předchozími 3 roky?
- Jaký je celkový obrat v prodeji nemovitostí v jednotlivých pobočkách?
- Jaký je vztah mezi ziskem jednotlivých poboček a celkovým počtem pracovníků pobočky za roky 2005-2012?

Celkový ER model realitní kanceláře s vyznačením business procesů



Zdroj: Connolly T., Begg C.: Database systems, 2010

Dimenzionální model realitní kanceláře



Zdroj: Connolly T., Begg C.: Database systems, 2010

Porovnání metodologií

Metodologie	Hlavní výhoda	Hlavní nevýhoda
Inmon	Potenciální vytvoření konzistentního a úplného pohledu na všechna data organizace.	Komplexní, rozsáhlý projekt, který může selhat aniž by přinesl očekávané výsledky v plánovaném čase nebo po vynaložení plánovaných prostředků.
Kimball	Rozdělení projektu na etapy umožní demonstrovat přínos prvního datového tržiště v plánovaném čase nebo po vynaložení plánovaných prostředků.	Datová tržiště, vytvářena postupně mohou být budována případně i jinými vývojovými týmy a při použití jiných vývojových systémů. Konečného cíle, kterým je komplexní datový sklad, nemusí být v plné míře nikdy dosaženo.