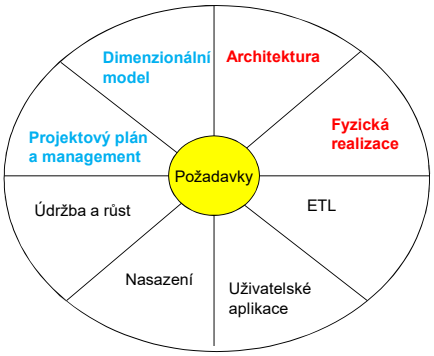
Databázové systémy a metody zpracování dat

Návrh technické architektury a infrastruktury

8.přednáška

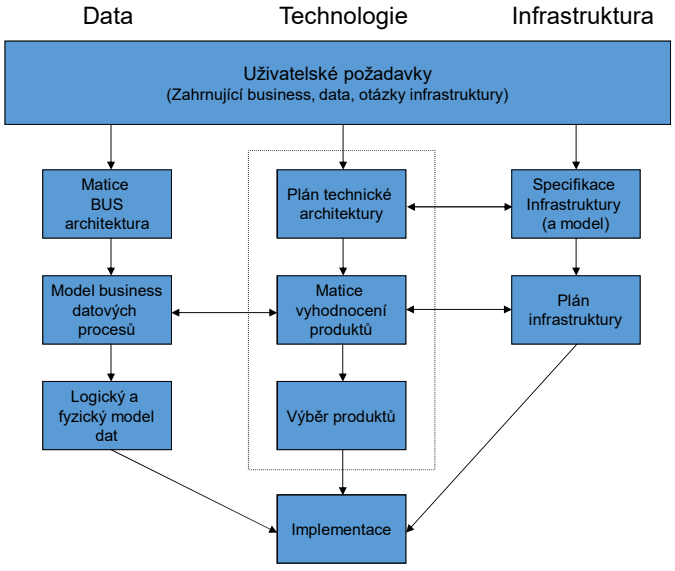
Architektura



Architektura

- Výstupy návrhu architektury
  - Plán technické architektury
    - Shrne požadavky uživatelů
    - Návrh budoucí architektury DW
  - Plán infrastruktury
    - Někdy součástí plánu technické architektury
    - Popisuje servery, desktopy, síť

Architektura



Plán technické architektury

- Během interview získat požadavky na architektury
  - Její první návrh lze udělat už v průběhu interview nebo před a následně upřesňovat
- Spolupráce s vybranými pracovníky (IS) na oponentuře a tvorbě architektury
  - Vypracovat draft plánu
  - Nechat oponovat
- Vytvoření modelu architektury (grafického)
  - Vhodný pro komunikaci
- Návrh postupu implementace architektury
  - Výběr prioritní oblasti
  - Odhad času a zdrojů potřebných pro implementaci
- Provedení revize dokumentu s managementem

Plán technické architektury

Dotazník pro interview – Dodatečné otázky na architekturu a infrastrukturu

---

A. IS business role

- Jak důležitá je analýza dat při podpoře rozhodování managementu ve Vaší firmě?
- Jakou roli hraje IS při podpoře rozhodování (analýzy dat)?
- Mění se toto? (Vlivem konkurenčního prostředí, organizační struktury, …)

B. Technologické směrování

- Jaký je přístup Vaší firmy k IS/ICT v předchozích letech (striktně Klient/Server, web-base aplikace, ERP, …)?
- Existuje plán, specifikace, která určuje požadavky na softwarovou infrastrukturu (DCOM, CORBA, objektová orientace, …)?
- Jaké jsou Vaše plány pro nejbližší budoucnost?
- Jaké plány a záměry v oblasti infrastruktury budou mít dopad na přístup k datům (přesun dat, načasování úloh, jména serverů, bezpečnost, distribuce software, …)
- Existuje specifická role pro metadata? Jak jsou řízené?
- Jaké jsou standardní firemní produkty dnes? Jaké platformy, OS SW, DBMS, klientský SW, utility. Bude to tak i v následujících letech?
- Jaké jsou nejužší místa a otázky v oblasti infrastruktury?
- Kdo odpovídá za architekturu? Existuje nějaká dokumentace?

C. Infrastruktura

- Kdo všechno je zahrnut do nákupu, instalace a podpoře nové infrastruktury (servery, Sw, připojení)? Kdo odpovídá za bezpečnostní architekturu?
- Existuje centrální správa uživatelů, dat, bezpečnosti v organizaci?

Plán technické architektury

- Plán – nemusí být doveden na úroveň konkrétního produktu, ale měl by vycházet z potřeb uživatelů
  - Vize do budoucnosti
  - Vypsat typy uživatelů, jejich potřeby na přístup k informacím, požadavky na reporting, …

Plán technické architektury

Plán infrastruktury

- Tři základní oblasti
  - Server (HW a OS)
    - CPU, OS, Disk, Přírůstek dat, paměť
  - Síť
    - Od OLTP do DW
    - Od DW k uživatelům
    - Zabezpečení, frekvence přenosu
  - Pracovní stanice
    - HW, OS, konektivitu (ODBC, OLE DB, …)
- Plán je postupně upřesňován na základě vybraného SW podle plánu technické architektury

Výběr produktů

- Výběr na základě návrhu architektury, uživatelských požadavků
  - HW
  - DBMS
  - ETL tool
  - Nástroje pro prezentaci a přístup k datům
- Vhodné je testovat vybrané HW a SW komponenty
  - Vytvořit prototyp DW a testovat produkty
  - Nikdy je ale nelze otestovat na reálné zatížení
  - Vhodné je např. domluvit s dodavatelem 90 denní testovací období a potom teprve podepsat smlouvu

Výběr produktů

- Vytvořit Matici vyhodnocení produktů
  - Nastavit priority (váhu) jednotlivým kritériím
    - Musí mít … bylo by dobré
  - Spolupráce s výrobcem (objasnění potřebné informace)
- Provedení průzkumu trhu
  - WWW stránky
  - Publikace, časopisy
  - DW konference, fórum, portály
- Kritéria
  - Dle uživatelských potřeb pro daný typ nástroje
  - Výrobce
    - Podpora
    - Dokumentace
    - Školení
    - Konzultace
    - Externí podpora (nezávislé forum na webu o produktu, …)
    - Spolupráce s výrobce (dobrá, nekomunikativní, …)
    - Velikost, budoucnost
    - Reference

Výběr produktů

- Vybrat pro porovnání maximálně 5 produktů
- Uspořádat prezentaci výrobců
  - Získat informace na vyplnění Matici vyhodnocení produktů
  - Ukázka práce produktu
  - Domluvit ukázku u stávajícího zákazníka s podobným řešením
- Není-li možné rozhodnout – vytvoř prototyp (2 – 3 výrobci) a otestuj
  - Vhodné přenést na výrobce nebo s ním spolupracovat

Výběr produktů

- Otestovat produkty v plné šíři, ne jen omezeně
  - Jednoduché i komplexní funkce
  - Málo i hodně dat
  - Jeden a více uživatelů
- Tvorba prototypu – 4 až 6 týdnů
- Po 2 – 3 letech je vhodné výběr opakovat pro možný upgrade na nové technologie

Výběr produktů

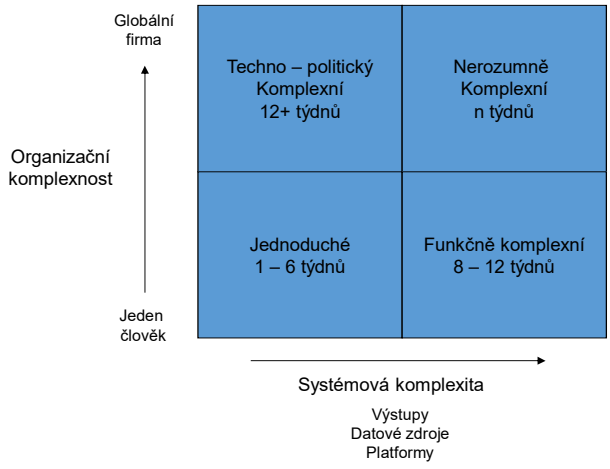
- Při tvorbě prototypu:
  - Definovat rozsah prototypu
  - Rozsah by neměl být moc velký
  - Vzít rozumnou velikost dat
  - Načíst data v původním, originálním stavu (neztrácet čas čištěním)
  - V rámci tvorby prototypu získat co nejvíce poznatků o provozních systémech, problémech v datech, HW, SW
  - Získat názory uživatelů – ukázat jim výstupy
  - Nenechat se zaslepit očekáváním

Výběr produktů

- Instalace závisí na výběru HW a SW
- Nutno dobře nainstalovat – může dosti ovlivnit výkonnost řešení
  - Spolupracovat se specialisty, dodavatelem, výrobcem
- Nezapomenout na:
  - Příprava místa na HW (servery)
  - Školení administrátorů
  - Testování po instalaci

Výběr produktů

- Čas potřebný na vytvoření plánu architektury



Fyzický design

Fyzický design - Agregace

- Agregace mohou významně zvýšit výkonnost
  - Někdy nahrazeno OLAP databází
- Uchování agregací vedle detailních dat v databázi DW
- Agregace – obvyklé souhrny podle daných dimensí
- Je třeba vložit vrstvu, která dokáže rozpoznat, zda uspokojit uživatelský požadavek přímo z detailních dat nebo existuje agregace
  - Umí některé reportingové nástroje
  - Nebo je třeba vyvinout
  - Je to největší obtíž a výzva směrem k agregacím

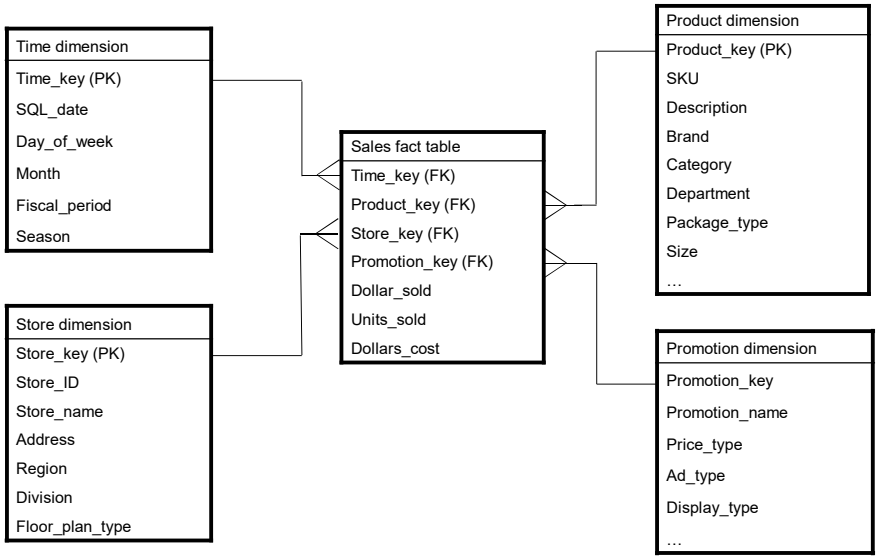
Fyzický design - Agregace

- Přidat je rozumné agregace – zvyšuje nároky na prostor na disku
- Výběr dat pro agregaci
  - Dle potřeb uživatelů
    - Např. na základě existujících reportů
    - Vybrat atributy z dimenzí používané pro seskupení (region, kategorie produktu)
    - Určit jejich kombinace – které atributy jsou používány společně
  - Dle statistické distribuce dat v DW
    - Např. produktu je 1 000 000
    - Kategorií – 500 000 – nemá moc smysl agregovat (stále mnoho řádků)
    - Kategorii – 1 500 – agregovat
    - Počítat frekvence výskytu řádků pro kombinace hodnot atributů dimenzí
      - Např. 12 měsíců x 256 produktů = XY možných řádků v faktové
  - V čase se mění – interaktivní cyklus
    - Mažou nepoužívané, přidávají nové

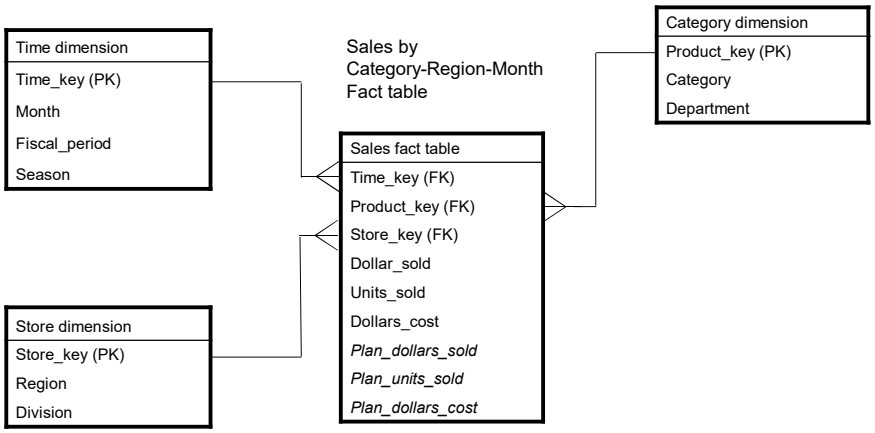
Fyzický design - Agregace

- Všechny agregace v součtu by měly být optimálně asi stejně objemově velké jako původní tabulka
- Každá agregace má svoji faktovou tabulku
  - A zmenšené verze dimenzionálních tabulek (dle úrovně agregace)
  - Některá fakta musí být vypuštěna – mohou dávat smysl pouze na detailní úrovni
- Agregace pak často umožňují přímé porovnání s uloženými daty plánu (plány jsou často na agregované úrovni)
  - Lze uložit do jedné tabulky
- Agregované tabulky mohou být často rozšířeny o fakta typu min_prodej_kč, max_, count_

Fyzický design - Agregace



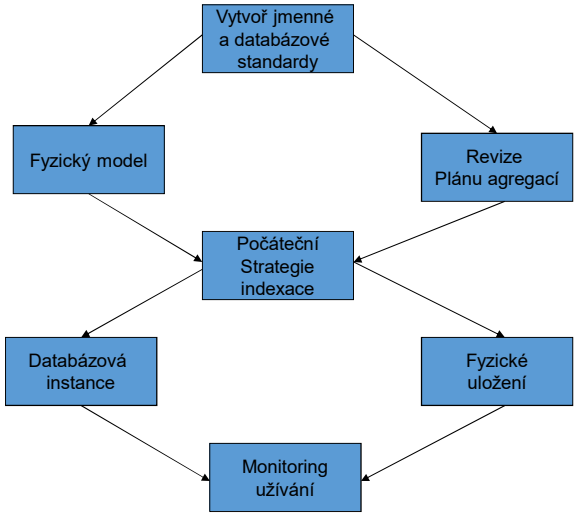
Fyzický design - Agregace



Fyzický design

- Je ovlivněno:
  - Logický datový model
  - Zvolený RDBMS
  - Objemem dat
  - Způsobem využití dat
  - Nástroji pro přístup k datům
- Je třeba:
  - Vytvořit plán fyzické implementace
  - Vytvořit a dodržovat standardy

Fyzický design



Fyzický design

- Vytvořit standardy
  - Jména databázových objektů
  - Jména a cesty k fyzickým souborům
- Lze převzít (a případně modifikovat) již existující firemní standardy
- Příklad konvence – jméno složené ze tří částí
  - Hlavní část – co je to? (např. zákazník, produkt, účet)
  - Třída – typ objektu (např. průměr, počet, datum, flag)
  - Vlastnost – volitelný, popisuje předchozí dvě vlastnosti (např. počátek, konec, primární, sekundární)
  - Konvence: hlavni_vlastnost_třída
    - Ucet_počateční_datum
    - Prodeje_průměr
- Zvážit rozlišit logická a fyzická jména
  - Doporučeno stejné – co nejvíce popisné

Fyzický design - standardy

- Dohodnout se na seznamu slov (Hlavních části) s uživateli
  - Vytvořit seznam tříd a vlastností
- Vytvořit seznam používaných zkratek (např. desc = description)
- Vytvořit standardy pro pojmenování pracovních tabulek pro ETL
- Zvážit poměr mezi popisností názvu a jeho délkou
  - my_company_billingDW_customer_ID
- Pozor zda databáze je case senzitive
  - Brát jako by byla i když není
  - Připraveno pro migraci na novou databázi, která by mohla být

Fyzický design - standardy

- Využít pro názvy tabulek synonyma
  - Je pak jednoduší při změně struktury tabulky jenom změnit odkaz synonyma než měnit aplikace
- Alternativně lze využít view
  - Nebo materializované view
  - Zvážit výkonnost view tvořené z více tabulek
- Vytvořit standardy (adresářovou strukturu a jmenné konvence) pro umístění souboru
  - Zdrojové kódy
  - Skripty
  - Binární soubory
  - Databázové soubory
  - Modely
  - Dokumentace

Fyzický design - standardy

DISK A RAID 1	
Adresáře	
RDBMS	Obsahuje databázi
ETL	Obsahuje ETL nástroj
LOG	Obsahuje log soubory
SCRIPT_PROD	Obsahuje produkční scripty
METADATA	Obsahuje skripty pro metadata
DIMENZE	Obsahuje scripty pro dimenze
ZAKAZNIK	Obsahuje scripty pro dimenzi zákazník
crt_zakaznik.sql	DDL - vytáří tabulku zákazník
crt_cust_stage.sql	DDL - vytáří data staging tabulky pro zákazníka
crx_customer.sql	DDL - indexy nad tabulkou zakazník
drx_customer.sql	DDL - drop indexy
customer_stage.sql	Script pro 0.vrstvu
upd_customer.sql	Script pro načtení do 1. vrstvy
readme	Popis obsahu adresáře
….	
FAKTA	Obsahuje scripty pro faktové tabulky
SCRIPT_DEV	Obsahuje vývojové scripty

DRIVE B RAID 5	
DATABASE	Obsahuje databázové soubory (vlastní data)
DRIVE C NO RAID	
DATASTAGE	Obsahuje flat files
TEMPDAT	Temp místo pro databázi
JOBLOGS	Logy z proběhlých úloh a scriptů

Fyzický design

- Fyzický model vychází z logického
  - Některé změny vlivem zvoleného RDBMS
  - Přidány pomocné tabulky (většinou nejsou součástí logického modelu)
  - Detailní nastavení datových typů, partition, specifikace umístění tabulky, umístění na disku databáze (souborů)
- Vhodné využít modelovacího (case) nástroje
  - Většinou je možné i využít pro tvorbu dokumentace (např. technický popis z jakých zdrojů se plní daný atribut, jaký je typ, transformace, …)
- Definovat entitní, doménovou a referenční integritu, null hodnoty
  - Někdy je výhodnější neimplementovat (když jsou čistá data) – nezatěžuje RDBMS – ale pozor na konzistenci dat

Fyzický design

- Model by měl obsahovat i indexy
- Modifikovat model dle potřeb uživatelských nástrojů
  - Např. vyžadují snow-flake
- Provedení přibližného odhadu velikosti databáze
  - Možno využít schopností modelovacího nástroje
  - „Ruční výpočet“
    - Délka (velikost) řádku
    - Počet řádek, počet řádek přírůstku pro jeden load
    - Pro indexy přidej stejně místa jak pro tabulku
    - Temp space – pro budování indexu musí být dvojnásobný jak index
      - Pro třídění alespoň velký jak tabulka
    - Započítat metadata tabulky
    - Připočítej agregační tabulky (obecně velikost v souhrnu jako base tabulka)
    - Obecně platí, že DW zabere v součtu 3 až 4 tolik místa jak atomická data
    - Potřeba zahrnout i místo pro testování, vývoj, ETL

Fyzický design

- Nejvíce místa zabírají
  - Faktové tabulky
  - Indexy na nich
- Dimenze jsou v porovnání s faktovou tabulkou obecně zanedbatelné
  - Výjimka např. velké dimenze zákazníků

Fyzický design

- Vytvořit počáteční plán indexace
  - Bude se měnit v průběhu využívání DW – podle analýzy dotazů, doby odezvy, …
- Potřeba porozumět jak zvolený RDBMS využívá indexy a jak tvoří plán provedení dotazu
- B-tree indexy
  - Pro atributy s velku kardinalitou (např. customer_ID)
  - Klastrované vs. Neklastrované
  - Na jednom nebo více sloupcích
- Bitmapové indexy
  - Pro atributy s nízkou kardinalitou (např. pohlaví)
- Některé RDBMS disponují speciálními indexy
- Některé mají zabudovanou podporu star schéma (násobné join)

Fyzický design

- Faktová tabulka
  - Indexy na klíčích
    - Jeden pro jeden klíč – RDBMS podporuje využití více indexů pro jeden dotaz
    - Několik složených indexů – podle cesty dotazu
  - Lze v případě potřeby i indexy na faktech (když hodně dotazů typu prodej > 1000)
- Dimenzionální tabulka
  - Na primárním klíči
  - Bitmapové indexy na atributech dimenze
  - Na atributech sloužících pro join, filtrování, group by

Fyzický design

- Nastavit a zvážit indexy i pro efektivní ETL
- Při loadu dat
  - Zvětší-li se loadem tabulka o 10 až 20 procent je efektivnější smazat a znovu vytvořit index
- Kontroluj indexy a statistiky po loadu

Fyzický design

- Instalace a nastavení databáze
  - Zdokumentovat nastavení databáze plus důvod
- DW je náročný na paměť
- Blocksize
  - Záleží na potřebách
- Uložit scripty pro nastavení databáze
- Nastavit partition tabulek
  - Většinou podle atributu datum
- Nastavit umístění souborů na disku
  - Doporučuje se využít RAID disky (RAID 1 až 5)
  - Optimálně databáze a OS jeden disk, zdrojová data (flat soubory) druhý disk, tabulky a indexy další dva disky, transakční log další disk, temp další disk
  - Z hlediska dotazů je vhodné aby fakt na jednom disku a dimenze na jiném

Fyzický design

- Potřeba vybudovat systém pro monitoring využití DW
  - Load dat
  - Dotazy
  - Běh procesů
  - Využití zdrojů
- Důvody pro monitoring:
  - Výkonnost
    - Ladění dotazů, indexy, agregace
    - Výběr testovacích dotazů
  - Podpora uživatelů
    - Sledovat vytíženost, logování uživatelů
    - Pro plánování školení
    - Proč se uživatel už dlouho nepřihlásil – neví jak, nemůže se připojit k databázi
  - Marketing
    - Že DW je stále více využíván
    - Kdo z uživatelů využívá nejvíce – konkurence mezi uživateli
  - Plánování
    - Další rozvoj DW dle vzrůstajícího počtu uživatelů, konkurenčních dotazů, času loadu, velikosti databáze,….