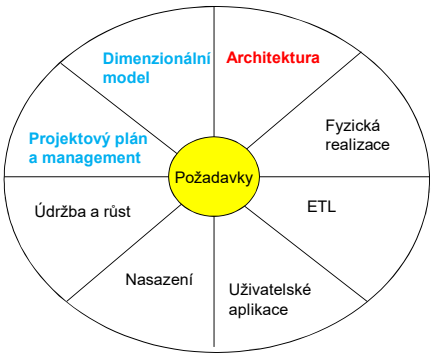
Databázové systémy a metody zpracování dat

Architektura – 1.část

7.přednáška

Architektura



Architektura

- Vždy je nutno stanovit plán (projekt)
  - „Nelze stavět dům bez plánu“
  - DW je finančně náročný projekt, nutno mít vše podchyceno v návrhu architektury
- Návrh architektury slouží
  - Pro komunikaci v rámci týmu
  - Určení plánu
  - Podklad pro učení nových členů týmu
  - Zaručuje flexibilitu a snadnost údržby
  - Pro znovu použití (na dalších projektech)

Architektura

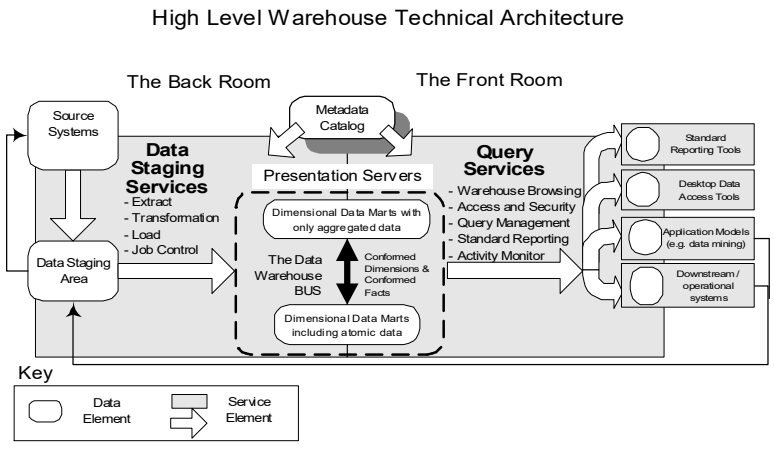
- Využít Top-Down přístup
  - Od obecného k detailními
- Architektura musí vycházet z uživatelských požadavků

Architektura

Úroveň	Data (co)	Technologie (jak)		Infrastruktura (kde) (bezpečnost, síť, metadata, …)
		ETL	Prezentace	
Uživatelské požadavky a audit	Jaké informace potřebujeme k rozhodování? Jak zapadají do matice BUS architektury (DM a dimenze)?	Jak můžeme získat data, jak je transformovat, jak je dostat k uživatelům? Jak se to děje dnes?	Jak chceme analyzovat data? Jaké jsou hlavní business otázky? Jak měříme výkonnost firmy?	Jaké HW a SW komponenty potřebujeme k úspěchu? Jaké používáme dnes?
Model a Dokumentace architektury	Dimensionální model: Jaké jsou hlavní fakta a dimenze které chceme sledovat? Jaké jsou vztahy mezi nimi? Jak tyto entity mají být strukturovány?	Jaké specifické komponenty potřebujeme k dostání dat v užitečné formě na potřebné místo v potřebném čase? Jaké budou hlavní úložiště dat a kde budou situovány?	Co budou uživatelé potřebovat k získání informací v užitečné podobě? Jaké druhy analýz a reportingu budeme potřebovat? Jaké jsou priority?	Odkud a kam proudí data? Máme kapacity na přesun dat a na jejich uložení?Co za specifické komponenty potřebujeme? Kdo za ně bude odpovídat?
Detailní model a specifikace	Logický a fyzický model: Jednotlivé atributy, datové typy, domény, pravidla pro odvození atributů. Jak se zdroje dat mapují na cíle dat?	Jaké produkty podporují potřebné vlastnosti? Jak je lze integrovat? Standardy pro psaní kódu, jmenné konvence, …	Jaké jsou požadavky na reporty (šablony) – pro řádky, sloupce, hlavičky, filtry, …? Kdo potřebuje jaké? Jak často? Jak je třeba je distribuovat	Jak je možné integrovat nástroje? Jak volat jejich utility, API, …?
Implementace	Vytvořit databáze, indexy, backup, … Dokumentace.	Napsat ETL kód. Automatizace procesu. Dokumentace.	Implementace reportingového a analytického řešení, vytvořit první sadu standardních reportů, školení uživatelů. Dokumentace.	Instalace a testování komponent infrastruktury. Propoj zdroje dat a DW s uživateli. Dokumentace.

Architektura

- Základní framework architektury (logický model)



Architektura

- Architektura by měla být řízena metadaty (metadata catalog)
  - Poskytuje parametry a informace pro všechny procesy (ETL, reporty, …)
  - Informace o DW, zdrojových systémech, ETL, …
  - Např. procedury pro zakládaní tabulek, indexů, uživatelů, spuštění skriptů, … (informace o tabulce a atributech v metadata katalogu, …)
  - Umožňuje např. rychlou změnu při upgrade provozních systémů, …

Architektura

- Data staging area (DSA) je vhodným místem pro archivování dat z provozních systémů
- Datový model DSA by měl být navržen tak, aby podporoval výkonnost ETL procesu a snadnost vývoje
  - ERD modely (kopie OLTP systémů)
  - Dimenzionální model
  - Pomocné tabulky
  - DSA – převod ERD modelů do dimenzionálních určených k načtení do 1. vrstvy DW (prezentační část)
  - Otázka zpracování dat ze zdrojů uložených pomocí RDF
  - Integrace různých schémat, popř. jak zakomponovat data bez schémat

Architektura

- Nástroje ETL lze získat dvojím způsobem
  - Samostatný vývoj
  - Zakoupení hotového nástroje
- Komerční ETL nástroje jsou často poměrně drahé, ale zaručují vysokou produktivitu (při správném využití)
  - Je doporučeno první etapu udělat ručně
  - Po schválení výsledků lze použít ETL nástroj

Architektura

- ETL proces – zabírá až 60 procent celkového vývojového času na projektu
- Potřeba vybrat a integrovat data
- Problémy
  - Různé názvy sloupců
  - Chybné hodnoty
  - Neexistující referenční integrita
  - …

Architektura

- Možnosti získat data
  - Export z provozního systémů (flat file)
    - Vhodné dohodnout rozhraní a přenechat odpovědnost na provozovateli OLTP systému
  - Přímé napojení
  - Replikace
  - Někdy je vhodné data komprimovat pro přenos a zase dekomprinovat

Architektura

- Typy extrakce
  - Inkrementální load
    - Načtení dát nových (změněných) od posledního loadu
    - Nová (změněná) data identifikována většinou flag v provozním systému nebo dle datumu (času) transakce, trigery, log file, …
    - Datum (čas) transakce posledního loadu nutné uložit do metadata katalogu
    - Většinou tak načítáme velké fakt tabulky
  - Full load
    - Malé dimenze
    - Při sledování SCD je nutné porovnat s původní dimenzí
    - I v případě velkým faktových tabulek – jestliže není možné identifikovat v provozním systému co se změnilo od posledního loadu

Architektura

- Po načtení dat následuje transformace
- Transformace zahrnuje:
  - Integraci
  - SCD
  - Kontrola referenční integrity
  - Denormalizace
  - Čištění dat, spojování (merge), rozpojování
  - Konverze datových typů (např. ASCII, EBCDIC)
  - Odvození nových atributů, alokace hodnot na nižší granualitu (např. náklady) – využití business pravidel
  - Agregace
  - Audit obsahu dat (kontrolní součty, počty řádků, …)
  - Plnění auditních tabulek a metadat
  - Transformace dat pro specifické potřeby (např. pro Data miningové nástroje)
  - Null hodnoty – identifikace (např. speciálních kódu pro null hodnoty), nahrazení

Architektura

- Během načítání dat je třeba kontrolovat a monitorovat:
  - Naplánování úloh
  - Monitorovat proces načtení – ukládat potřebné informace např. pro možnost recovery (znovunačtení)
  - Řízení výjimky a špatná data identifikovaná při procesu ETL (např. duplicity)
    - Exportovat, předat k vyřešení, načíst
    - Identifikovat odpovědnou osobu za data (kvalitu)
  - Navrhnout postupy řešení chyb při načítání (např. výpadek proudu, DB, …)
  - Zaslání upozornění na chyby při loadu (nečekat až do rána, že neproběhlo)

Architektura

- Potřeba nastavit způsoby zálohování a recovery
  - Potřeba zajistit rychlost a jednoduchost zálohování a recovery
- Archivace
  - Fyzická – databáze
  - Logická – modely, skripty, …
- Seznam všech tabulek v DW (0. i 1. vrstva)
  - Zhodnotit jak náročné je bude obnovit
  - Které by nešly (obsahují historická dat)
  - Vybrat a rozhodnout co zálohovat
  - Nastavit procesy
  - Vyzkoušet recovery

Architektura

- Otázky bezpečnosti nejsou v 0. vrstvě (ETL) kritické – není-li určena pro dotazy
  - Pozor při přenášení dat po sítí při loadu

Architektura

- V metadata katalogu je vhodné udržovat metadata pro uživatele při získávání dat z DW
  - Kde co naleznou, popis, v jakých reportech se vyskytuje, …
- V prezentační vrstvě je třeba zajistit bezpečnost dat
  - Authentication
    - Systém hesel, fyzické kontroly (otisky prstů, rohovky), omezení přístupu na základě IP adres, …
    - Bezpečnost přístupu k databázi, reportům, exportovaným souborům
  - Authorization
    - Co kdo může vidět – někdy vyžaduje hodně práce zabezpečit

Architektura

- Monitorovat využití DW
- Využití pro:
  - Zvýšení výkonnosti (agregace, indexy, view, …)
  - Podporu uživatelů, školení
  - Marketing DW
  - Plánování (rozvoje, upgrade architektury, …)
- Dobré reportovací nástroje by měly podporovat
  - Uživatelskou formulaci dotazů (business vrstva mezi daty a uživatelem)
  - Podporu využití agregací
  - Rozklad složitých SQL na více dotazů a jejich propojení ve výsledků
  - Tvorbu odvozených atributů
  - Ochranu před složitými dotazy, které mohou „položit“ DW

Architektura

- Reportingové nástroje
  - Desktop
  - Aplikační server
  - Součást RDBMS
- Reportingové řešení by mělo podporovat
  - Uživatelskou tvorbu reportů
  - Reportovací server
  - Parametrizované reporty
  - Plánovač plnění reportů (scheduler)
  - Interaktivní reporty
  - Propojení reportů
  - Report delivery (pull, push, email, web, file system)
  - Metadata
  - Pokročilá prezentace (např. barevně označení položek nad určitý limit)
  - Podpora práce s semiaditivními ukazateli
  - Přímý zápis SQL plus SQL builder
  - API

Architektura

- Dále:
  - Calculace položet (if then, …)
  - Pivoting (cross tab)
  - Sorting
  - Kompletní formátování (mix tabulek, obrázků, grafů, report s více sekcemi, …)
  - Podpora grafů
  - Export výstupů (excel, word, …)
  - Uživatelsky přívětivý
  - Drop and down – tvorba reportů
  - Multitasking (nemuset čekat až report doběhne)
  - Cancel query
  - Podpora přístupy k různým zdrojům
  - Bezpečnost
  - Administrace (desktop, web)
  - …

Architektura

- Existuje několik skupin uživatelů z hlediska jejich technických znalostí

Oblast	Typ uživatele			
	Papírový uživatel	**Tlačítkový uživatel**	**Jednoduchý ad-hoc**	**Pokročilý uživatel**
Používání počítače	Žádná	E-mail, Word	Word, Excel, PowerPoint	Makra, tvorba WWW
DW	Spoléhá na pomoc ostatních	Standardní reporty, default parametry, EIS	Vytvoří jednoduché dotazy (QBE), modifikace existujících dotazů, změna parametrů, navigace v hierarchiích	Tvorba dotazů, přímý přístup k databázi

Architektura

- Informační požadavky se rovněž liší

Informační potřeby – kategorie	Role uživatele	Přístup k datům - kategorie	Nástroje	Počet osob
Monitoring na vysoké úrovni - klíčové metriky, flags	Senior management	Přístup na "tlačítko"	EIS - styl nástrojů	Malý
Sledování businessu - trh, produkty, zákaznicí, drill-down k detailu	Střední management, marketing, prodejci, podpora zákazníků	Standardní reporty – parametrizované	OLAP style, reportingové nástroje, časování tvorby	Velký
Průzkum . Výjimky, nové problémy a příležitosti, vývoj obchodních případů (business case)	Jako předchozí plus business analytici	Ad hoc analýza	OLAP style, reporting, analytické nástroje pro ad-hoc dotazování	Střední
Kompletní analýza – kombinace dotazů, statistická analýza, vývoj modelů	Business analytici a experti na analýzu	Data mining - pokročilé analýzy	Statistické nástroje, data mining nástroje, pokročilé analytické nástroje	Malý

Architektura

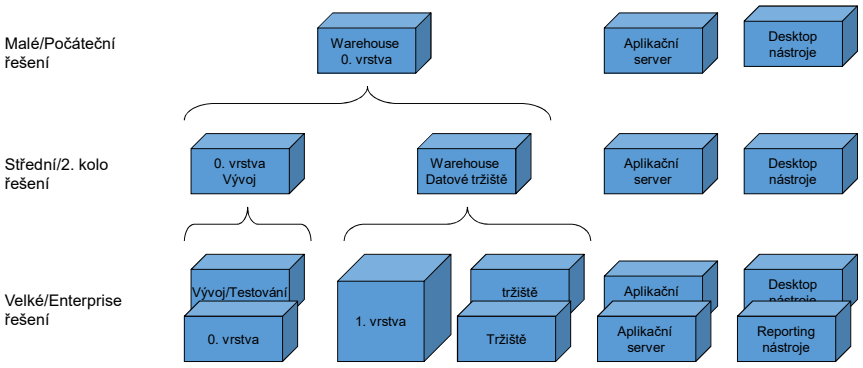
- Potřeba poznat potřeby uživatelů -> navrhnout správné nástroje pro přístup k datům
- Vycházet z priorit nejdůležitějších uživatelů DW
- Vhodné využít web
  - Např. pro přístup k metadatům, reportům, …

Architektura

- Otázky infrastruktury je vhodné řešit s expertem z dané firmy (kde se buduje DW)
- Navržení HW
  - DW většinou prudce roste v prvních dvou letech (objem dat, dotazy)

Architektura

- Existuje mnoho variant řešení:



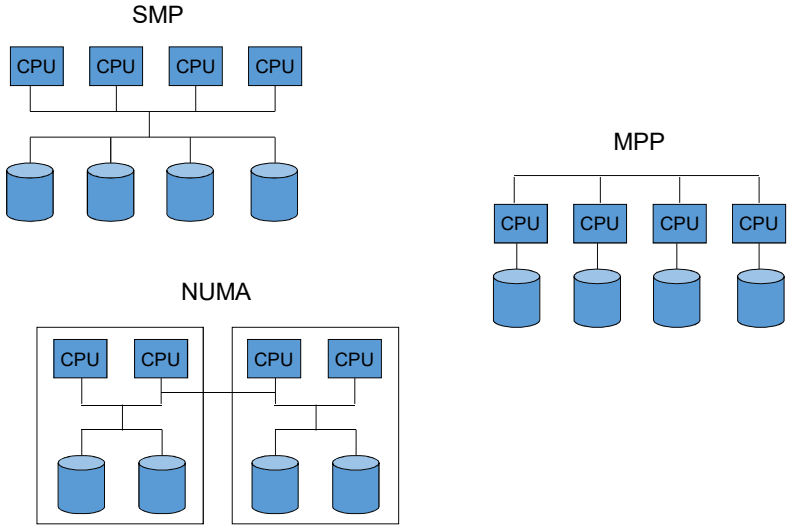
Architektura

- Třeba vzít v úvahu business požadavky
- Faktory
  - Velikost dat
  - Přírůstky, změny v datech, frekvence loadů, čas pro load
  - Počet uživatelů, jejich aktivita, typy analýz, kolik pracuje konkurenčně, rozložení časové zátěže
  - Složitost business – řešení
  - Požadavky na dostupnost (regionální, časovou)
  - Znalosti IS pracovníků (umí UNIX, Windows, DBA, …)
  - Finanční zdroje

Architektura

- Využití paralelního zpracování
  - Symmetric multiprocessing (SMP)
    - Jeden počítač, sdílejí disk, paměť
  - Massively parrallel processing (MPP)
    - Více počítačů, každý paralelně, např. tabulky přes několik počítačů, full table scan paralelně
  - Non-uniform memory architecture (NUMA)
    - Kombinace SMP a NUMA
- Musí to HW a SW (OS, RDBMS) podporovat
- Při výběru OS zvážit i jaké jsou pro něj dostupné aplikace

Architektura



Architektura

- HW (nutno konzultovat s odborníky ze zdrojových systémů):
  - Disky (velikost, rychlost, RAID 1-5)
  - Paměť (čím více tím lépe)
- Nezapomenout na zdroje pro back-up
- Volba
  - OS:
    - UNIX
    - Windows
  - RDBMS
  - OLAP databáze
  - ETL nástroje

Architektura

- RDBMS – hlavní faktory pro uvažování:
  - Podpora DW
  - Bitmapové indexy
  - Optimalizátor dotazů
- Vhodné je se poučit z jiných DW projektů nebo dopředu otestovat výkonnost RDBMS
- Existují speciální relační databáze určené pro DW a ne pro transakční
  - Sybase IQ

Architektura

- Pro front-room nástroje zjistit nároky:
  - Paměť
  - Disk
  - Platforma
- U desktop aplikací vyřešit problém upgrade na nové verze, OS desktop stanic, paměť
  - Záleží na potřebách uživatele co bude za aplikace používat
  - Stanovit minimální požadavky dle skupin uživatelů (pasivní, analytici, …)
  - Pozor na bezpečnost při systémové integraci

Architektura

- Potřeba zvážit nároky na síť
  - Kapacita, zatížení, množství přenesených dat
  - Bezpečnost – přenos přes síť (ssh)
- Databázová konektivita
  - Nativní ovladače
  - OBDC
  - OLE DB
  - JDBC
- Využití DNS serveru (ne přímý zápis IP adres), Active directory - metadata

Architektura

- **Metadata – data o datech**
  - **„Je to užitečná definice, ale je nutno hlavně chápat význam.“**
- **Metadata**
  - **Pro ETL**
  - **Pro prezentaci**
- Ohledně metadat je třeba:
  - Sepsat všechna metadata
  - Určit jejich důležitost
  - Určit odpovědnou osobu
  - Rozhodnout zda není vhodné koupit speciální nástroj na podporu metadat
  - Určit jejich uložení a back-up a recovery
  - Vystavit je uživatelům (procesům), kteří je potřebují
  - Zaručit jejich kvalitu a aktuálnost
  - Kontrola

Architektura

- Metadata:
  - Zdrojové systémy
    - Umístění, modely, formát, vlastník, popis, frekvence změn, limity použití, dostupnost, způsob přístupu, práva přístupu, heslo, privilegia, extrakty z provozních systémů (formát, čas, zodpovědnost)
  - ETL
    - Plánování (schedule) ETL pump, výsledky pump, jakých dat se týkalo (time stamp – pro inkrementální load), čas loadu, definice tabulek (dimenze, fakta, pracovní), skripty – popis, SCD pro každý atribut, umělé klíče pro produkční klíče (mapování), specifikace postupů čištění dat, transformace - popis, datový model, skripty – popis, agregace popis, auditní dimenze, nastavení bezpečnosti, specifikace back-up, recovery
  - RDBMS
    - Model (logický, fyzický), partition, indexy, disk využití, bezpečnost, uživatelé, práva, view definice, uložené procedury, administrační skripty, back-up, recovery

Architektura

- Metadata
  - Prezentace
    - Business jména sloupců, tabulek, hierarchií, reporty, business skupiny požadavků (marketing, obchod, …), uživatelská dokumentace, bezpečnost – privilegia, počty login, délka, počet dotazů, nejvíce využívané reporty (obecně, uživatelem)
- Metadata – efektivní využití DW je založeno na metadatech

Architektura

- Je potřeba zajistit publikaci dat při striktním dodržování ochrany dat a bezpečnosti
- Bezpečnost:
- Fyzické komponenty
  - Servery, stanice, kabely, switch, …
  - Nebezpečí
    - Krádež
    - Zničení/vandalismus
    - Oheň
    - Vlhkost
    - Voda
    - Prach
    - Slunce, chemie
    - Elektrické výpadky, zkraty
    - Magnetismus

Architektura

- Informace: data, finanční dopady, reputace
  - Data, metadata, emaily, dokumenty, kopie, zálohy
  - Nebezpečí
    - Vyzrazení strategie, plánů, rozpočtu, …
    - Vyzrazení citlivých dat (čísla účtů, stavy účtů, …)
    - Modifikace dat
  - Hrozby
    - Odposlech
    - Hacker útoky
    - Fyzická krádež (notebook, dokumenty)
    - Sociotechnika
    - Trojské koně

Architektura

- Software
  - Nebezpečí
    - Krádež kódu
    - Krádež softwaru
    - Viry a další škodlivé kódy
- Ataky na business
  - Nemožnost poskytovat služby (např. zahlcení serveru)
  - Nemožnost rekonstruovat data (např. objednávku)
  - Terorismus

Architektura

- Zajistit bezpečnost přenosu informací přes internet
  - Kryptování
    - Symetrické
    - Asymetrické
- HW prvky k využití:
  - Routery (pro filtraci paketů)
  - Firewall
- Důležitý systém hesel
  - Neodhadnutelný
  - Odolný proti útoku hrubou silou (slovník slov)
    - Heslo – písmena a číslice
  - Člověk – neopatrným jednáním přestavuje podstatné nebezpečí pro bezpečnost
  - Biometrické ochranné prostředky (otisky, sítnice, průsvit prstu)

Architektura

- Bezpečnost je proces
  - Potřeba zakomponovat do firemní kultury
- Je potřeba
  - Definovat bezpečnostní politiku
  - Získat podporu vedení
  - Školit a stále připomínat
  - Být ostražití, stále aktualizovat bezpečnostní pravidla
  - Nedůvěřivý (Proč chtějí uživatelé vidět tyto data, pozorovat log, …)

Architektura

- Kroky
  - Taktické
    - Využití antivirových programů (server, desktop), pravidelná aktualizace
    - Instalace firewallu, zákaz připojení přes modem ze stanic uživatelů, nastav bezpečnou komunikaci na internet (přes proxy server)
    - Kontrola SW instalovaného na počítačích
    - Nastavení a kontrola politiky hesel
    - Školení uživatelů v otázkách bezpečnosti
    - Nastavení kontrolních mechanizmů – kontrola logů, chybných přihlášení
    - Zabezpečení back-up media
    - Kontrola nastavení práv uživatelů
    - Fyzicky zabezpečené servery
    - Skartace starých záloh, …

Architektura

- Strategické
  - Nezávislý bezpečnostní audit
  - Využití asymetrického kódování
  - Nahrazení hesel pokročilými technikami (otisky prstů, sítnice, magnetické karty)
- Z pohledu DW – využít bezpečnostní politiku, kterou firma implementovala dosud

Architektura

- Návrh a implementace architektury
  - Vycházet z principu 80-20
  - 20 procent úsilí přinese 80 procent přínosu
  - Nastavit priority a na ty se zaměřit
  - Interaktivní proces – zpřesňování požadavků a návrhu architektury