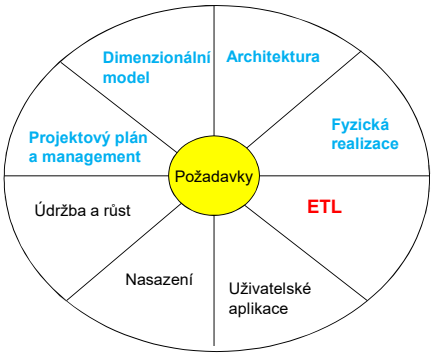
Databázové systémy a metody zpracování dat

Proces ETL (Extrakce, Transformace, Loading)

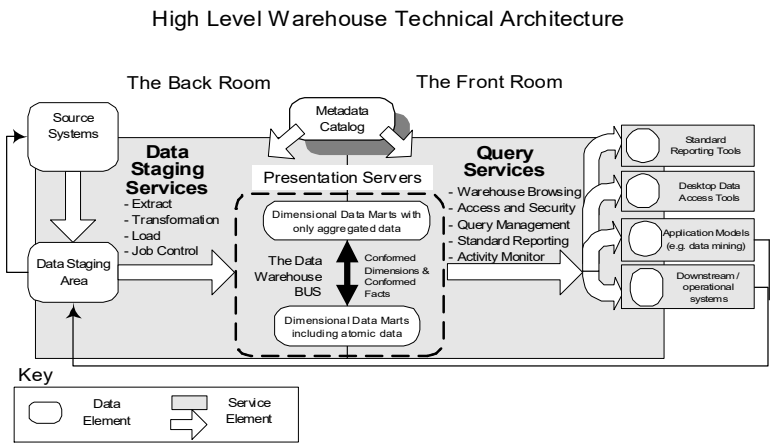
9.přednáška

Proces ETL (Extrakce, Transformace, Load)



Architektura

- Základní framework architektury (logický model)



ETL - datová pumpa

- Informace se do datových skladů ukládají pomocí datových pump z provozních databází (tím jsou myšleny relační databáze z podnikových informačních systémů–ERP, CRM atd.).
- Nástroje datové pumpy se také označují jako ETL nástroje(zkratka slov „Extraction“, „Transformation“ a „Loading“).
- Příkladem nástroje ETL (datové pumpy) může být například DTS (Data Transformation Services) firmy Microsoft (je součástí instalace MS SQL Serveru) nebo Oracle Data Mart Builder.

ETL

- Hlavní základní část dobře fungujícího DW
- V prvním kroku je třeba vytvořit plán ETL
  - Plán:
    - Konceptuální model zdroj-cíl proudění dat (na jednu stránku)
    - Testovat, implementovat nástroj pro ETL nebo využít SQL
    - Graficky zobrazit všechny komplexní transformace, generování umělých klíčů, SDC. Vytvořit prvotní plán sekvenčních kroků
  - Dimenze
    - Vytvoření a testování statického load dimensionální tabulky. Primárním cílem je otestovat infrastrukturu (připojení, přenos souborů, bezpečnost – práva)
    - Vytvoření a testování SCD procesu pro jednu dimenzi
    - Vytvoření load pro zbývající dimenze
  - Fakta a automatizace
    - Vytvoření load historických dat do faktových tabulek (zahrnuje management umělých klíčů, a jejich substituci)
    - Vytvoření a testování inkrementálního procesu
    - Vytvoření a testování load agregací nebo load do OLAP vrstvy
    - Vytvoření a testování automatizace celého procesu

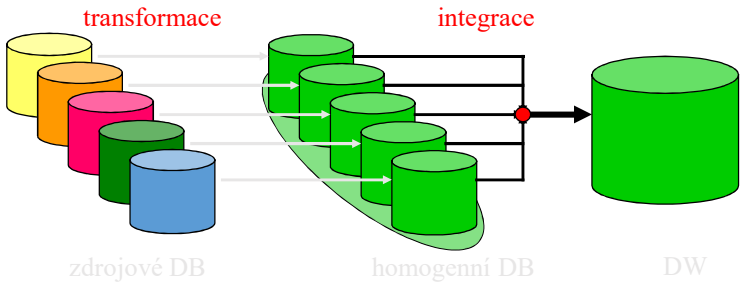
Vrstvy datového skladu

- Datový sklad se většinou staví ve vrstvách:

Vrstva	Popis	ETL náročnost
0. vrstva	V nulté vrstvě se uchovávají data z jednotlivých provozních databází. Jedná se většinou o kopie provozních dat 1:1. Data neslouží přímo pro analýzy, ale jako vstup pro další vrstvy.	Převod dat v zásadě 1:1, základní transformační kroky.
1. vrstva	Data jsou uložena v datovém modelu datového skladu (tabulky fakt a dimenzí). Na data jsou aplikována integritní omezení. Tato vrstva slouží pro analýzy. Data jsou očištěná a konzistentní.	Náročné transformace a čištění dat, mapování dat z 0. vrstvy na datový model datového skladu.
2. vrstva	Speciálně připravená data pro podporu speciálních aplikací. V podstatě se jedná o jednotlivá datová tržiště.	Náročné transformace z 0. a 1. vrstvy (speciální algoritmy).

ETL

- Obecně je nutné vyřešit dva základní problémy:
  - Transformace z různorodých zdrojů
  - Integrace data do datového skladu



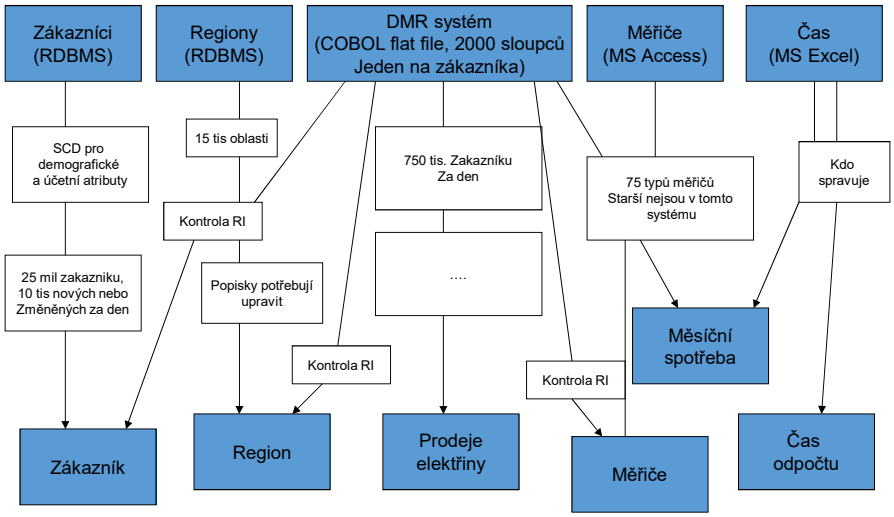
ETL

- Je třeba zajistit dostupnost zdrojových systémů pro potřeby loadu
  - Přímé napojení
  - Extrakty s danou strukturou
- Vychází se z vytvořeného dokumentu mapující vazby zdroj – cíl
- Je potřeba
  - Dodržovat standardy (jmenné, psaní kódu)
  - Psát srozumitelné komentáře
  - Hlavičky skriptů
  - Vytvořit knihovny obecně využívaných funkcí
  - Testovat funkčnost
  - Dokumentace

ETL

- Krok 1 – konceptuální model ETL
  - Základní, na jednu stránku
  - Mapování zdroj – cíl, poznámky k hlavním bodům
  - Je-li jeden hlavní systém – zdroje logické seskupení zdrojových tabulek
  - Tři základní fáze ETL
    - Extrakce – ze zdrojových dat
    - Transformace
    - Load – do 1. vrstvy DW

ETL



ETL

- Krok 2 – nástroj na ETL
  - Možnosti řešení
    - Kód
      - T-SQL, PL/SQL, Delphi, …(programováno v různých prostředích, mnohdy historických)
    - Nástroj
      - Grafické rozhraní, zrychlení procesů
      - Repositář dat, paralelismus
      - Dražší řešení (cenově, další náklady)
      - Zástupci: Informatika, DTS, warehouse Builder
  - Většinou na první fázi dělat manuálně („ručně“)
    - Nezvyšovat náklady na DW
    - Záleží na podmínkách

ETL

- Krok 3 – detailní plán
  - Detailně rozpracovat jednotlivé kroky
  - Rozhodnout se nad sekvencí kroků
    - Nejprve dimenze
    - Pak fakta (plus look-up na umělé klíče)
  - Zde jedna (i více) stránka pro jednu tabulku v DW
    - Někdy smysl vycházet ze zdrojové tabulky
  - Doplnit pseudokódem pro transformaci
  - 0. vrstva = DSA - data staging area
    - Místo, kde jsou načtená data čištěna, kombinována, archivována, transformována a přenášena do prezentačních vrstev (1. vrstva DW)

ETL

- Dimenze
  - Statická
  - SCD
  - Umělé – nejsou v datech (Časová dimenze)
- Krok 4 – naplnit jednoduchou statickou (ne SCD) dimenzionální tabulku
  - Load
    - Ze souboru – výhoda, že soubory lze zálohovat a tak znovu použít při recovery, lze je při přenosu kryptovat, zapakovat
    - Přímé napojení - stream

ETL

- Krok 4 – pokračování
  - I jednoduchá tabulka potřebuje
    - Čištění dat
    - Přiřazení umělých klíčů
      - Je třeba uchovávat tabulku mapování přirozených klíčů na umělé
      - Využívá se později i pro faktové tabulky
      - Obvykle umělý klíč – integer
      - Možno využít sekvencí
  - Hlavní transformace – konverze datových typů, kódování – čeština
  - Jsou-li zdrojová data pro dimenzi z více zdrojů (např. zákazník)
    - Potřeba namapovat na sebe
    - Někdy těžko lze – fuzzy logika (jméno, adresa, …)
    - Existují na to nástroje
    - Uložit do tabulky umělých klíčů přirozené klíče ze všech spojených záznamů (ze všech zdrojů)
  - Testovat zda vztahy mezi atributy dimenze jsou opravdu 1:1 nebo 1:N

ETL

- Krok 4 – pokračování
  - Pro load dat využít bulk funkci
    - I pro load do 1. vrstvy je možné
    - Insert into je pomalé a zapisuje se do log souboru – problém při velkých loadech
    - Pro load dat do prezenční vrstvy – doporučení
      - Vypnout loggování
      - Pre-sort data dle primárního klíče (rychleji se načte při indexu na primární klíč)
      - Při full refresh tabulky – smaž původní data pomocí truncate table
        - Nelogguje se
      - Přidává-li se více jak 10 procent dat do tabulky je smysluplné drop a recreate index (záleží na podmínkách, počtu indexů, …)
      - I při ponechání indexů je dobré je časem přebudovat pro zamezení přílišné fragmentace (fillfactror na maximum)

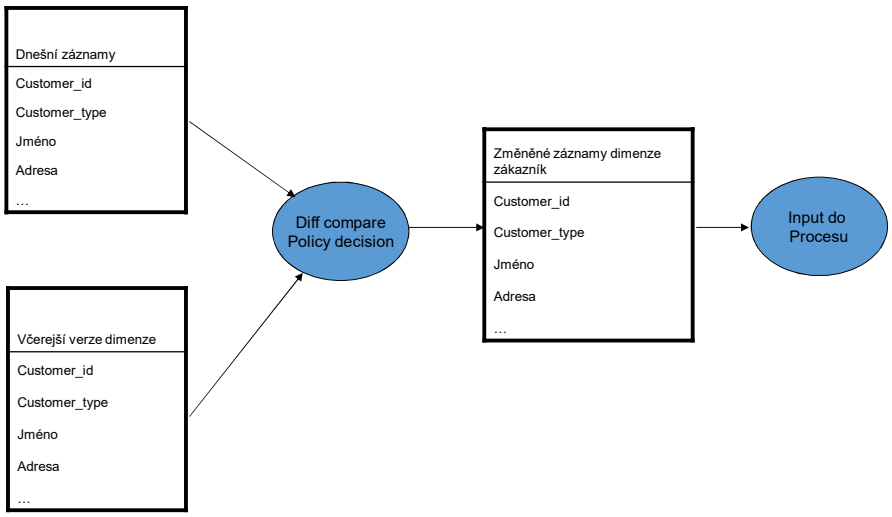
ETL

- Krok 5 – SCD
  - Hlavně se používá technika Typ 2
  - Přístup
    - Načíst všechna data a dívat se co se změnilo
    - Využít inkrementálního načítání (načíst jen změny od minulého loadu) – viz dále u faktové tabulky
      - Hlavně u velkých tabulek
      - SCD typ 1 – je vlastně inkrementální načtení dimenzionální tabulky (jako full ale inkrementálně)
    - Problém s položky smazanými v OLTP – byly smazány tak nemohou být načteny – pro DW to ale potřebujeme vědět
      - Triggery
      - Změny v OLTP – nemazat, deaktivovat
    - Výhodně je načítat jen změny
      - A ještě více pokud OLTP udržuje informaci o typu změny

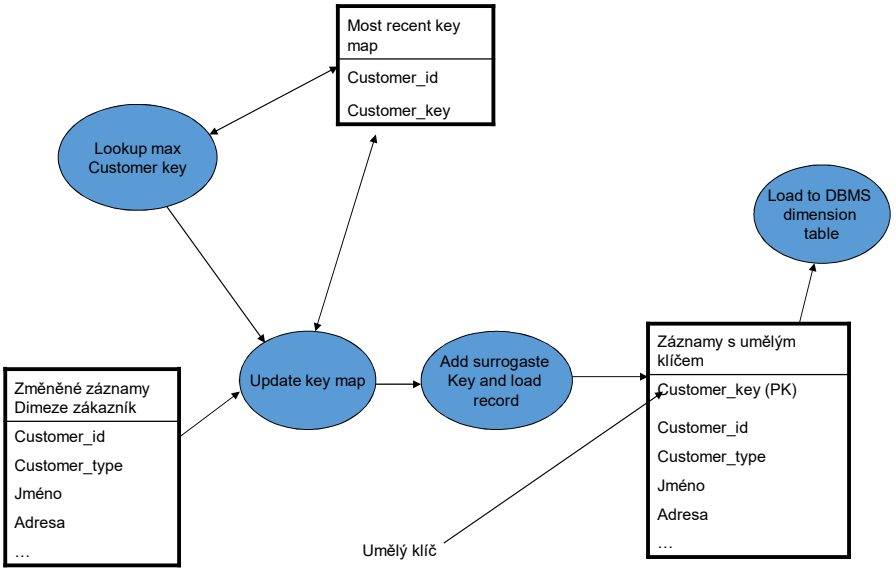
ETL

- Krok 5 – pokračování
  - Identifikovat změny
  - Porovnat s stávající dimenzí
    - Neexistuje-li záznam – vložit
    - Existuje-li – SCD (dle typu SCD)
  - Jestliže v dimenzi některé položky mají SCD typ 1 a některé SDC typ 2 musím u položek s typem 1 každou změnu promítnout do všech záznamů pro daný objekt v dimenzionální tabulce
  - Velké dimenze se zpracovávají podobně jako faktové tabulky

ETL



ETL



ETL

- Krok 6 – naplnit zbývající dimenze
  - Obdobně jako předchozí načíst i další tabulky
  - Vytvořit skript pro načtení Časové dimenze
    - Někdy se využívá tabulkový procesor

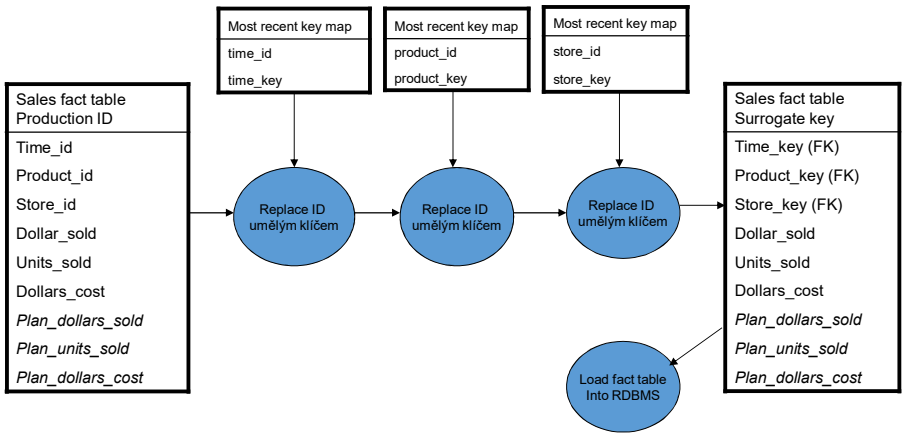
ETL

- Faktové tabulky
  - Výhodné načítat inkrementálně
    - Jen záznamy změněné od posledního loadu
  - Podobně i velké dimenze
- Krok 7 – načtení historických údajů
  - Vytvořit pumpy pro načtení historických dat
  - Interaktivní proces – nebude pravděpodobně dobře na první pokus
    - V praxi řada výjimek (co započítat, co ne, …) – potřeba identifikovat tyto business pravidla
    - Potřeba auditovat součty, počty, … a porovnat s výstupy z provozních systémů – zvyšuje důvěryhodnost DW

ETL

- Krok 7 – pokračování
  - Nahrazení přirozených klíčů umělými
  - Vhodné je zjišťovat aktuální umělé klíče ze speciálních tabulek, ne přímo z dimenzí (může být pomalé)
    - Lze řešit v dimenze flagem – aktuální záznam a bitmapovým indexem nad ním
  - Jestliže se načítají faktová tabulka historicky musí se i historicky přiřadit umělé klíče (platné v čase transakce zachycené v faktové tabulce), ne vzít aktuální umělý klíč!!!

ETL



ETL

- Krok 7 – pokračování
  - Je-li null hodnota v klíči do dimenzionální tabulky – nahradit klíčem k speciálnímu záznamu v dimenzionální tabulce („Neuvedeno“)
  - Odvozená fakta je možné uložit fyzicky (jsou-li často přistupována, chceme index nad nimi) nebo vypočítáme až ve view

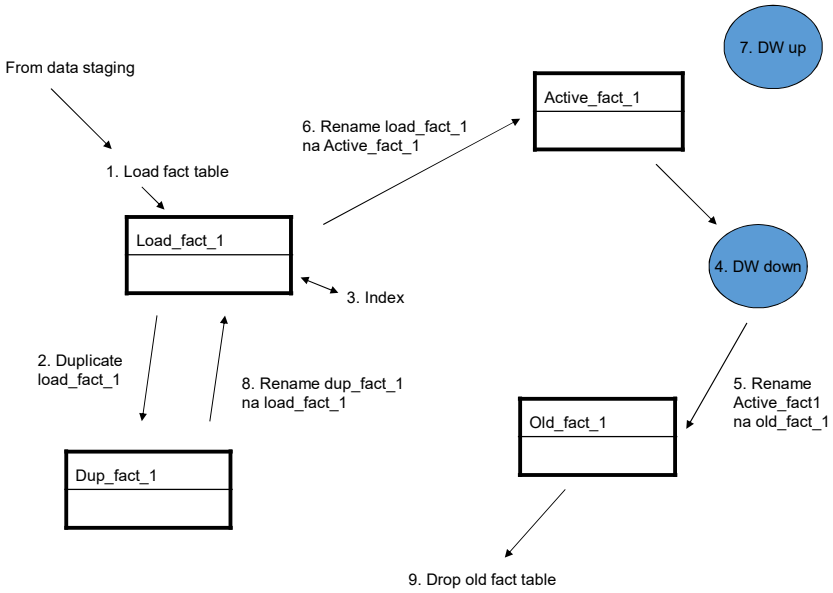
ETL

- Krok 8 – inkrementální načtení
  - Identifikovat, co se stalo nového
    - Nová transakce
      - Přidat záznam do fakt tabulky
    - Update transakce
      - Změnit záznam ve fakt tabulce
      - Přidat změnový záznam
    - Smazání transakce
      - Smazat záznam ve fakt tabulce

ETL

- Technika pro zajištění maximální dostupnosti DW
- Vhodná pro menší DW
  - Budou existovat 3 kopie faktové tabulky

ETL



ETL

- Krok 9 – Agregace a OLAP
  - Agregace – někdy třeba rovněž vytvářet inkrementálně (full proces moc náročný časově)
  - Je-li agregován čas (z dnu např. na měsíce) volby:
    - Nezahrnovat dosud neskončený měsíc
    - Zahrnout – hodnota month-to-day (každý den se přepočítává)
  - OLAP – viz dále

ETL

- Krok 10 – automatizace
  - Načasovat jednotlivé kroky
    - Lze na sebe jednotlivé kroky navázat (např. zápis do tabulky metadat, že už jeden proces skončil a druhý může začít, existence souboru, …)
  - Získat potřebná metadata
    - Proces
    - Start
    - Konec
    - Doba běhu
    - Počet přesunutých řádků
    - Status dokončení (úspěšně/neúspěšně)
    - Diskové operace, CPU, …
    - Viz Auditní dimenze u faktové tabulky

ETL

- Krok 10 – pokračování
  - Možný postup:
    - Extrakce dimenzí a zápis metadat
    - Extrakce faktů a zápis metadat
    - Procesování dimenzí
      - Umělé klíče/SCD/….
      - Čištění dat, zápis metadata
    - Procesování fakt
      - Umělé klíče, zápis nevyhovujících záznamů
      - Zápis nevyhovujících záznamů
      - Transformace dat
    - Load dimenze
    - Load fakta
    - Agregace, OLAP
    - Validace loadu proti metadatům
    - Záměna serverů (pro 24 hod DW)
    - Načtení dat pro datové tržiště
    - Aktualizace metadat
    - Zápis metadat o loadu
    - Otestování správnosti a úplnosti loadu

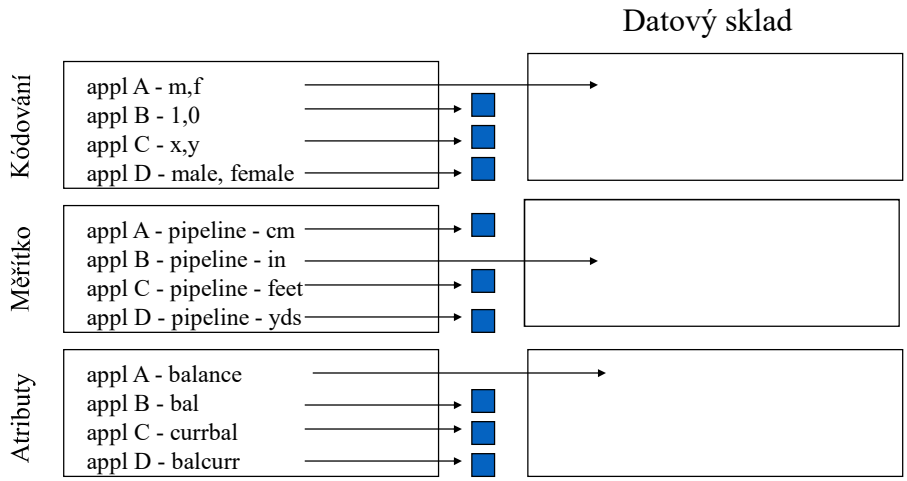
ETL

- Hlavním problémem je kvalita zdrojových dat
- Obsáhlý problém (Data quality and cleaning)
- Kvalitní data
  - Přesnost
  - Kompletnost
  - Konzistence
  - Jedinečnost (stejné názvy pro atributy se stejnou informací)
  - Včasnost
- Kvalitní data – „pravda, jenom pravda a nic než pravda“

ETL

- Problémy v datech
  - Nekonsistentní používání kódů (Ano, true, T, …)
  - Jeden atribut uchovává více informací
  - Význam atributu záleží na hodnotě druhého atributu
  - Chybějící hodnoty
  - Duplicity
  - Chybné hodnoty
  - Překlepy
- OLTP systémy pro podporu transakcí
  - Zcela jiná priorita – především důraz na transakce
  - Kvalita dat není na prvním místě
  - Validace dat by mohla neúměrně zdržet zápis transakce
  - Problémy při ručním vkládání
- DW – přínos – ukazuje na kvalitu dat
- Pro zvýšení kvality potřeba získat podporu managementu

ETL – příklad transformace



ETL

- Nejběžnějším problémem je integrace zákazníků
  - Standardizace jmen a adres
  - Householding – identifikace ekonomické jednotky
  - Na tento problém existují specializované nástroje
- Doporučení ke kvalitě dat
  - Je-li možnost (více potencionálních zdrojů dat) vybrat nejkvalitnější zdroj
  - Prozkoumat data – hledat možné problémy
  - Sdělit problémy s kvalitou (nekvalitou) dat managementu – velmi často existuje názor, že data jsou nekvalitní jen v DW
  - Nezbytná spolupráce s OLTP správci na odstranění chyb
  - Nezbytná spolupráce s uživateli na definici pravidel pro čištění dat
  - Doporučeno využívat specializované nástroje na čištění dat
  - Odpovědnost za čištění dat – je-li to možné - přenést na provozovatele OLTP (čisté extrakty pro DW)

ETL

- Kontrola správnosti dat
  - Dotazy vůči provozním systémů
    - Porovnání s DW
    - Možnost automatizovat a ukládat do metadat
  - Manuální prohlídka
    - Nejde-li testovat vůči provozním (např. informace v DW z více zdrojů)
    - Hledat odchylky, možné chyby
    - Spolupracovat s uživateli (experty)
  - Výsledek není zaručen

ETL

- 0. vrstva
  - Často místo pro zálohy
  - Obsahuje detailní data
  - Je možné z nich rekonstruovat další vrstvy (?inkrementální load fakt a dimenzí?)
    - Není třeba je zálohovat (data jejich, modely ano)
  - Může sloužit pro nové transformace (např. pro data mining)
- Potřeba kontrolovat dostupné místo na disku
  - Např. i místo přidělené pro růst databáze, zda nedojde k jeho vyčerpání během ETL – problém
    - Někdy vhodné vytvořit speciální opravné pumpy

Praktický příklad

- Tvorba modelu a naplnění 1. vrstvy
  - Využití DTS
  - SQL skripty
    - Create
    - Naplnění