Datové sklady

Datové sklady (Data Warehouses) představují hlavní datovou základnu pro aplikace BI. Cílem kapitoly je seznámení se s významem a základními principy jejich architektury, ale i s jedním z možných pohledů, dle kterého je na datové sklady v této práci nahlíženo.

V souvislosti s prudkým rozvojem informačních technologií představují podniková data pro většinu společností cenný zdroj informací. Tato data jsou často roztříštěna po různých podnikových systémech, jejichž funkcionalita se většinou liší natolik, že i data v nich obsažená mají odlišnou strukturu a existuje u nich vzájemná nekompatibilita. S narůstajícím objemem dat se však tento problém začal stále více prohlubovat, protože získání požadovaných dat napříč celým podnikem bylo čím dál více složitější.

Definice datového skladu

Termín *„datový sklad"* poprvé definoval v roce 1991 William Harvey Inmon, který je často nazýván „otcem datových skladů". Jeho definice zní následovně:

„Datový sklad je integrovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu."

Pojmy z Inmonovy definice se dají interpretovat takto

- **Subjektová orientace** - data jsou rozdělována podle jejich typu, ne podle aplikací, ve kterých vznikla.
- **Integrovanost** - data jsou ukládána v rámci celého podniku, a ne pouze v rámci jednotlivých oddělení.
- **Stálost** - datové sklady jsou koncipovány jako "Read Only", což znamená, že zde žádná data nevznikají ručním pořízením, a nelze je ani žádnými uživatelskými nástroji měnit.
- **Časová rozlišenost -** aby bylo možné provádět analýzy za určitá období, je nutné, aby byla do datového skladu uložena i historie dat. Načítaná data s sebou tedy musí nést i informaci o dimenzi času.

Obecně může existovat více pohledů, dle kterých lze na datové sklady nahlížet. Jejich hlavní rozdíl spočívá v rozsahu komponent a procesů, které jsou vnímány jako dílčí součást skladu. Na celkové koncepci se však prakticky nic nemění, protože počet a provázanost komponent samotných zůstává nezměněn. Jeden z možných pohledů je do značné míry vystižen přímo v obrázku č. 1 na straně 12. Tento pohled vnímá datový sklad jako hlavní datové úložiště ve vrstvě databázových komponent. V poněkud širším vymezení je tento pojem chápán v druhém pohledu, kde jsou zahrnuty procesy od ETL zpracování dat až po jejich postupné převedení do jednotlivých datových tržišť. Datový sklad je tak v tomto případě tvořen prvními dvěma vrstvami architektury BI.

Osobně jsem si již zvyknul při své praxi v bance datový sklad vnímat tímto druhým pohledem, jelikož právě takto je tam na něj nahlíženo. Konkrétně na mém pracovišti jsou jako součást datového skladu označeni také zaměstnanci, kteří se starají o jeho provoz, protože bez jejich každodenní práce a údržby celého skladu by zkrátka nemohl fungovat a přinášet firmě užitek. I nadále bych si proto dovolil nahlížet na tuto problematiku tímto způsobem.

Důležité pojmy v teorii datových skladů

Dříve, než se pustím do popisu samotné architektury datového skladu, bude třeba blíže představit některé komponenty a procesy, které jsou součástí dané problematiky. Tyto pojmy byly již zmíněny v kapitole 1.3 a setkáme se s nimi minimálně ještě v následující kapitole v rámci jednotlivých architektonických přístupů k zavádění datových skladů. Jedná se o následující výrazy:

ETL (Extaction Transformation Loading) – tzv. *„datová pumpa"*. Jedná se o sled komplexních algoritmů, které slouží k požadované úpravě a převedení dat ze zdrojových systémů do datového skladu. Pro tento účel jsou často využívány specializované ETL nástroje (např. MS SQL Server Integration Services, Informatica).

Dočasné úložiště (Data Staging Area) – jedná se speciální typ relační databáze, jejíž význam spočívá v rychlé extrakci (extrakční fáze ETL) dat ze zdrojových systémů, aby mohla být následně vyčištěna a transformována (transformační fáze ETL) bez zbytečného zatěžování těchto systémů, které jsou již tak dost vytíženy běžným provozem. Data jsou zde uložena pouze krátkodobě, obvykle v rozmezí jednoho dne až měsíce a jsou tak obrazem dat ve zdrojových systémech za tento časový interval.

Datové tržiště (Data Mart) – je podmnožinou datového skladu, která se typicky orientuje na jednu konkrétní část podnikového zaměření. Každé datové tržiště obsahuje data vztahující se k dané části a je tedy primárně určeno pro potřeby odpovídající skupině uživatelů, kteří jsou na okruhu těchto dat závislí.

Podle Marka Humpriese [6, s. 35] by se dala definice datového tržiště z předchozí strany shrnout i následovně:

„Data pro datová tržiště jsou vybírána s cílem vyhovět specifickým požadavkům částem organizace. Není neobvyklé najít datová tržiště vyvinutá a implementovaná pro oddělení, divizi či geografickou lokaci."

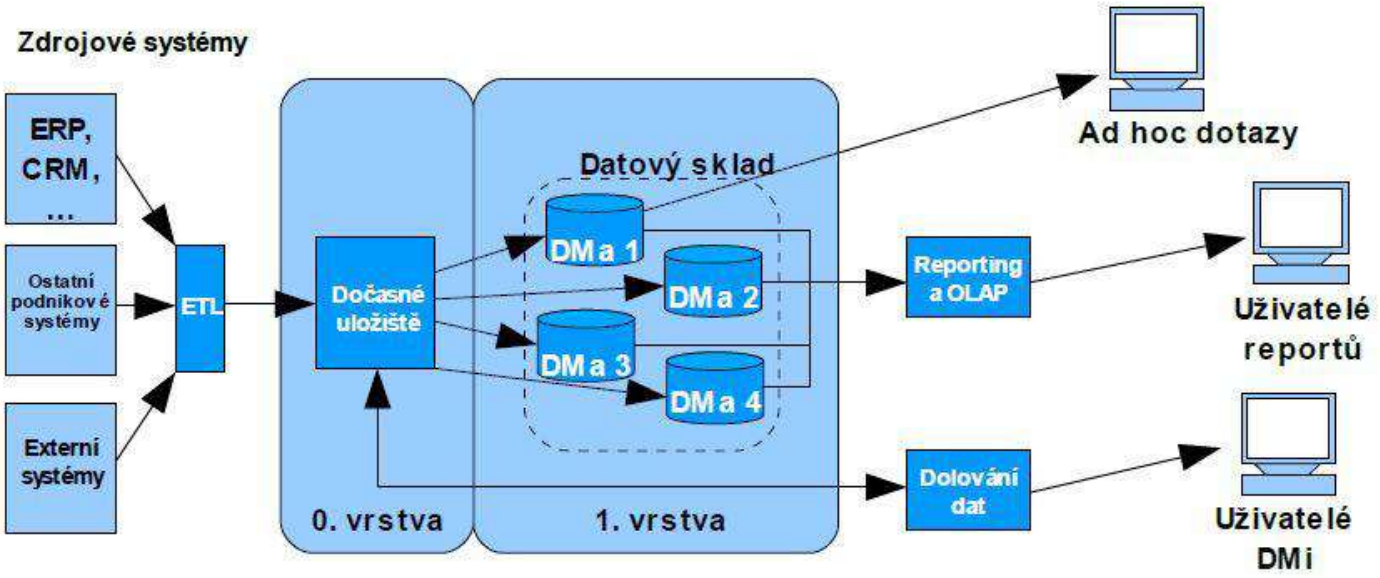
Architektura datového skladu

Existují rozdílné přístupy, koncepce k tvorbě datových skladů. Přístupy se liší zvolenou architekturou, složitostí a nákladností implementace, možnostmi a složitostí škálovatelnosti. Obvykle se můžeme setkat se třemi přístupy k řešení

- Postupné budování datových tržišť, založené na architektuře nezávislých datových tržišť – tzv. *„dvouvrstvá architektura"*
- Jednorázové vybudování celkového řešení, založené na architektuře konsolidovaného datového skladu – tzv. *„třívrstvá architektura"*
- Přírůstkový přístup založený na architektuře konsolidovaného datového skladu

Dvouvrstvá architektura

Tato architektura byla navržena Ralphem Kimballem a její princip je založen na konceptu vzájemně nezávislých datových tržišť. Datový sklad se buduje postupně po jednotlivých tržištích a nejen výsledky, ale i finanční prostředky na vývoj jsou rozloženy v čase. Tento přístup je volen především tehdy, pokud je třeba upřednostnit konkrétní oddělení či pobočku a dodat první výstupy z datového skladu v co nejkratším možném čase. [4]



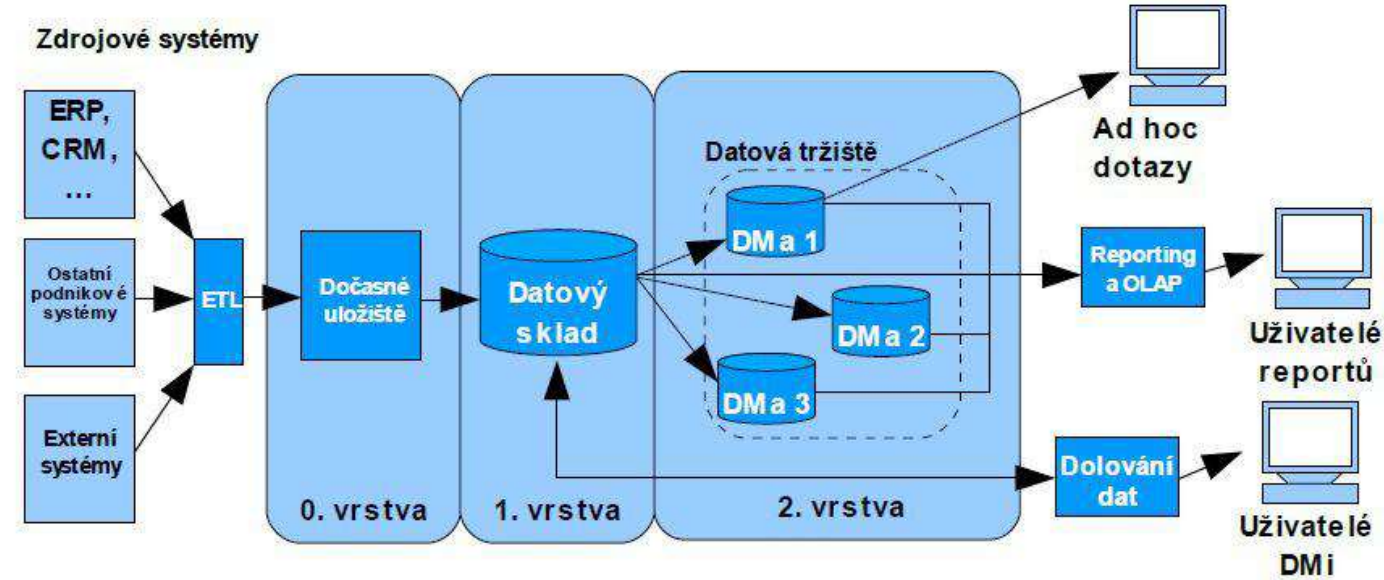
Obrázek č. 2: Postupné budování datových tržišť, převzato z [8]

Dvouvrstvá architektura však spolu nese i určitá rizika. Při postupné integraci nových datových tržišť může dojít k jejich částečnému překrývání, kdy se jeden atribut vyskytuje na několika místech a jakákoli změna musí být provedena ve všech instancích tohoto atributu, což sebou nese i vyšší náklady na údržbu datového skladu. Tento přístup je vhodné používat v následujících situacích

- Není technologicky možné nebo není potřebné budovat celopodnikové řešení založené na třívrstvé architektuře.
- Je potřeba vybudovat rychle řešení pro několik vzájemně nezávislých oddělení, přičemž se neočekává do budoucna potřeba celkové integrace řešení.
- Zadavatel nemá nebo není ochoten vynakládat finanční prostředky na počáteční integrační činnosti spojené s architekturou konsolidovaného datového skladu.

Třívrstvá architektura

Zakladatelem třívrstvé architektury je již výše zmíněný William Harvey Inmon. Její podstatou je jednorázové vybudování celého databázového řešení, které pokryje všechny požadované analytické potřeby firmy. Tato koncepce s sebou nese vyšší počáteční náklady a ve většině případů také značně dlouhou dobu na kompletní realizaci.



Obrázek č. 3: Architektura konsolidovaného datového skladu, převzato z [8]

Oproti předchozímu má tento přístup následující výhody

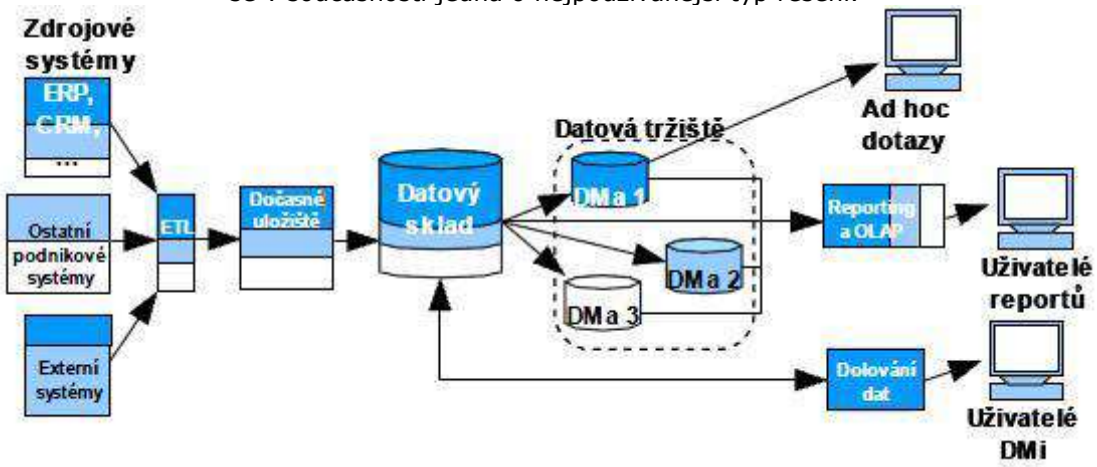
- Architektura je dostatečně flexibilní a integrovaná pro podporu náročných analytických úloh požadujících nejen agregovaná, ale i detailní data.

- Vzhledem ke komplexnosti uložených dat lze jednoduše budovat téměř neomezené množství datových tržišť pro potřeby různých uživatelů.
- Díky udržování normalizovaných transakčních dat na úrovni datového skladu je možno tato data použít i pro jiné typy analytických úloh, než pro jaké byly původně určeny.

- Datový sklad přímo podporuje tvorbu specializovaných datových úložišť i v jiné než dimenzionální formě (např. pro dolování dat).

Přírůstkový přístup

Tento typ je stejně jako předcházející přístup založen na principu třívrstvé architektury a jedná se o nejmladšího zástupce z již uvedených architektur, který se snaží skloubit výhody obou předchozích přístupů. Jeho asi vůbec největší výhodou je dle fakt, že jednotlivá řešení jsou dodávána postupně a proto i finanční náklady jsou více rozprostřeny v čase, což umožňuje pružnější sledování návratnosti těchto investic. Neméně důležitým kladem je i to, že je zde díky principu postupných přírůstků více prostoru pro případné změny v konceptu. Není tedy divu, že se v současnosti jedná o nejpoužívanější typ řešení.



Obrázek č. 4: Sklad založený na přírůstkovém principu, převzato z [8]

Tento přístup je vhodné používat například v následujících situacích

- Zadavatel chce vybudovat konsolidované řešení, díky čemuž bude mít možnost zajišťovat případné změny v koncepci datového skladu, neboť očekává signifikantní rozvoj v uživatelských požadavcích.
- Zadavatel je ochoten investovat počáteční časové i finanční prostředky k vytvoření celkové strategie a následně budovat malá řešení přinášející okamžitý užitek