ARCHITEKTURA DATOVÉHO SKLADU A PŘÍSTUP K DATŮM V REÁLNÉM ČASE

Dušan Kajzar

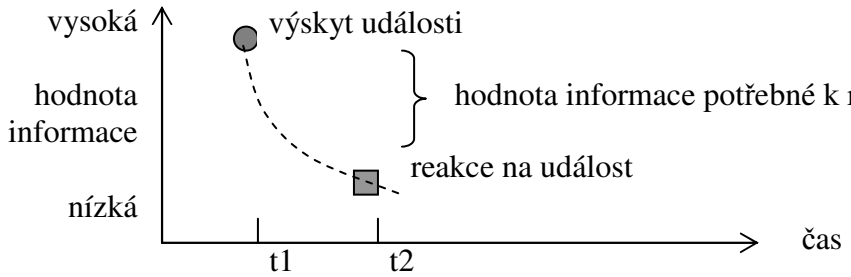
Slezská univerzita v Opavě, Filozoficko - přírodovědecká fakulta,
Ústav informatiky, Bezručovo nám. 13, 746 00 Opava, e-mail: kajzar@c-box.cz

Abstrakt
Autor článku se zabývá problematikou datového skladu, jehož úkolem je poskytovat podnikovým manažerům podklady nejen k rozhodování strategického charakteru, ale i k rozhodování na taktické či operativní úrovni. Nejprve jsou diskutovány možné architektury datového skladu splňující uvedené požadavky, dále jsou zmíněné podpůrné technologie pro přenosy dat a zajištění dostupnosti systému, v poslední části článku je diskutován pohled na potřebu informací „v reálném čase“ a „ve správném čase“.

1. ÚVOD

Datové sklady (Data Warehouse) jsou technologií, která v dnešní době již v mnoha podnicích zdomácněla a podnikový management si svoji práci bez analýz pomocí Business Intelligence nástrojů nedovede představit. Je tedy logické, že rostou nároky podnikového managementu na výstupy podporující nejen strategickou, ale i taktickou či dokonce operativní úroveň řízení. Jinými slovy - objevují se požadavky na výstupy (reporty) obsahující **aktuální data**, požadavky na přístup k datům **v reálném čase**.
Pro architekturu datového skladu umožňující analýzu dat v „reálném čase“ se v literatuře používají pojmy „**Real-Time Data Warehouse**“ resp. „**Active Data Warehouse**“.

Koncepce „Real-Time Data Warehouse“ resp. „Active Data Warehouse“ řeší problém exponenciálně klesající hodnoty informace v závislosti na stáří této informace [3]:



Jak ovšem skloubit klasickou architekturu datového skladu (založeného na datech ustálených, neměnných, dávkově periodicky aktualizovaných) s požadavky na přístup k aktuálním (ještě neustáleným, měnícím se) datům, tj. k datům, která jsou doménou provozních systémů?

V dalším textu chci nastínit řešení datových skladů, která do větší či menší míry mohou splnit požadavky na poskytování informací v reálném čase. Půjde o:

- využití klasické koncepce datových tržišť (Data Marts);
- koncepci operativních datových úložišť (Operational Data Store - ODS), začlenění ODS do architektury datového skladu;
- koncepci systému „Active Data Warehouse“.

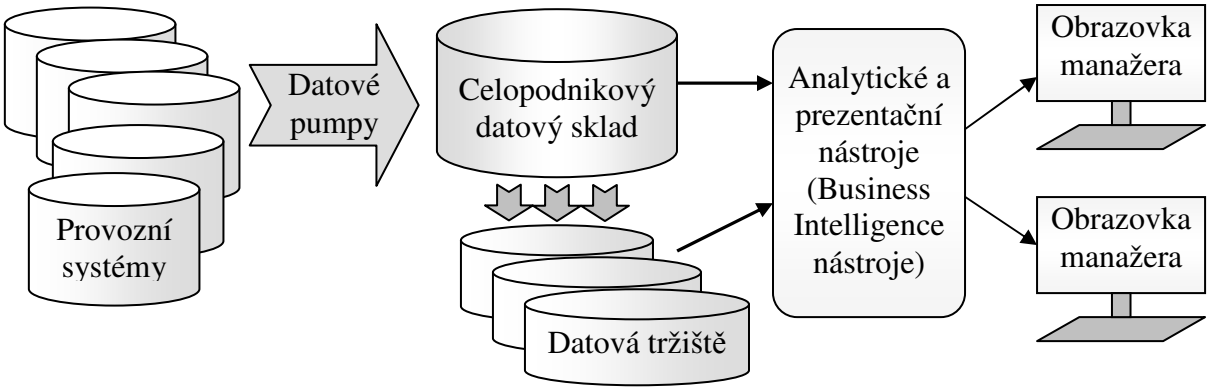
Součástí článku je také zamyšlení nad souvisejícími technologiemi, které uvedený trend provázejí a nad potřebou informací v reálném čase („in real time“) versus informací ve správném čase („in right time“).

2. REAL-TIME DATA WAREHOUSE A ZÁKLADNÍ ARCHITEKTURY

V této kapitole zmíním a stručně vysvětlím základní používané architektury datových skladů se zaměřením na jejich možnosti zahrnout do analýz data „v reálném čase“.

2.1 Klasická koncepce datového skladu a datových tržišť

Následující obrázek zachycuje klasickou architekturu Data Warehousingu, tj. procesu získávání, zpracování a skladování dat z různých provozních aplikací tak, aby tato data byla přístupná pro potřebné analýzy - pro analýzy sloužící jako podklady k manažerskému rozhodování.



Obr. 1 Klasická koncepce datového skladu a datových tržišť

Analytické a prezentační nástroje (Business Intelligence nástroje) mohou čerpat data pro zpracování potřebných analýz z celopodnikového datového skladu nebo z datových tržišť. Datové tržiště (Data Mart) je tématicky orientovaný datový sklad určený ke zprostředkování informací pro určitý útvar podniku, např. pro marketing, finanční řízení apod.

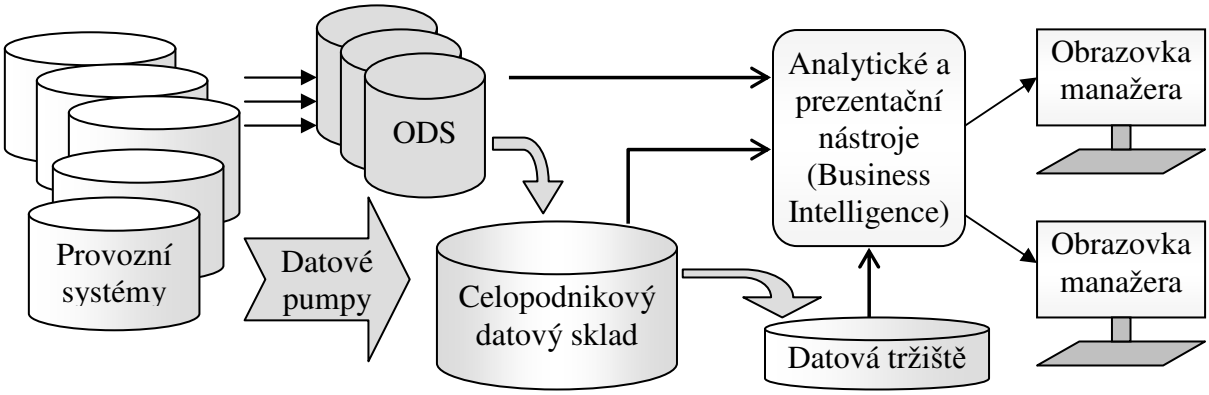
Problémem datového (informačního) toku z provozních systémů přes datový sklad, datová tržiště až na obrazovku manažera, je ovšem jeho časová náročnost. Datové pumpy (jinými slovy ETL nástroje - Extraction, Transformation, Loading) pracují dávkově a transformační proces jimi realizovaný trvá jistý čas. Po dobu transformace obvykle bývá blokováno i využití databáze datového skladu. Neméně časově náročným procesem jsou i výpočty agregací a generování vícedimenzionálních kostek OLAP (On-Line Analytical Processing) serverů, které realizují datová tržiště.

O využití dat v „reálném čase“ v této architektuře proto není možno hovořit. Zvýšit denní četnost běhu datových pump situaci neřeší, jelikož se zvyšuje režie systému pro návazné procesy - tj. pro přepočítávání agregací a opětovné generování vícedimenzionálních kostek. Systém je takto neúměrně zatěžován, blokován transformačními a dalšími systémovými procesy údržbového charakteru. Zvýšení četnosti běhu datových pump není vhodné ještě z dalšího důvodu - datový sklad by měl obsahovat data již provozně zpracovaná, hodnotově stálá, neměnná.

Praxe tedy hledala další řešení, která by umožnila Business Intelligence nástrojům přístup k aktuálnějším datům - jedním z řešení je aplikace tzv. operativních datových úložišť.

2.2 Operativní datová úložiště

Operativní datová úložiště (Operational Data Store - ODS) jsou komponentou architektury datových skladů, která slouží jako místo datové integrace aktuálních dat provozních systémů. ODS jsou databáze, které podporují vkládání a modifikaci dat v reálném čase, např. pomocí replikací z provozních systémů resp. pomocí integračního serveru (Enterprise Transformation Integration - EAI). ODS jsou takto zdrojem konsolidovaných agregovaných dat s minimální dobou odezvy, zdrojem vhodným pro využití dat „téměř v reálném čase“ [3].
Analytické a prezentační nástroje (Business Intelligence nástroje) v této architektuře mohou čerpat data pro zpracování potřebných analýz jak z celopodnikového datového skladu a datových tržišť, tak z ODS - v případě požadavku na „near real time“ přístup k datům.



Obr. 2 Architektura s využitím operativních datových úložišť (ODS)

Koncepce ODS jistě může v mnoha podnicích vyhovět požadavkům na zpracování a poskytování informací v reálném či spíše téměř reálném čase. Ovšem v některých případech koncepce ODS vyhovovat nemusí. Navíc - koncepce ODS bývá některými autory kritizována. Podle těchto kritiků nejsou ODS ničím jiným, než „maskovanými“ tzv. nezávislými datovým tržišti. Nezávislá datová tržiště jsou budována přímo nad provozními systémy, datový sklad pak vzniká integrací dat těchto datových tržišť. Taková koncepce je dnes považována mnohými experty za překonanou.
Přednost se dává vybudování celopodnikového datového skladu (Enterprise Data Warehouse - EDW), který je plněn daty pomocí datových pump přímo z provozních systémů. Teprve nad celopodnikovým datovým skladem jsou konstruována podle potřeby tzv. závislá datová tržiště - odvozením (výběrem a agregací) potřebných dat z jednotné datové základny, tj. z celopodnikového datového skladu.

Návrháři architektury datových skladů si tedy kladou otázku: Je možno se vyhnout budování ODS a v případě potřeby efektivně přistupovat přímo k datům provozních systémů? Odpověď přináší další koncepce, tzv. koncepce Active Data Warehouse.

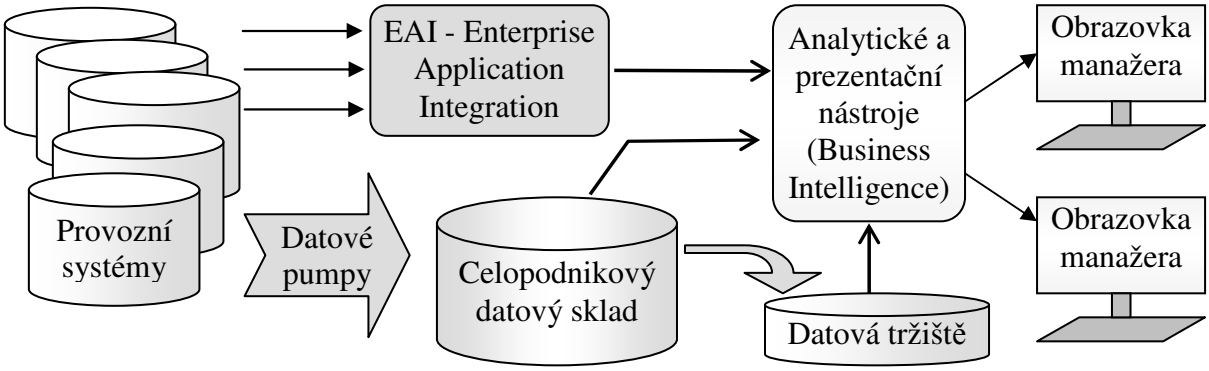
2.3 Koncepce Active Data Warehouse

Koncepce Active Data Warehouse [6] je založena na propojení celopodnikového datového skladu a provozních systémů v aplikační vrstvě tak, aby jednotlivé klientské aplikace (analytické a prezentační nástroje - Business Intelligence nástroje) byly schopny v případě potřeby využít přímo data provozních systémů. Toto propojení přitom musí být pro koncového uživatele transparentní. Koncovému uživateli je zcela jedno, odkud jeho klientská aplikace data čerpá, zajímají ho správné výsledky.

Technologie Active Data Warehouse se vyznačuje možností využívat aktuální, detailní data a poskytovat podklady pro:

- strategická rozhodnutí (rozhodnutí dlouhodobého charakteru zahrnující obvykle celopodnikovou tématiku);
- taktická rozhodnutí (krátkodobější charakter dopadu rozhodnutí, obvykle zaměřeno na některé téma či podnikovou oblast);
- operativní rozhodnutí (reakce na určité události).

Podklady pro strategická rozhodnutí jsou získávány na základě analýz zhotovených z historických, neměnných dat. Hlavním zdrojem dat pro analýzy je celopodnikový datový sklad případně datová tržiště.
Získání podkladů pro rozhodování taktického charakteru již vyžaduje přístup jak k datům historickým a neměnným, tak i k datům aktuálním, i když třeba ještě se částečně měnícím.
Získání podkladů pro rozhodování operativního charakteru se již neobejde bez přístupu k datům „in real time“.



Obr. 3 Architektura s využitím integračního serveru (EAI)

Koncepcí Active Data Warehouse se datový sklad posunuje svým významem směrem k provozním systémům. Otázka tedy zní: Jak vhodně integrovat datový sklad s ostatními službami poskytovanými provozními systémy?
Business Intelligence nástroje v této architektuře musí být schopny spolupráce se standardními službami integračního serveru (EAI - Enterprise Application Integration), kterými jsou např. Web Services, Java, .Net, resp. spolupráce pomocí tradiční klient/server technologie.
Je však potřeba si uvědomit, že dotazy zasílané analytickými a prezentačními nástroji do databází provozního charakteru mohou výkonnostně konkurovat standardnímu provoznímu zpracování. Může takto docházet ke kolizím mezi procesy typu DSS (Decission Support) a procesy typu OLTP (On-Line Transaction Processing).

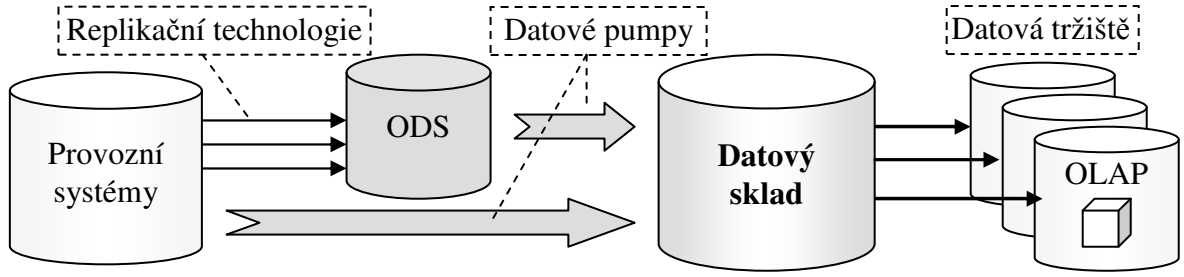
3. SOUVISEJÍCÍ TECHNOLOGIE

V předchozí kapitole jsem charakterizoval architektury datových skladů z hlediska jejich možností zahrnout do analýz data „in real time“ resp. „in near real time“. O architektuře typu Real-Time Data Warehouse však není možno uvažovat izolovaně - nedílnou součástí návrhu architektury musí být související technologie, bez kterých by celá koncepce Real-Time Data Warehouse v praxi selhala. Souvisejícími technologiemi zde mám na mysli:

- infrastrukturu datových přenosů;
- technologii pro zajištění dostupnosti systému.

3.1 Infrastruktura datových přenosů

Aby navržená architektura datového skladu skutečně plnila svou funkci, je nutné zajistit kvalitní oboustranné propojení všech navržených částí, tj. provozních systémů, integračního serveru (EAI) resp. operativních datových úložišť (ODS), celopodnikového datového skladu, datových tržišť (OLAP serverů), pomocí vhodné infrastruktury datových přenosů. Nemám zde na mysli pouze dostatečně propustnou a spolehlivou počítačovou síť LAN či WAN, ale i vlastní technologii, která datové toky realizuje.
Datové pumpy zajišťují datový tok (spojený s potřebnou transformací dat) z prostředí provozních systémů resp. z prostředí ODS do celopodnikového datového skladu. Datové pumpy mohou mít podobu relativně jednoduchých jednoúčelových programů (tzv. programových skriptů), ale také výkonných grafických nástrojů, mezi které patří např. DTS nástroje (Data Transformation Services) firmy Microsoft nebo systém Informatica firmy Informatica Corporation.
Datový tok z celopodnikového datového skladu do databáze datového tržiště je pak realizován pomocí služeb OLAP (On-Line Analytical Processing) serverů.



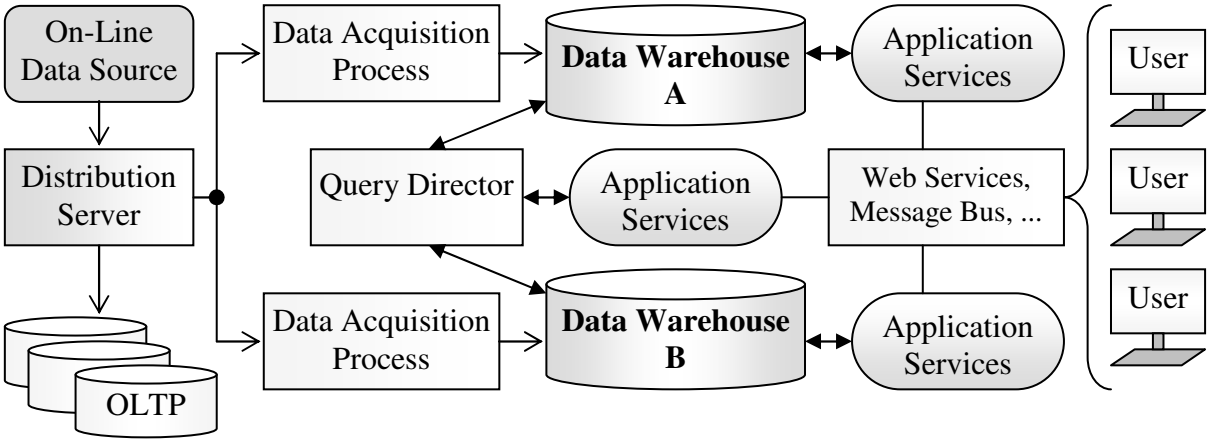
Obr. 4 Technologie datových toků - replikační systém a datové pumpy

Pomocí jaké technologie je vhodné realizovat datový tok z provozních systémů do ODS? Tento datový tok by totiž měl operativnímu datovému úložišti dodávat data průběžně, nikoliv dávkově, jak je tomu u datových pump. Vhodnou technologií je proto replikační mechanismus, jenž je v současné době součástí všech významných databázových systémů. Replikační mechanismus (replikační systém) umožňuje definovat přenosy dat ze zdrojové do cílové databáze na základě tzv. replikačních předpisů (je možno definovat, která data se mají přenášet - které tabulky, sloupce tabulek, data vyhovující dané podmínce atd.). Replikační systémy mnohdy umožňují přenášet data i mezi různými databázovými platformami (např. Sybase - Oracle apod.).
V souvislosti s datovým skladem tedy můžeme replikační systém prezentovat jako technologii, která stojí v pozadí budování datového skladu jakožto páteřní systém pro přenosy dat. Replikační systémy tak hrají významnou roli v procesu včasných dodávek dat do centrálního datového skladu.

Výkonnost celé přenosové infrastruktury a všech použitých přenosových technologií umožňuje propojení provozního a analytického prostředí v přenosové infrastruktuře tak, jak je tomu v architektuře Active Data Warehouse.

3.2 Zajištění požadované dostupnosti systému

Využívání datového skladu pro oblast taktického či dokonce operativního rozhodování posouvá tyto systémy blíže ke kategorii systémů provozních. Z datových skladů jakožto systémů podpůrného charakteru se takto mohou stát systémy se strategickým významem pro podnik, tj. systémy, při jejichž třeba i krátkodobém výpadku jsou ohroženy či znemožněny základní aktivity podniku.
V souvislosti s koncepcí Real-Time Data Warehouse je proto potřeba navrhnout i vhodnou architekturu pro zajištění požadované dostupnosti systému, pro eliminaci případných výpadků funkčnosti.
Dodavatelé technologií datových skladů se samozřejmě touto problematikou zabývají. Na datové sklady je dnes možno aplikovat clusterové technologie, které umožňují pokrýt jak výpadek některé komponenty datového skladu, tak i rozložit zátěž mezi paralelně pracujícími uzly. Jako příklad uvádím schéma technologie zajišťující rozložení zátěže a eliminaci výpadku jednoho z uzlů tak, jak jej prezentuje firma NCR pro datový sklad Teradata (převzato z [6]):



Obr. 5 Zajištění dostupnosti datového skladu (zpracováno podle [6])

V levé části obrázku č. 5 je znázorněno souběžné plnění provozních systémů (OLTP) a datových skladů A a B daty z On-Line datového zdroje. Směrem zprava je pak znázorněn tok dotazů do datových skladů ze strany uživatelů prostřednictvím komponent Application Services a Query Director.

4. INFORMACE V REÁLNÉM ČASE VS. VE SPRÁVNÉM ČASE

Zamysleme se nyní spolu s [7] nad potřebou přístupu k datům v reálném čase („in real time“) a potřebou dat (informací) ve správném čase („in right time“). Z jistého úhlu pohledu můžeme rozlišit dva typy Real-Time Data Warehouse:

A) Data Warehouse, který **poskytuje** přístup k datům (informacím) „in real time“;
B) Data Warehouse, který **získává** data (podklady pro zpracování informací) „in real time“.

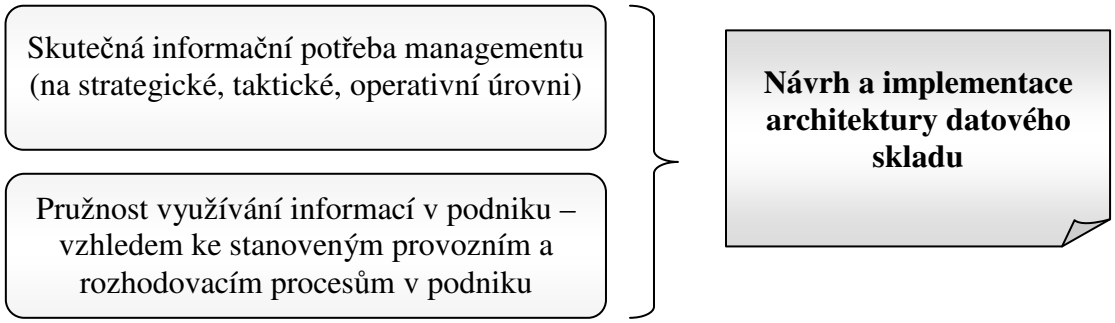
Data Warehouse typu „B“ principiálně zahrnuje Data Warehouse typu „A“. Získávat data „in real time“ implikuje také možnost poskytování informací „in real time“.
Opačně:
Poskytování informací „in real time“, i když podklady pro ně nebyly získány „in real time“, je v praxi také velmi důležité. Jde totiž o to, aby manažer měl potřebné informace v době, kdy je skutečně potřebuje, kdy je to nutné z hlediska jeho odpovědnosti a rozhodovací pravomoci.

To samozřejmě nemusí znamenat „ihned“. Předčasně a za každou cenu „chrlené“ informace mohou vést k informačnímu zahlcení manažera, navíc mohou být nepřesné, jelikož systém při jejich zpracování vychází z ještě neusazených, neúplných a měnících se dat. Spíše tedy bychom měli hovořit o poskytování informací „in right time“, než „in real time“.
Chybou analytiků při rozboru informačních potřeb manažerů může být neadekvátní zaměření se na potřebu informací „in real time“ místo „in right time“.

Požadavky na úroveň služeb poskytovaných systémem Real-Time Data Warehouse by měly vždy vycházet z potřeb podnikových procesů, nikoliv z možností informačních technologií. Navržená architektura Real-Time Data Warehouse by sice měla do budoucna počítat s možností „real time“ charakteristik (např. zahrnovat možnost budoucího rozšíření funkčnosti, škálovatelnost apod.), implementována by však měla být podle skutečných požadavků business procesů a rozhodovacích procesů.

Technická dokonalost architektury Real-Time Data Warehouse bude k ničemu, pokud poskytované informace podnikový management nedokáže využít. Informace na „přístrojové desce“ systému poskytované „in real time“ jsou zbytečné, pokud v podniku fungují zkostnatělé a nepružné produkční i řídicí procesy a manažeři nejsou schopni na základě poskytovaných informací „in real time“ reagovat.

Dále - nikoliv každý manažer potřebuje analýzy na základě dat „in real time“. Víme, že pro podporu strategické úrovně řízení jsou podstatná historická neměnná data. Propojení historických dat s daty provozních systémů „in real time“ resp. „in near real time“ pak bývá potřebné spíše u systémů podporujících taktickou či operativní úroveň řízení.



Obr. 6 Faktory pro návrh a implementaci architektury datového skladu

Víme již, že architektura Real-Time Data Warehouse (Active Data Warehouse) umožňuje integrovat aktuální („in real time“) data provozních systémů do kontextu historických a již neměnných dat datového skladu. Analýzy provedené na základě takto integrovaných dat mohou napomoci managementu k lepšímu pochopení aktuální situace v podniku a trendů zejména v oblasti taktického rozhodování a řízení. Taková integrace může podpořit optimalizaci rozhodovacího procesu příslušného manažera.
Je samozřejmě zbytečné, aby manažerské informační systémy určené pro podporu taktického řízení počítaly potřebné agregace a historické údaje, jestliže tyto jsou k dispozici v prostředí datového skladu – zde je na místě propojení datových skladů s provozními systémy.

Je však nutné si uvědomit, že datový sklad nenahrazuje ani nemůže nahradit provozní (tj. OLTP) systémy. Datový sklad je zaměřen na podporu řízení podniku, tj. na analýzy integrovaných historických dat. Architektura datového skladu (v podobě Real-Time Data Warehouse resp. Active Data Warehouse) může být integrována s prostředím provozních

systémů, vyžaduje-li to řešení podpory rozhodovacích procesů. Účelem Real-Time Data Warehouse ovšem v žádném případě není nahrazení provozních systémů realizujících rutinní provozní podnikové technologie.

Z výše uvedeného plyne – na propojení dat datového skladu s daty provozních systémů je vhodnější se podívat spíše z úhlu potřeby informací „in right time“ než „in real time“. To znamená pochopit potřebu informací ve správný čas vzhledem k požadavkům na podporu rozhodovacích procesů (na úrovni strategické, taktické, operativní) a vzhledem k možnostem využití (tj. rychlosti, pružnosti, …) získaných informací v daném podnikovém prostředí.
Potřebu informací „in right time“ pak specifikujeme při návrhu a implementaci architektury datového skladu v daném podniku.

5. ZÁVĚR

V článku jsem se zabýval určitým trendem, který se v současné praxi budování datových skladů rozvíjí – tj. možností přístupu k aktuálním datům, k datům v reálném čase. Tento trend souvisí s využíváním technologií datového skladu a Business Intelligence nástrojů nejen pro podporu rozhodování strategického, ale i taktického či dokonce operativního. Trend je podporován architekturou tzv. Real-Time Data Warehouse resp. Active Data Warehouse.
V podstatě jde o propojení datového skladu a provozních systémů tak, aby analytické a prezentační nástroje byly v případě potřeby schopné zahrnout do analýz nejen historická a neměnící se data, ale i data aktuální, „ještě v pohybu“, tj. data provozních systémů.

V kapitole 2 jsem uvedl příklady tří architektur, které jsou do větší či menší míry schopné požadavky na využití dat v reálném čase uspokojit.
Kapitola 3 upozorňuje na související technologie, bez nichž se návrh architektury Real-Time Data Warehouse jistě neobejde.
Kapitola 4 je pak úvahou nad opodstatněností požadavků na poskytování informací „in real time“ resp. „in near real time“ a posouvá úhel pohledu k informacím „in right time“.

LITERATURA

[1] Humphries M.: Data warehousing - návrh a implementace. Computer Press, 2001.
[2] Lacko L.: Datové sklady, analýza OLAP a dolování dat. Computer Press, 2003.
[3] Novotný O., Pour J., Slánský D.: Business Intelligence. Grada Publishing, 2004.
[4] http://www.adastra.cz
[5] http://www.teradata.com
[6] Walter T.: Active Data Warehouse in the Real-Time Enterprise. Referát na konferenci Teradata, PARTNERS User Group Conference, Seattle, 2003.
[7] Brobst S.A.: Top Ten Mistakes to Avoid when Constructing a Real-Time Data Warehouse. Referát na konferenci Teradata PARTNERS User Group Conference, Seattle, 2003.