ÚLOŽIŠTĚ DAT A BEZPEČNOST DATABÁZÍ

Miloslav Špunda

Anotace

Příspěvek se zabývá problematikou užití centrálního úložiště dat jako dalšího stupně rozvoje informačního systému (IS) instituce, zejména vzhledem k specifickému prostředí vysoké školy (fakulty). Objasňuje pojem datového skladu jako možného technického řešení. Je diskutována problematika ukládání statických dat z provozu IS a zajištění jejich bezpečnosti při přechodu na užití centrálního úložiště.

Klíčová slova

centrální úložiště, úložiště dat, datový sklad, relační databáze, ochrana dat, řízení přístupu

Úvod

Rozvoj IT technologií a jejich stále se rozšiřující užití při řízení chodu instituce znamená i postupný nárůst ukládaných dat. Tato data mají původně převahu dat dynamických, potřebných pro rychlé ukládání a aktualizaci informací (ekonomické řízení, finanční systém, spisová služba, datové schránky). Postupně však narůstá i objem dat statických. Je to způsobeno implementací dalších aplikací, které přímo s každodenním chodem instituce nesouvisí, ale jsou z právních a dalších hledisek nezbytné (majetek, registr smluv, archivace, IT podpora výběrových řízení).

Tento nárůst aplikací podporujících chod instituce a s tím souvisící nárůst objemů příslušných ukládaných dat postupně zatěžuje provozní databáze IS, přitom některá uložená data (datové soubory) nejsou opakovaně aktualizována. Jejich uložení v rámci **provozního systému** (základní aplikace pro rychlé ukládání a aktualizaci dat) je tak nevýhodné, může způsobovat i zpomalení systému a komplikace s potřebným prostorem pro základní databáze IS.

Řešení této situace znamená rozvíjet IS instituce dalším směrem a to zřízením centrálního úložiště dat instituce. Znamená to zároveň nutnost vybudovat i aparát přístupu k uloženým datům (datovým souborům), ale z jiných hledisek než je jejich rychlé ukládání a aktualizace. Vedoucím pracovníkům by takový systém měl nabídnout globální pohledy potřebné k řízení s možností analýzy a dotazů. Centrální úložiště dat lze řešit implementací **datového skladu** (Data Warehouse – DWH), který kromě uložení dat statického charakteru mimo prostor provozních databází umožňuje i strukturovaný přístup k uloženým souborům a analytické dotazování.

Specifickým problémem je zde ochrana uložených dat, převzatých do datového skladu z provozního systému s přísným řízením přístupu a ověřováním oprávnění. Pro datový sklad je třeba implementovat takové řešení zpřístupňující data uživatelům, které respektuje původní charakter ochrany dat v provozním systému.

Datový sklad v informačním systému

Datový sklad je obvykle budován jako centrální úložiště pro všechna data (datové soubory) instituce. Smyslem je, kromě technického hlediska odlehčení datového prostoru provozních databází, poskytovat uživatelům ucelená data o provozu instituce. Datový sklad zároveň vytváří datovou základnu pro možné detailní analýzy dat.

Součástí SW podpory datového skladu jsou proto nástroje pro import dat z provozních databází, transformaci těchto dat do struktur datového skladu, přípravu dat pro reporting a analýzu a příslušné nástroje pro řízení přístupu uživatelů k uloženým datům. Data tak nejsou uchovávána roztříštěně v databázích provozních systémů či jejich částech daných často nesourodými částmi základní aplikace (např. řešení fakturace, registr smluv).

Datový sklad data integruje na jednom místě, navíc v konsolidované formě (sloučení informací z více zdrojů, odstranění duplicit, sjednocení číselníkových hodnot, atd.). Datový sklad uchovává kromě dat aktuálních i data historická (je možno sledovat vývoj), což provozní systém obvykle nepodporuje. Příkladem může být návaznost na spisovou službu (komponenta provozního systému), kdy datový sklad podporuje archivaci dokumentů a usnadňuje splnění právních náležitostí.

Základní rozdíl mezi datovým skladem a provozním systémem v IS instituce je v tom, že provozní systém podporuje rychlé ukládání a aktualizaci online přicházejících dat (např. faktury, objednávky, platby, průběh řešení grantových projektů, aktuální pohyb dokumentů v instituci, aj.). Podpora reportingu je malá a systém zatěžuje. Datový sklad naproti tomu je navržen a optimalizován zejména vzhledem k možnostem rychlého vyhledávání a odpovědí na dotazy uživatelů.

Architektura datového skladu obvykle vychází z koncepce, která datový sklad rozděluje do tří datových vrstev:

- dočasné úložiště (staging area), které je základní vrstvou datového skladu, slouží pro dočasné uložení a zpracování dat přenesených z provozního systému
- centrální úložiště dat sloužící k uchování konsolidovaných dat (hlavní vrstva datového skladu)
- datová tržiště (data marts) je vrstva obsahující data transformovaná z centrálního úložiště do forem vhodných pro analýzy, reporty, data mining, aplikace Business Intelligence (BI)

Data jsou mezi jednotlivými vrstvami přenášena jednosměrně ze základní vrstvy přes centrální úložiště do datamartů.

Součástí datového skladu jsou **metadata**, která slouží k popisu datového skladu (klasifikaci uložených dat) a k jeho řízení.

Vstupní data z různých zdrojů se do datového skladu importují, obvykle v pravidelných intervalech (např. denní frekvence importu). Import je řízen procesy **ETL** (Extract – Transform – Load), které extrahují data z provozního

systému a uloží je do dočasného úložiště. Data jsou následně transformována do vhodné struktury a uložena do centrálního úložiště. Data v centrálním úložišti nejsou měněna, nedochází k přepisu daty novějšími, obvykle ani k jejich výmazu (jen výjimečně jako technická akce). Data se udržují historicky, při změně se původní stav dat označí (jako aktuálně neplatný) a jako aktuální se uloží nová hodnota. Toto řešení dovoluje analyzovat vývoj sledovaných entit, porovnávat minulý a současný stav nebo odhadnout další vývoj.

Využití dat z centrálního úložiště detailně je vzhledem ke způsobu uložení obvykle složité a časově náročné na zpracování. Z tohoto důvodu jsou data obvykle hned po importu z provozních systémů transformována do tvaru vhodného pro reporting a analytické zpracování a uložena do datových tržišť (datamarts). Datamarty jsou navrhovány se zaměřením na další využití uložených dat, např. podle oblastí, kterých se týkají zamýšlené reporty (finanční data, výběrová řízení, smlouvy, apod.)

Výše zmíněná metadata jsou určena k uložení informací o jednotlivých entitách datového skladu. Popisují způsob naplnění, pravidla užitá při transformacích a jsou základní součástí architektury datového skladu. Obvykle jsou dvojího typu: klasifikační (typ entity, atributy, popis reportů, způsob naplnění z provozního systému, aj.) a technická (popisy procesů ETL, informace pro řízení datových toků, informace pro konsolidaci dat, aj.).

Základními funkcemi datového skladu jsou **reporting a analýza dat**, tedy zpřístupnění uložených dat koncovým uživatelům v dobře srozumitelné formě. SW datového skladu proto podporuje tvorbu předem definovaných statických reportů, které jsou poté zpřístupněny oprávněným uživatelům. Pro analýzu dat se obvykle užívá složitějších nástrojů, jakou jsou OLAP (On-line Analytical Processing) a data mining.

Technologie OLAP je určena pro rychlou analýzu rozsáhlých dat. Data pro OLAP jsou uložena v multidimenzionálních datových strukturách s různým stupněm agregace, zdrojem pro OLAP jsou datamarty. Na rozdíl od provozní databáze je zde uživateli poskytnuta možnost kombinace kritérií podle aktuální potřeby s rychlou odezvou na dotaz.

Data mining představuje pokročilejší způsob analýzy dat. Dovoluje analýzu dat zdánlivě nesourodých, odhadovat možný vývoj, hledání souvisících segmentů dat s možností promítnutí výsledků zpětně do dat v datovém skladu. Jsou zde uplatněny složitější matematické a statistické metody, data pro data mining jsou obvykle uložena ve specializovaných datamartech.

Implementace datového skladu obecně znamená značné zkvalitnění informačního systému instituce jako celku. Dovoluje využívat data z provozních databází pro plánování, hodnocení instituce ze strategických hledisek a podporuje pohled na instituci nezbytný z hlediska jejího řízení.

Implementace datového skladu a ochrana dat

Zajištění dostatečné ochrany dat vznikajících při práci s aplikacemi nad provozními databázemi představuje při implementaci datového skladu

jeden ze základních problémů. Oprávnění k přístupu při práci s aplikacemi je obvykle řízeno při administraci, jednotlivě pro každého uživatele vzhledem k jeho pracovní činnosti a odpovědnostem. Údržba systému oprávnění a jeho aktualizace je pracovně náročná, ale ve výsledku je ochrana dat důsledná a adresná. Představa o ochraně dat navržená při analýze, která předchází návrhu a vytvoření aplikace, tak může být při provozu plně respektována.

Přechodem k využívání datového skladu vzniká problém hodnocení oprávnění přístupu k datům, která původně vznikla za zcela konkrétních představ o jejich ochraně. Přenesením dat do datového skladu z provozního systému vzniká potřeba nové definice jejich ochrany.

Řízením přístupu k datům v datovém skladu se z hlediska SW podpory zabývají systémy access manager, které obsahují také nástroje pro řízení přístupu koncových uživatelů.

Pro praxi prostředí fakulty se jeví jako možné řešení koncept automatického přidělování rolí na základě personálních dat. Jakmile pracovník je pověřen konkrétní funkcí (např. proděkan) je tato skutečnost zaznamenána v personálním systému v provozním systému fakulty. Na tomto základě access manager automaticky vygeneruje odpovídající rozsah oprávnění k přístupu z hlediska souborů uložených v datovém skladu. Podobně při ukončení funkce pracovníka je toto oprávnění automaticky zrušeno. Je zřejmé, že funkčnost takového řešení je přímo závislá na aktuálnosti dat v provozním systému.

Závěr

Při přechodu na užívání centrálního úložiště dat (datový sklad) je nutno uvážit okolnost postavení instituce v širším kontextu. Fakulta jako instituce funguje autonomně, avšak z právního hlediska obvykle nemá subjektivitu, patří k vyššímu celku, například škole. Škola má ovšem jako celek vlastní koncept IS včetně ochrany dat a přístupu k nim. Některá citlivá data fakulty uložená v jejím datovém skladu mohou být pro řízení školy důležitá, přitom však jejich ochrana musí být i při jejich zpřístupnění pro IS nadřízené instituce zachována. Tyto otázky mají spíše koncepční a politický charakter, technicky je řešení možné. Toto řešení je nutno koncipovat předem s jasnými kompetencemi přístupu k datům držitele datového skladu tedy fakulty.

Dalším důležitým prvkem je otázka užité SW podpory. Prostředí vysoké školy je z hlediska koncipování IS charakterizováno množstvím dat a jejich vazeb, které se v jiných oblastech nevyskytují. Řešení z oblasti Business Intelligence (BI) obvykle bez rozsáhlejších modifikací nelze užít. Je třeba z ekonomického hlediska zvážit možnost vytvoření specifické SW podpory s návazností na již užívané aplikace v provozním systému.

Literatura

[1.] Lacko L.: Datové sklady, analýza OLAP a dolování dat. Computer Press, 2003.

[2.] Novotný O., Pour J., Slánský D.: Business Intelligence. Grada Publishing, 2004.

Kontakt:

Doc. Ing. Miloslav Špunda, CSc.
Ústav biofyziky a informatiky UK 1. LF
Kateřinská 32, Praha 2
e-mail: *miloslav.spunda@lf1.cuni.cz*