

Ukladáme firemné dáta efektívne?

V dnešnom svete sú dáta doslova na každom kroku. Máme k dispozícii on-line objednávky tovarov všetkého druhu, cestovné poriadky, bankové operácie v reálnom čase, on-line predaj lístkov atď. Všetky tieto výdobytky modernej doby môžeme využívať vďaka mnohým systémom, ktoré neustále pracujú a spracúvajú ohromné množstvá údajov. Za každým takýmto systémom sa skrýva nejaké úložisko dát – v druhej väčšine relačný databázový systém (RDBMS). Zvykli sme si, že tradičné relačné databázy vždy spoľahlivo vykonávajú svoju prácu. Sú však tým najlepším riešením v každej situácii?

Tradičný prístup

RDBMS boli vyvíjané presne podľa požiadaviek takéhoto typu systémov – veľa používateľov súčasne aktívne pracujúcich nad tou istou množinou údajov. RDBMS majú preto zapracované rôzne technológie na zabezpečenie vylúčenia vzájomných kolízií používateľov, ktorí súčasne manipulujú s tými istými údajmi. Dosahujú to rozdelením práce používateľov na isté atómicke časti – transakcie. Preto sú takéto systémy nazývané aj OnLine Transaction Processing (OLTP). A treba povedať, že RDBMS perfektne plnia úlohy a nároky kladené systémami OLTP.

Dáta vo svete BI

Čo však v prípadoch, keď sa chceme na naše dáta pozrieť s istým odstupom, takpovediac hlbším spôsobom analyzovať doterajšie údaje, nájsť v nich štatistické ukazovatele a určiť trendy vývoja do budúcnosti? Túto oblasť prístupu k dátam vo všeobecnosti nazývame Business Intelligence (BI). V ostatnom čase sa požiadavky tohto druhu vyskytujú čoraz viac. Preto sú vyvíjané tzv. systémy na podporu rozhodovania (Decision Support Systems – DSS). Tie na svoju činnosť vyžadujú relevantné, vyčistené a skonsolidované údaje z primárnych systémov OLTP. Navyše na efektívnu prácu s týmito údajmi je nevyhnutné mať dáta uložené vo vhodnej forme. Práve na tento účel budujú podniky dátové sklady (DWH), ktoré obsahujú integrované údaje z primárnych systémov. Keďže primárnych systémov môže byť relatívne veľa (nezriedka i desiatky v rámci jedného podniku) a v dátových skladoch sa uchovávajú nielen aktuálne, ale aj historické údaje, ktoré tu zostávajú pre ďalšie analýzy, DWH sú práve tie miesta v podniku, ktoré koncentrujú najväčšie objemy dát.

Už zo samotného faktu, že používatelia a aplikácie pristupujú a spracúvajú také veľké množstvá údajov, vyplýva, že reakcie na požiadavky môžu byť – a často aj sú – veľmi pomalé. Ak k tomu pridáme ďalšiu skutočnosť, že nad

týmito dátami zvyčajne prebiehajú zložité analýzy a generovanie komplikovaných reportov, nespokojní business používatelia sú na svete.

Už pri uvažovaní nad stratégiou budovania podnikového dátového skladu má teda veľký význam zaoberať sa výberom správnej technológie na ukladanie veľkého objemu dát.

Kde je hlavný problém?

Nemálo spoločností pri budovaní dátových skladov siahne po osvedčených systémoch na ukladanie dát – RDBMS. A majú na to svoje (objektívne) dôvody od získaných skúseností s danou platformou RDBMS cez vyškolených administrátorov až po motiváciu zachovať v podniku čo najmenší počet rôznych systémov.

Treba však mať na pamäti, že spôsob, akým sa pristupuje k dátam v systémoch DSS, je úplne odlišný od spôsobu manipulácie s dátami v primárnych systémoch OLTP. Kde sú rozdiely?

V prvom rade transakcie môžeme považovať za malé – pracujúce s malými kúskami dát, typicky pristupujúce k jednému záznamu. Naproti tomu query v systémoch DSS bývajú často veľké – pristupujúce k veľkým blokom dát, typicky prechádzajúce celú tabuľku alebo tabuľky obsahujúce aj niekoľko miliónov záznamov. Dôsledkom tejto skutočnosti je fakt, že transakcie v OLTP prebiehajú veľmi rýchlo, ale query v DSS sú oveľa náročnejšie na spracovanie.

Po druhé transakcie v primárnych systémoch obyčajne menia údaje (pridávajú, menia a odstraňujú záznamy), ale úlohy nad dátami v systémoch DSS sa primárne zaoberajú čítaním uložených údajov.

A do tretice v primárnych systémoch OLTP existuje zvyčajne veľa ľudí alebo procesov spúšťajúcich transakcie, ale relatívne málo ľudí alebo procesov pozerajúcich sa na výsledky týchto transakcií. V systémoch DSS alebo Business Intelligence vo všeobecnosti je situácia práve opačná, teda len relatívne málo procesov naozaj spúšťa query a vytváranie reportov, ale konečné výsledky sú doručené mnohým ľuďom na ďalšiu analýzu.

Používatelia DSS teda na rozdiel od používateľov OLTP nevykonávajú mnoho malých transakcií meniacich dáta, ale kladú oveľa menej, zato rozsiahlych a náročných požiadaviek na čítanie uložených dát.

Rozmenené na drobné

Pozrime sa na niektoré z funkcií, ktoré sú nevyhnutné pre databázové systémy využívané v OLTP, ktoré však zároveň nemajú (také) uplatnenie vo svete BI.

UZAMYKANIE ZAZNAMOV

– v prostredí mnohých súčasne bežiacich transakcií sa môže stať (a nezriedka sa i stáva), že viacerí používatelia chcú v jednom čase pristupovať k tým istým uloženým údajom, ba dokonca ich meniť. Aby v takýchto prípadoch nedochádzalo k nepredvídateľným situáciám, databázové systémy implementujú rôzne techniky chránenia záznamov pred súčasnou zmenou viacerými používateľmi, resp. transakciami.

Všetky takéto techniky, ktoré zabraňujú konfliktným situáciám, ako sú zamykanie dát na úrovni riadkov alebo detegovanie deadlockov, strácajú význam v prostredí, kde sa dáta v jednom čase len čítajú.

COMMIT

– transakcia je vo svete databázových systémov považovaná za atómicú, ďalej nedeliteľnú operáciu nad dátami. Keďže v rámci jednej transakcie možno manipulovať s viacerými dátami (rôzne záznamy alebo aj rôzne tabuľky), požadované zmeny nie sú vykonávané hneď, ale až na konci transakcie – po jej potvrdení tzv. commit. V transakčných prostrediach je technika potvrdzovania uskutočnených transakcií alfou a omegou konzistencie dát.

Naopak, v prostrediach, kde sa vo väčšine prípadov dáta len čítajú, stráca táto pomerne komplikovaná technika zmysel – nehovoriac o možnostiach do seba vnorených transakcií alebo transakcií roztrúsených v distribuovanom prostredí cez viaceré databázové systémy, keď hovoríme už o tzv. 2-phase commit.

RECOVERY

– schopnosť zotavenia sa databázového systému z takmer akéhokoľvek výpadku je dnes na pomerne slušnej úrovni. Okrem iného je to možné vďaka prepracovanému systému vytvárania tzv. kontrolných bodov (checkpoints), keď je databáza spolu so svojimi údajmi v konzistentnom stave vzhľadom na prebiehajúce transakcie. Medzi takými kontrolnými bodmi sa zaznamenávajú všetky transakcie, resp. zmeny nad databázou. Po prípadnom výpadku sa tieto údaje o zmenách využívajú na vrátenie databázy do pôvodného stavu pred výpadkom.

Opäť v prípade prostredia DWH, kde sa dáta s výnimkou nejakého dávkového loadovania len čítajú, môže byť takáto funkcionality minimálne značne zjednodušená.

Uviedli sme tu zopár funkcií, ktoré sú implicitne súčasťou tradičných relačných databáz, využívaných nielen v primárnych systémoch OLTP, ale aj ako dátová platforma na budovanie

dátových skladov. Ako sa ukázalo, tieto funkcie nemajú uplatnenie v prostrediach, kde k dátam pristupujeme s cieľom čítania a analyzovania.

Sú tu však aj ďalšie vlastnosti tradičných databázových systémov, ktoré treba zmeniť a vyladiť v prípade použitia pri budovaní DWH.

BUFFER/PAGE SIZE – buffery sú časti pamäte využívané pri práci s dátami. Keďže systémy OLTP prevažne s veľkým množstvom malých transakcií, buffery sú nastavené na malú veľkosť a veľký počet. Pri DWH optimálne nastavenie spočíva v menšom počte bufferov väčšej veľkosti. Rovnaká situácia je v prípade nastavenia tzv. Page Size, čo je v skutočnosti najmenšia jednotka (množstvo dát), s ktorou príslušný databázový systém pracuje.

INDEX – jedna zo zložitých súčastí administrácie databázových systémov je správne ladenie indexov, ktoré slúžia na rýchlejšie vyhľadávanie dát. Existuje viacero mechanizmov vytvárania a spravovania indexov, ktoré treba starostlivo vyberať podľa konkrétneho použitia v aplikáciách prístupujúcich k databáze. Keďže spôsoby prístupu k dátam sú rozdielne v systémoch OLTP a DSS, aj administráciu indexov treba príslušným spôsobom vyladiť.

Čo z toho plynie?

Pri použití RDBMS ako dátovej platformy pre DWH, ktorý je základom pre BI, resp. systémy DSS, tak mnohé zapracované technológie zostávajú nevyužitú, naopak, môžu dokonca zapríčiniť neefektívny prístup k dátam, a teda pomalšie reakcie na požiadavky používateľov. Napriek tomu, že RDBMS poskytujú možnosti na individuálne nastavenia a prispôsobenie konkrétnemu využitiu v príslušnom podnikovom prostredí, ktorými sa dajú dosiahnuť výrazné zlepšenia časových reakcií na požiadavky používateľov, stále trpia na dôsledky konceptu využívania, pre ktoré boli navrhované (teda pre transakčné systémy).

Na druhej strane bohatá funkcionalita, ktorú dokážu ponúknuť tradičné relačné databázové systémy, ich predurčuje na použitie v každej oblasti, kde ide o ukladanie a spracúvanie dát. Navyše do súčasných RDBMS sú dopĺňané stále ďalšie a ďalšie funkcie, ktoré si vyžaduje doba. Príkladmi môžu byť interná podpora formátu XML a príslušného hierarchického indexovania podľa štruktúry XML, vnorené tabuľky, špeciálne dátové typy na podporu audio, video, textových a geografických dát. Ruka v ruke s takými rozšíreniami idú i rozšírenia jazyka SQL, aby bolo vôbec možné rozumným spôsobom využívať tieto funkcie.

Ak sa na uvedené fakty pozrieme v kontexte požiadavky na vybudovanie dátového skladu ako dátovej základne pre systémy na podporu rozhodovania, nevyhnutne sa vynárajú neodbytné otázky. Do akej miery a či vôbec existujú požiadavky na bohaté funkcie poskytované tradičnými RDBMS v oblasti BI? **Je vôbec možné, aby jedna technológia bola vhodná pre také rozdielne oblasti práce s dátami, ako sú OLTP a BI?**

Pred časom panoval medzi IT odborníkmi názor, že napriek uvedeným známym skutočnostiam RDBMS po príslušných vyladeniach poskytujú výkon dosť dobrý aj pre oblasť BI. Cena zvýšenej administrácie bola akceptovateľná a vyvažovaná zachovaním doteraz používaných známych databázových platforiem, a teda znovuvyužitím skúseností získaných z administrácie systémov OLTP. Doba však prináša čoraz väčšie objemy dát, ktoré končia v dátových skladoch. **Pri týchto narastajúcich dátach, a teda aj rastúcich požiadavkách na databázové systémy sa v stále väčšej miere prejavujú dôsledky toho, že RDBMS neboli vyvíjané na použitie mimo transakčného sveta.**

Existuje riešenie?

Riešenie tejto situácie poskytujú dátové platformy, ktoré sa nesnažia prispôbovať existujúce RDBMS na podmienky DSS, ale sú od začiatku navrhnuté a budované na používanie v systémoch DSS. Neumožňujú síce ich využívanie v prostrediach OLTP, ale poskytujú vysoko efek-

tívne prostredie na ukladanie veľkého množstva dát, ktoré sú pripravené na následné rozsiahle analytické query pri súčasnej minimalizácii časovej reakcie.

Akým spôsobom teda možno dosiahnuť efektívny spôsob ukladania dát v prostrediach využívaných v BI/DSS? Použitím technológie, ktorá

- eliminuje náročné techniky, nevyhnutné v prostrediach OLTP, ale spomaľujúce v BI,
- prednastaví parametre na použitie v BI,
- optimalizuje algoritmy na používanie v prostredí BI,
- použije vhodné dátové štruktúry na ukladanie veľkého objemu dát.

Aké sú hlavné výhody, ktoré môžeme očakávať pri použití takejto alternatívnej technológie namiesto tradičných databázových systémov, použiteľných aj v OLTP?

- Zvýšenie výkonu, teda podstatné skrátenie času potrebného na spracovanie rozsiahlych query a analytických úloh
- Zjednodušenie administrácie – použitie technológie, ktorá eliminuje veľa nevyužitelných vlastností, môže významne znížiť nároky na administráciu

Jedinou otázkou pre manažerov zodpovedných za stratégiu implementácie riešenia BI, budovania dátových skladov a systémov na podporu rozhodovania teda zostáva, či dokážu na základe objektívnych kritérií opustiť tradičné prístupy a zvoliť správnu technológiu, ktorá zabezpečí vysoký výkon pri súčasnom znížení nákladov na prevádzku. Technológie sú pripravené, ste pripravení aj vy?



■ DUŠAN KRCHO, SAS Institute
Dusan.Krcho@svk.sas.com

SAS Intelligence Storage

Takmer všetky podniky riešia stále páčivejšie problémy súvisiace s enormným nárastom objemu dát: rastúce nároky na úložný priestor, dlhotrvajúce reporty, predlžujúce sa reakcie na ad hoc query, neustála, čoraz zložitejšia a nákladnejšia administrácia.

Keďže najväčšie objemy dát sa v podnikoch sústreďujú v dátových skladoch a datamartoch a slúžia tak ako vstup pre aplikácie analýzy dát a podpory rozhodovania, tieto problémy sa stávajú vo svete Business Intelligence (BI) ešte vypuklejšími.

Niektorí IT manažéri sa už presvedčili, že investície do nákupu nového hardvéru neriešia problémy databázových systémov navrhovaných primárne pre transakčné systémy. Hľadajú preto systémové riešenie problematiky ukladania veľkých objemov dát pre oblasti BI a analýzy.

Spoločnosť SAS ponúka technológiu na ukladanie veľkého objemu dát, od začiatku navrhovanú na použí-

tie v oblasti BI. Špecifický návrh umožnil maximálne zjednodušiť a zefektívniť architektúru systému na ukladanie dát, čoho výsledkom sú výhody v podobe rýchlejších query, menších nárokov na hardvér, menších nárokov na úložný priestor a zníženej administrácie. Všetky tieto faktory vedú k zníženiu celkových nákladov na prevádzku pri súčasnom poskytovaní vyššieho výkonu pre koncové systémy.

Vnútorňa architektúra je postavená tak, aby mohla v maximálnej miere využívať paralelný prístup, či už pri spracúvaní dát, alebo ich fyzickom uložení. Takto dokáže naplno využiť možnosti súčasných multiprocessorových hardvérových serverov a vyhýbať sa potenciálnym úzkym miestam, najmä pokiaľ ide o priepustnosť dát. Navyše špecificky navrhnutá architektúra je plne škálovateľná, čo zabezpečuje plynulý rast výkonu spolu s rastúcim počtom používateľov a ich požiadavkami.

Spoločnosť SAS ponúka takúto špecializovanú technológiu nielen na prácu s dátami relačného charakteru, ale aj na spracúvanie dát v multidimenzionálnej štruktúre OLAP. Obe technológie pod spoločným označením

SAS Intelligence Storage sú pre koncového používateľa prístupné transparentným spôsobom cez štandardné rozhrania. Napriek skutočnosti, že vnútri je architektúra navrhnutá podľa špecifických požiadaviek, navonok sú dáta prostredníctvom štandardných rozhraní (ODBC, JDBC, OLE DB) k dispozícii presne tak, akoby boli uložené v tradičných relačných databázových systémoch (RDBMS).

Príklady z praxe ukazujú, že nahradením systémov RDBMS v oblasti BI špecializovanou technológiou možno pri zachovaní rovnakého hardvéru zrýchliť komplexné query až niekoľko desiatok krát. Pre business používateľov to znamená, že na výsledky svojich analýz nemusia čakať niekoľko hodín, ale len pár minút. Rovnako nároky na diskový priestor možno zredukovať zhruba na tretinu, čo šetrí priame náklady, keďže kapacita súčasných diskových polí aj pri aktuálnom náraste objemu dát vystačí dlhšie, ako sa predpokladalo. Výrazné zníženie administrácie umožňuje IT oddeleniu venovať sa činnostiam, ktoré prinášajú podniku vyššiu pridanú hodnotu.