Databázové architektury: současné trendy a jejich vztah k novým požadavkům praxe

Jaroslav Pokorný

KSI MFF UK
Malostranské nám. 25, 118 00 Praha
Tel: 221914265
e-mail: pokorny@ksi.ms.mff.cuni.cz
www: http://kocour.ms.mff.cuni.cz/~pokorny/

Klíčová slova: databázové architektury, vrstevnatá architektura databázového systému, proudy dat, nejistá data, nepřesná data, dolování dat, OLAP, bezdrátové vysílání dat, mobilní počítání, prostor dat

Abstrakt: Ve vývoji databázových systémů se dnes uvádí dvě hnací síly: web a jednotlivé vědy či obory, jako jsou fyzika, biologie, medicína a inženýrství. Objevují se data naměřená pomocí čidel či monitorovacími systémy, která ovlivňují chod komputerizované společnosti. Správa takových dat je založena buď na tradičních souborových technikách nebo využívá komerční SŘBD. S produkcí rozsáhlých kolekcí dat a růzností jejich zpracování, často v reálném čase, nároky na jejich uložení a zpracování vzrůstají. Nové ITC technologie a metody podporující tyto požadavky zahrnují senzorové sítě, zpracování proudů dat, zpracování nejistých a nepřesných dat, práci se znalostmi a inteligentní analýzu dat, jakož i bezdrátový přenos dat a mobilní počítání. Ukazuje se, že tradiční architektura univerzálního SŘBD pouze obtížně vyhovuje těmto trendům a že je třeba nalézat nová řešení. Objevují se specializované architektury, hybridní či konfigurovatelné SŘBD spojované do sítí. Článek diskutuje současné trendy a pokroky v těchto směrech a pokouší se je osvětlit na příkladech z praxe.

1 Úvod

Svět dat se bezpochyby mění, zvláště pak podstata zdrojů informací. Všechny tyto změny mají význačný vliv na databázové potřeby a v důsledku toho na otázky, kde se oblast databází nachází a kam by měl směřovat její vývoj. Abiteboul et. al. ve zprávě [1] zdůrazňují dvě hlavní hnací síly v oblasti databází: Internet a jednotlivé vědy, jako jsou fyzika, biologie, medicína a inženýrství. Tyto vědy produkují rozsáhlé[1] a složité množiny dat, které požadují pokročilejší databázovou podporu než poskytují současné komerční systémy. Objem dat se přibližně zdvojnásobuje každým rokem a pohybuje se ve škále, kde jednotkou objemu je 1 PB [8]. Objevuje se také potřeba nových mechanismů pro integraci informací.

Další trend, existující již od 60. let, se týká průmyslových odvětví reagujících na stále se zvyšující environmentální požadavky od zákazníků, autorit a vládních organizací. Zvyšují se nároky na bezpečnost. V obou případech směřuje vývoj ke vzniku nových monitorovacích systémů. Nové funkce jsou integrovány do běžných podnikových řídících systémů, systémů státní správy, ale i dalších, jako jsou např. systémy řízení digitálních domácností.

[1] Např. databáze BaBar (obsahuje nukleární data), považovaná za největší na světě, měla 5. 11. 2004 vice než 895 TB dat uložených v 847149 souborech.

Problémy se zpracováním uvažovaných dat lze dobře dokumentovat na příkladu environmentálních dat. Chtějí-li uživatelé vyhledávat a používat environmentální informace, zjišťují podle [20], že:

(1) data neexistují nebo nejsou postačující; někdy je požadována syntéza nebo reprodukce dat.

(2) data nejsou dodavateli dat popsána, je tedy obtížné je lokalizovat, nebo se na ně odkazuje podle specifických klasifikačních kritérií, která závisí na aplikační doméně.

(3) k datům je obtížný přistup; mohou být soukromá, nebo jsou drahá, nebo vyžadují drahou úpravu formátu.

(4) přistupované kolekce dat je obtížné použít, protože jsou nekonzistentní nebo nekompatibilní; např. pro přístup k dlouhým časovým řadám nemohou být použity standardní techniky pro běžné kolekce dat, což činí související časové řady nekompatibilní.

(5) kvalitu vybraných dat je obtížné ohodnotit. Často je obtížné porovnat data vytvořená použitím různých vědeckých modelů kvůli nedostatečné dokumentaci výpočetního procesu v pozadí.

Databázová komunita se zaměřuje na uložení informací, jejich organizaci, správu a přístup v softwarových architekturách - *systémech řízení bází dat* (SŘBD). Vývoj SŘBD je vždy řízen novými aplikacemi, technologickými trendy a novou synergií mezi vztaženými oblastmi, ale i inovací v oblasti samotné. Problémy (1) - (5) jsou přirozenou součástí dnešního databázového výzkumu a vývoje. Nové databázové technologie nasazené v praxi mohou pomoci při překonávání těchto problémů.

Vývoj SŘBD ovlivňuje několik technologických aspektů. Zaměříme se na chvíli na vědecká data. Senzorové sítě produkující část těchto dat sestávají z velkého množství levných zařízení, z nichž každé je zdrojem dat měřícím nějakou kvantitativní veličinu, např. umístění objektů nebo okolní teplotu. Zpracování takových dat je obvykle úplně odlišné od dat uložených v podnikových databázích. Data se objevují ve vysokorychlostních proudech a dotazy nad těmito proudy potřebují být zpracovány online způsobem, který umožňuje odezvu v reálném čase. Ve srovnání se zpracováním podnikových dat jsou tato data navíc nejistá nebo nepřesná. Vědecká měření totiž obsahují zcela běžně chyby. Další aspekty těchto dat zahrnují nejasnou formulaci dotazů založenou na známých technikách, jaké se používají např. v klasických databázích. Často nejsme schopni formulovat dotaz v SQL i když jsme přesvědčeni, že v našich datech je skryto něco zajímavého a že je tedy možné se na to zeptat. V takových situacích zjevně schází sémantika. Popsat sémantiku dat na úrovni metadat (pokud možno formálně) je více než žádoucí. V kontextu analýzy dat pomocí dotazování mohou také pomoci techniky dolování dat a OLAP.

Samozřejmě není překvapující, že v mnoha případech jsou k dispozici pouze tradiční souborově orientovaná řešení. Např. systém CORIE (Columbia River Estuary) produkuje ve svých simulacích 5GB předpovědí každý den [5]. Jeho repozitář metadat nicméně neobsahuje schéma dat, ani formáty souborů, knihovny pro databázový přístup či dokonce XML schémata, na kterých by se uživatelé mohli dohodnout. Ve spojení s Internetem, webovými službami se taková řešení nezdají udržitelná.

Účelem článku je prezentovat hlavní směry ve vývoji databázových architektur v souvislosti se vznikem nových typů dat a nových typů jejich zpracování. V sekci 2 stručně popíšeme vrstevnatou architekturu SŘBD tak, jak ji navrhli v 80. letech Härder a Reuter [9]. Lze ji považovat za referenční, protože veškerý další vývoj databází znamenal uspět v její realizaci. V sekcích 3-7 stručně diskutujeme technologie a požadavky praxe, které zásadně ovlivňují dnešní databázové architektury, tj. senzorová data a senzorové sítě, zpracování proudů dat, uvažování nejistých a nepřesných dat, dolování dat a OLAP a taktéž bezdrátové vysílání dat a mobilní počítání. Sekce 8 je již věnována novým databázovým architekturám. Zmíníme některé příklady architektur a popíšeme krátce jejich charakteristiky. V sekci 9 zobecníme úvahy nad architekturami SŘBD pomocí pojmu datového prostoru. Dostaneme se tak do širšího kontextu současného rozvoje zpracování a správy dat. Závěry stručně shrnují diskutované myšlenky.

2 Vrstevnatá architektura SŘBD

Každý, kdo používá relační databáze, si je vědom toho, že tabulky dat vyskytující se na vrcholu nějakého databázového systému jsou v jistém smyslu virtuální. Konkrétněji, poskytují logickou datovou strukturu vhodnou pro uživatelsky orientované zpracování dat v databázi. Jde pouze o jednu, nejvíce viditelnou vrstvu databázového systému. Härder a Reuter [9] navrhli mapovací model sestávající z pěti vrstev. Tabulka 1 upravená na základě [10] ukazuje těchto pět vrstev podrobněji. Na každé úrovni abstrakce můžeme pozorovat objekty, se kterými se má pracovat, a jednotlivé funkce implementující zobrazení mezi dvěma po sobě následujícími vrstvami. Např. neprocedurální přístup ve vrstvě L5 poskytuje tabulky a příkazy pro jejich manipulaci formulované obvykle v jazyku SQL. Vrstva L2 zajišťuje rozdělení lineárního adresového prostoru na vnější paměti do různých typů stránek. Mezi objekty ve vrstvě L3 můžeme nalézt datové struktury podporující indexaci, např. známé B-stromy pro řetězce znaků a čísla nebo R-stromy pro prostorová data. Jdeme-li po vrstvách směrem nahoru, objekty a asociované operace se stávají složitější, mohou se také vyskytnout přídavná integritní omezení.

	Úroveň abstrakce	**Objekty**	**Pomocná zobrazení dat**
L5	neprocedurální přístup	tabulky, pohledy, řádky	popis logického schématu
L4	záznamově-orientovaný navigační přístup	záznamy, množiny, hierarchie, sítě záznamů	popis logického a fyzického schématu
L3	správa záznamů a přístupových cest	fyzické záznamy, přístupové cesty	tabulky volného prostoru, DB klíčů, tabulky pro překlad
L2	rozdělení do stránek	segmenty, stránky	buffery, tabulky stránek
L1	správa souborů	soubory, bloky	direktoráře

Tab. 1: Popis hierarchie zobrazení pěti vrstev SŘBD

Koncept vícevrstvé architektury uvažuje svou ideální implementaci pomocí stroje, který má *k* vrstev. Ačkoliv počet pět je v architektuře považován za dobrý kompromis, v praxi se vyskytují problémy s provozem takové architektury. Zjednodušení složitosti vrstev na jedné straně zvyšuje režii při běhu systému na straně druhé. V důsledku toho jsou vyvíjeny různé způsoby optimalizace provozu SŘBD a počet vrstev je pro některé systémové funkce redukován.

Vývoj vrstvy L5 v posledních 10 letech vyústil do specifikace tzv. objektově relačního (OR) datového modelu. Jeho část je standardizována ve standardu SQL:1999 [11] resp. SQL:2003 [12]. V OR modelu mohou mít tabulky strukturované komponenty svých řádků, sloupce mohou dokonce mít uživatelsky definovaný typ. Do této kategorie patří prostorová data, časové řady nebo texty. Pro některé datové typy, např. VITA (video, image, text, audio), existují pro manipulaci jejich instancí standardizované množiny predikátů a funkcí. Tento "rozšiřitelný" přístup měl za následek vznik tzv. *univerzálních SŘBD* koncem 90. let. Jádro těchto databázových strojů bylo rozšiřováno volně spřaženými přídavnými moduly (komponentami) pro každý nový datový typ. Výrobci vedoucích SŘBD tyto komponenty nazývají extendery, datablady a cartridge. Připomeňme, že prostorové a textové komponenty patří mezi nejúspěšnější v tomto přístupu. Díky složitějším datovým strukturám než jsou řádky klasických relačních tabulek, poskytují OR SŘBD šanci pro využití v oblasti ukládání a zpracovaní vědeckých dat, kde pole (ARRAY) patří mezi klíčové datové struktury.

Možnost uživatelsky definovaných typů uvedla do implementace architektury SŘBD množství vážných problémů, zvláště pak v případě konceptuálně zcela různých datových typů, jako jsou VITA. U typů VITA je možné společně využít maximálně vrstvu L1, další je třeba implementovat pro každý typ zvlášť. Otevřeným problémem zůstává, jak integrovat tyto typy do společného rámce architektury SŘBD. Implementace nových přístupových cest, jako jsou speciální typy indexů, vede obvykle k modifikacím jádra SŘBD, např. kompilátoru SQL, optimalizátoru dotazů apod. Takové změny jsou v implementaci a testování velmi drahé, časově náročné a náchylné na chyby. Jako příklad uveďme nové přístupové metody a uživatelsky definované typy, např. pro sousedící tok dat z proudících datových zdrojů.

Každý výrobce používá k otevření architektury hostitelského systému do jistého stupně různý přístup. Cartridge Oracle jsou omezeny na integraci sekundárních indexů. V IBM DB2 extenderech existuje rámec pro indexaci nových typů dat omezený pouze na B-stromy. To znamená, že taková indexace může přinést zlepšení vyhodnocení jen pro některé typy dotazů. Jinými slovy řečeno, nová funkčnost je podporována, ale pouze pro omezenou třídu uživatelských požadavků.

Zdá se, že přínos takového software je patrný hlavně v případě požadavků, které mohou být dekomponovány do relativně nezávislých částí vyhodnocovaných odděleně v jádru SŘBD a v modulu, který implementuje specifický typ dat. Rámce zpracování jsou tedy buď příliš složité nebo ne dost pružné, aby se vypořádaly s širokým okruhem uživatelských požadavků na doménově specifické přístupové metody. S většinou těchto pokusů může být skutečně bezešvé integrace dosaženo obtížně. Dnešní implementace vrstevnaté SŘBD architektury tedy nejsou pro nové požadavky postačující a v případě univerzálních SŘBD selhávají.

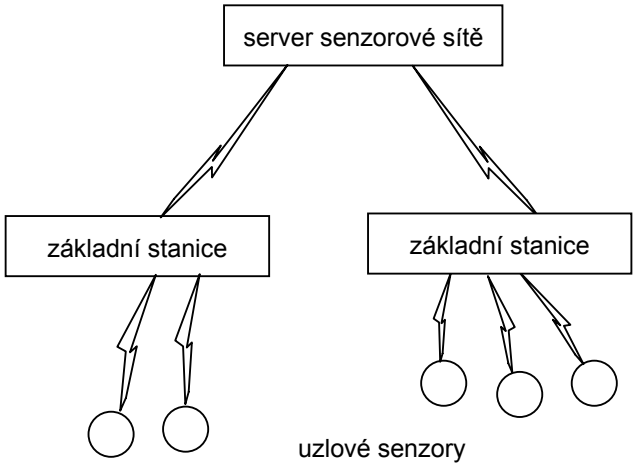
Dalším problémem tradičních řešení je, že jsou dostupná zejména pro statické aplikace. Tím, že např. environmentální data mají z definice časovou a prostorovou komponentu, mohl by být environmentální systém implementován na vrcholu prostorově temporálního SŘBD. Takový databázový software je dnes, bohužel, teprve ve vývoji.

3 Senzorová data a senzorové sítě

Mezi novými ITC technologiemi zaujímá v kontextu nových požadavků na databáze čelné místo levná mikrosenzorová technologie, která umožňuje většině objektů podávat v reálném čase zprávy o jejich atributech, jako je teplota, tlak, stav nebo umístění, (např. pomocí globálního pozičního systému). Tyto informace budou podporovat aplikace, jejichž hlavním účelem je to monitorovat tyto atributy [3].

Senzorové sítě produkují důležité zdroje dat a vytvářejí nové požadavky na správu dat. Senzor je bezdrátové zařízení s vlastním zdrojem. Takové zařízení spotřebuje více energie na komunikaci než na výpočty v jednotlivých uzlech. Ve skutečnosti se tak senzorová síť stává novým druhem databázového stroje, jehož optimální využití požaduje, aby operace byly tlačeny co možná nejblíže k datům. Ve složitějším případě mohou být senzory a/nebo uživatelé mobilní.

Zpracování senzorových informací nastoluje nejzajímavější databázové problémy do nového prostředí, s novými omezeními a příležitostmi. Rozsáhlé kolekce dat generované senzory, budou distribuovány po světě, přičemž data budou vznikat a zanikat dynamicky. Jinými slovy řečeno, senzory mohou produkovat nepřetržité, možná nekonečné, proudy dat.



Obr. 1: Vícevrstvá architektura senzorové sítě

Senzorové sítě jsou na první pohled podobné distribuovaným databázím rozšířeným o vlastnosti souvisejícími s využíváním reálného času. Důležitým rozdílem ale je, že míra vyhodnocování dat

vytvořených v síti senzorů je vyšší, než se typicky uvažuje u distribuovaných SŘBD. To láme tradiční paradigma integrace informací, protože neexistuje žádný praktický způsob extrakce a natahování dat do společné databáze ke každému jejich výskytu. Rovněž musí být redefinovány strategie zpracování a optimalizace databázových dotazů. Obrázek 1 ukazuje příklad vícevrstvé architektury senzorové sítě.

4 Zpracování proudů dat

Správa dat přicházejících ze senzorů založená výlučně na tradičním modelu „ukládání a dotazování“ obvykle nemůže efektivně zacházet s objemem a rychlostí proudících dat, jejichž hodnoty mohou existovat jen na krátký okamžik. Tradiční SŘBD jsou pro zacházení s proudy dat nevhodné z mnoha důvodů [3]:

- senzorové uzly produkují a odesílají data nepřetržitě bez ohledu na to, existují-li přímé požadavky na tato data,
- dotazy nad příslušnými kolekcemi dat mohou být méně časté než vkládání dat do kolekcí,
- vytvořená data jsou často zpracovávána v reálném čase, protože mohou reprezentovat události, které potřebují okamžitou reakci,
- dotazy se zpracovávají nepřetržitě, protože proudy dat nikdy nekončí, takže mohou „vidět“, jak se mění podmínky systému během jejich volání,
- kvůli omezením na paměť, nemůže být proud dat uložen celý na vnější médium,
- protože proudy dat jsou potencionálně nekonečné, mohou být použity pouze neblokující operátory,
- jestliže množina dat ze senzorů není jako vstup do operátoru celá dostupná, pak se operátor zablokuje.

V důsledku toho se objevily *systémy zpracování proudů dat* (Stream Data Management System - SDMS), viz např. [6]. Procesor (stroj) pro zpracování proudů dat je pak příkladem nové databázové architektury, která dovoluje volání dotazů, výpočtů a akcí na proudících datech v reálném čase. Procesor by měl akceptovat proudově orientované spojité dotazy zapsané v notaci SQL a spouštět je nad proudy událostí s výstupem v reálném čase. V SDMS je zpracování realizováno z větší části ve vnitřní paměti, operace čtení a zápisu na disk jsou volitelné a mohou být v mnoha případech zvládnuty asynchronně.

Např. v současném pilotním systému Streambase vyvinutém Stonebrakerem [19] v r. 2005 je možné analyzovat 140000 zpráv za sekundu, zatímco běžný relační SŘBD by zvládl v témže čase pouze 900 zpráv.

5 Přístup k nejistým a nepřesným datům

K problémům správy proudů dat se přidružují mnohé další problémy. Jakákoliv vědecká měření jsou běžně zatížena chybami. Např. údaje o umístění pohybujících se objektů zahrnují prvek nejistoty co se týče současné pozice objektů. Individuální senzory nejsou spolehlivé a v důsledku toho je celá bezdrátová komunikace nespolehlivá. Používají se proto přístupy, které poskytují přesnější odhady údajů o prostředí. Pro fúzi senzorových dat z více senzorů se někdy objevuje použití přístupů, jako jsou fuzzy množiny nebo Dempster-Shafer teorie a další techniky umělé inteligence [14].

Tradiční SŘBD se používaly pro zpracování podnikových dat, která jsou typicky reprezentována čísly a řetězci znaků. Údaje jsou přesnými veličinami - adresa, prodané množství, zůstatek na účtu, stav zásob apod. V důsledku toho nemají současné SŘBD žádné prostředky pro zpracování přibližných dat a nepřesných dotazů. Také posloupnosti a obrázky vyžadují aproximativní zpracování založené na podobnosti, metrikách apod.

Aby se zvýšila kvalita dat, objevuje se nový model dat, který zachovává původ dat a historii jejich zpracování, jakýsi *rodokmen* (*lineage*) objektů a procesů [4]. Pro zajištění, aby např. environmentální data byla co nejvíce a nejlépe využita, by měli producenti dat zahrnout údaje o původu (a hodnověrnosti informací) do metadat připojených k základním datům. Jinou aplikací rodokmenu dat může být systém personálních dat, kdy chceme zachovat různé verze dokumentu, nebo vztažené maily apod. To požaduje sofistikovanější techniky pro zpracování metadat na databázové úrovni.

6 Dolování dat a OLAP

Naměřená data většinou potřebují být analyzována, aby bylo možné obdržet informace nutné pro rozhodování, Typickým příkladem je environmentální management. Ve srovnání s jednoduchými formami pravidelností či nepravidelností zjišťovaných pomocí statistických metod, metody dolování dat mohou nalézt složitější hypotézy, které obsahují jak numerické tak logické podmínky.

Historicky vzato se dolování dat zaměřovalo na účinné způsoby objevování modelů existujících množin dat. Tyto modely mají odhalit nějaké užitečné aspekty dat při zakrytí detailů neužitečných pro danou aplikaci. Mnohé výzkumné komunity vyvinuly algoritmy, které provádějí operace jako klasifikaci, shlukování, objevování asociačních pravidel a sumarizaci. Tyto techniky se stávají novou částí produktů hlavních dodavatelů SŘBD a většina z nich je aplikovatelná i v oblasti vědeckých dat. Problémem ovšem je, že mnoho algoritmů je super-lineárních (např. pro zpracování n bodů mají složitost $O(n^2)$ nebo $O(n^3)$), což v případě kolekcí zmiňovaných v úvodu může být časově neúnosné. Řešením jsou např. aproximativní algoritmy a využití paralelismu.

Často jsou postačující OLAP techniky. Např. se požaduje zjistit trendy ve vývoji teploty a tlaku v nějakém prostředí. Odvození takových informací využívá typicky znalost minulých teplot a tlaků uložených v databázi a zpracovávaných podél časové dimenze.

Současný zájem o kombinaci technologie dolování dat a SŘBD směřuje k objevování nových přístupů k uložení množin dat, které mají být dolovány, aby se dolování dat optimalizovalo. Velikost kolekcí neni problémem pouze u vědeckých dat. Podle Gregova zákona se odhaduje, že objem podnikových dat vzrůstá na dvojnásobek každých 9 měsíců. Datové sklady v rozsahu TB nejsou již výjimkou.

Směry výzkumu zahrnují

(1) vícerozměrné OLAP pro objevování neobvyklých vzorů v proudech dat;

(2) dolování shluků a extrémních hodnot (outliers) v proudech dat; a

(3) jednoprůchodové klasifikační metody pro dolování proudů dat.

7 Bezdrátové vysílání dat a mobilní počítání

Vysílání dat je atraktivní alternativou k přístupu „on demand“, protože jím lze šířit data zároveň k velkému množství klientů za pevnou cenu. To je vhodné pro služby založené na umístění objektů, které vykazují silnou časovou a prostorovou lokalizaci, ve které sousedící klienti mají tendenci vyhledávat v jisté časové periodě stejný druh informací [21].

Např. environmentální data mají být často šířena k uživateli včas, vždy a kdekoliv. Pak v tomto kontextu nabývá na důležitosti mobilní prostředí. Např. při periodickém vysílání jsou data šířena periodicky po bezdrátovém kanálu. Mobilní klient naslouchá vysílání a stahuje z kanálu vhodná data v souladu s dotazem, který vydal uživatel nebo na základě uloženého zájmového profilu klienta. Sítě, ve kterých se tato zařízení umisťují, by samozřejmě měly být také schopné vydávat odpovědi na neperiodické dotazy.

Zdá se, že umístění v časoprostoru se stává velmi důležitou vlastností dat a zavádí novou dimenzi pro metody přístupu k datům. Tradiční metody přístupu k datům nejsou v tomto případě vhodné. Cílem

současného výzkumu je redefinovat některé dobře známé techniky, např. zpracování prostorových dotazů, do mobilního prostředí se speciálním důrazem na vysílání dat.

Data, která jsou vysílána, zahrnují také senzorová data. Senzory rozmístěné v prostředí mohou vysílat svá data periodicky nebo když nastane zajímavá událost. Na rozdíl od tradičního počítání nemohou klientská zařízení vytvářet požadavky na data od senzorů. Místo toho naslouchají klientská zařízení vysílacím kanálům pasivně. Senzory jsou tedy v komunikaci iniciativní. Senzory mohou vysílat data periodicky, jestliže měří nějaký kontinuální jev produkující data, nebo pouze když se vyskytne určitá událost. Tou může být situace, když se nějaká RFID značka[2] dostane do dosahu senzoru.

Senzory vyšší úrovně v senzorové síti mohou předzpracovat senzorová data nižší úrovně a vysílat pak tyto odvozené informace do klientských zařízení. Aby byla úspěšná, mohou taková zpracování požadovat modifikované databázové techniky.

Vedle toho představují mobilní zařízení ještě další kategorii aplikací [16]: kešování relevantních částí rozsáhlé množiny dat na menším zařízení s omezenou funkčností. Mobilní zařízení lze tedy považovat za keš pro globální množinu dat. Tento model má atraktivní vlastnosti - jako je schopnost rozšířit množinu lokálních dat o vstupy tak, jak jsou používány nebo jak jsou potřeba. Mobilní telefonní infrastruktura požaduje podobné kešovací schopnosti k udržování komunikačních kanálů, přičemž data samotná jsou zcela pomíjivá; mohou být ztracena a v případě potřeby opět regenerována.

8 K novým databázovým architekturám

Databázová technologie se zdá být základní pro nasazení technologií uvedených v sekcích 3-7 v kontextu nových aplikací. Některé pokusy ovlivnit vývoj systémů s naměřenými daty databázovými specialisty existují již dlouho. Např. projekt Sequoia 2000 [17] hovoří o spolupráci mezi informatiky a environmentálními vědci při návrhu informačního systému příští generace pro řízení dat pro výzkum globálních změn. Hlavní přínosy databázového přístupu by měly být

- flexibilita bez složitosti,
- jednoduchost použití.

Databázový přístup navíc přináší příležitost propojit všechna data dohromady na uživatelské úrovni a zjednodušit tak jejich veškerou analýzu, např. pomocí technologie jako je dolováni dat.

Společný náhled na zmíněné problémy se týká architektury SŘBD. Dnešní SŘBD poskytují prakticky univerzální architekturu aplikovatelnou na mnoho různých typů úloh, tj. slovy Stonebrakera a Çetinteme [18], existuje "*jedna míra na všechno*". V nových architekturách SŘBD se očekávají spíše oddělené databázové servery "*šité na míru*" v soulady s požadavky jednotlivých typů aplikací. Vedle tradičních oblastí, jako jsou OLAP, sklady dat a vyhledávání v textech, jsou kandidáty pro taková zařízení:

- zpracování proudů dat,
- senzorové sítě,
- databáze vědeckých dat,
- nativní XML databáze.

Pokusili jsme se osvětlit charakteristiky prvních třech technologií.

Uvažujeme-li nativní XML databáze, řešení se separátním strojem jsou dnes populární. Härder prezentuje v [10] XTC architekturu (XML Transaction Controller), která dokazuje, že nativní XML SŘBD může být implementován v intencích pětivrstvé architektury. Vyskytuje se také možnost

[2] RFID (Radio-frequency identification) je současná technologie založená na radiových vlnách pro přesnou identifikaci objektů.

hybridního stroje. Aby mohla být integrována relační a XML data, vyvíjejí v IBM nový hybridní DB2 SŘBD umožňující pracovat s opravdovým nativním XML úložištěm, které je umístěno vedle sebe s relačním repozitářem dat. Na vrcholu obou datových úložišť (relačním a XML) sedí jeden hybridní databázový stroj. Podobné řešení je používáno mnoha dodavateli, kteří kombinují SŘBD skladu dat a obvyklý transakční SŘBD tak, že jsou sjednoceny společným syntaktickým analyzátorem. Taková architektura může být také inspirující pro implementaci netradičních typů dat.

Další přístup rozvíjí původní myšlenku rozšiřitelnosti SŘBD. Acker et al [2] vyvinuli specifikaci manažeru přístupu k datům - nové programátorské rozhraní pro několik vrstev jádra SŘBD. To umožňuje programátorovi přidávat nové datové struktury do SŘBD s minimálním úsilím.

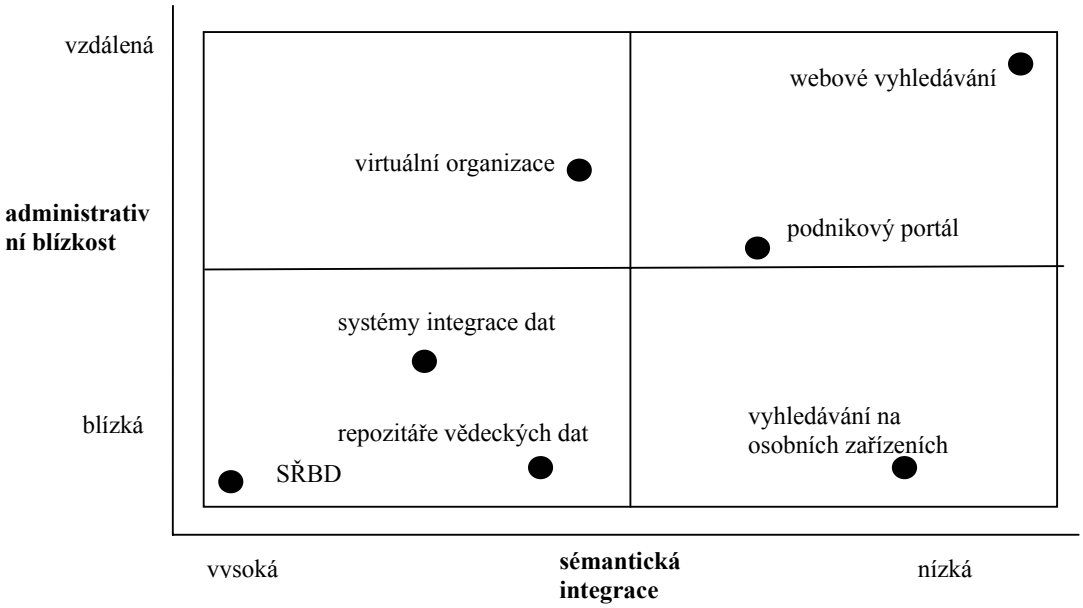
Existuje také třetí přístup, jak dosáhnout pružnosti ve zpracování data databázovým způsobem: vyprodukovat paměťový stroj, který je konfigurovatelný tak, že může být vyladěn podle požadavků individuální aplikace [16]. Existují v podstatě dvě vlastnosti, které řešení musí mít, aby se vypořádalo se širokým rozsahem aplikačních potřeb, které se dnes objevují:

- modularita a
- konfigurabilita.

Modulární SŘBD musí vývojáři umožnit použít nebo vyloučit některé své subsystémy v závislosti na tom, zdali je aplikace potřebuje. SŘBD musí být také konfigurovatelný vzhledem ke svému operačnímu prostředí: specifickému hardware, operačnímu systému a aplikacím, které ho používají.

9 Od databází k datovým prostorům

Veškeré snahy o nové architektury SŘBD naznačují, že současné požadavky na správu dat nelze řešit jejich umístěním do databáze jednoho (nejlépe relačního) SŘBD. Je tendence umisťovat data spíše do volně spřažených datových zdrojů, z nichž některé jsou řízeny relačním SŘBD, jiné však nikoliv. Datové zdroje se tak stávají součástí nějakého datového prostoru. Nejde ovšem o další přístup k integraci dat. Data v datovém prostoru spíše koexistují, sémantická integrace zde není podmínkou, aby části systému mohly být provozovány. Obrázek 2 převzaty z [7] ukazuje kategorizaci současných řešení správy dat ve dvou dimenzích. Administrativní blízkost indikuje, jak blízko jsou různé zdroje dat v termínech administrace. Sémantická integrace je míra, jak moc vztažena jsou schémata různých datových zdrojů, které se nějak integrují.



Obr. 2: Prostor řešení správy dat

Pojem datového prostoru je nová abstrakce popsaná v [7]. Vývoj odpovídajícího software – platformy pro podporu datového prostoru (PPDP) – je dnes uváděn jako hlavní položka programu oblasti správy dat, nebo jak se také říká, datového inženýrství.

PPDP nepřepokládá úplnou kontrolu nad daty v datovém prostoru. Dovoluje, aby data byla dále řízena jednotlivými systémy, pod které patří. Poskytuje vsak nove služby nad agregacemi těchto systémů. Součástí PPDP je katalog obsahující popisy *participantů* (zdrojů) v datovém prostoru a *vztahů* mezi nimi. Participanti jsou zdroje jako relační databáze, repozitáře XML dat, textové databáze, webové služby, senzory produkující data apod. Datový prostor předpokládá možnost modelovat libovolný vztah mezi participanty, v extrémním případě schéma výměny dat mezi zdroji. Nebo jde o jednoduchou závislost, kdy jeden zdroj je jen verzí jiného. Integrace datového prostoru se stále vyvíjí, a to na základě potřeby, nicméně data jsou stále přístupna např. v nejjednodušší formě pomocí klíčových slov.

Dotazování zahrnuje jak data, tak metadata, např. „Kde najdu něco o konferenci Moderní databáze?“, „Které zdroje mají atribut `délka`?“ Předpokládají se aproximativní odpovědi, či „co možná nejlepší“ výsledky v případě nedostupnosti některých zdrojů. PPDP by měla umožnit přejít v případě potřeby až k použití dotazovacího jazyka datového zdroje (je-li to možné) resp. přes placenou bránu.

Bylo by jistě zajímavé porovnat tento návrh s pojetím sémantického webu, který je založen na ontologiích a URL. V daném případě jde zřejmě o volnější integraci směřující spíše do hloubi ne nutně webových zdrojů. To však ukáže čas prostřednictvím projektů, které jsou momentálně řešeny.

10 Závěry

Našli bychom jistě další nová řešení databázové architektury, než ta, která jsme diskutovali v článku. Za zmínku stojí např. Service Oriented Database Architecture (SODA) vyvinutá Microsoftem a realizovaná v SQL Server DBMS 2005 [13], technologická řešeni jako je architektura Asymmetric Massively Parallel Processing (AMPP), nebo grid architektury.

Zmíníme už jen některé výzvy pro výzkum dat v následující dekádě, které formuluje známá specialistka IBM na optimalizaci dotazů Pat Selingerová [15]:

- znovu se zaměřit na vývoj architektury SŘBD a objevit způsob lepšího škálování, aniž by se obětovala dostupnost dat pro uživatele, či výkon,
- zkoumat, co všechno je správa obsahu, co je potřeba modelovat a jaké vytvářet nové modely,
- chápat zacházení s metadaty jako výzkum první kategorie**.**

Vše nasvědčuje tomu, že vývoj těchto témat a technologií má a bude mít důsledky, které ovlivní budoucí datově orientované systémy.

Poděkování

Práce byla částečně podporována Národním programem výzkumu – v rámci projektu Informační společnost 1ET100300419.

Literatura

[1] Abiteboul, S., et. al.: Lowell Database Research self-assessment. In: Communications of ACM, May 2005/Vol. 48, No. 5, 2005, s. 111-118.

[2] Acker, R., Pieringer, R., Bayer, R.: Towards Truly Extendible Database Systems. In: Proc. DEXA 2005 Conf., LNCS 3588, Springer-Verlag, 2005, s. 596-605.

[3] Amato, G., Caruso, A., Chessa, S., Masi V., Urpi, A.: State of the art and future directions in wireless sensor network's data management. 2004-TR-16, publikováno ISTI, 2004.

[4] Bose, R., Frew, J.: Lineage Retrieval for Scientific Data Processing: A Survey. ACM Computing Surveys, Vol. 37, No. 1, 2005, s. 1-28.

[5] Bright, L., Maier, D.: Deriving a Managing Data Products in an Environmental Observation and Forecasting System. In: Proc. Conference on Innovative Data Systems Research (CIDR), January 2005, s. 162-173.

[6] Carney, D., et al: Monitoring streams - a new class of data management applications. In: Proc. VLDB, 2002, s. 215-226.

[7] Franklin, M., Halevy, A., Maier, D.: From Databases to Dataspaces: A New Abstraction for Information Management. ACM SIGMOD Record, December 2005.

[8] Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A.S., DeWitt, D., Heber, G.: Scientific, Data Management in the Coming Decade. Microsoft Research, MSR-TR-2005-10, 2005.

[9] Härder, T., Reuter, A.: Concepts for Implementing a Centralized Database Management System. In: Proc. Int. Computing Symposium on Application Systems Development, March 1983, Nürnberg, B.G. Teubner-Verlag, 1983, s. 28-104.

[10] Härder, T.: DBMS Architecture - Still an Open Problem. In: Proc. BTW, Karlsruhe, March 2005, s. 2–28.

[11] ISO: Information technology -- Database languages -- SQL -- Part 1: Framework (SQL/Framework). ISO/IEC 9075-1:1999.

[12] ISO: Information technology – Database languages – SQL – Part 2: Foundation (SQL/Foundation). ISO/IEC 9075-2:2003.

[13] Kiely, D.: How SQL Server 2005 Enables Service-Oriented Database Architectures, 2006. Dostupné na: http://www.microsoft.com/technet/prodtechnol/sql/2005/sqlsoda.mspx#EHE

[14] Ramamritham, K., Son, S.H., Dipippo, L.C.: Real-Time Databases a Data Services. Real-Time Systems, Kluwer AP, 28, 2004, s. 179-215.

[15] Selinger, P.: Five Data Challenges for Next Decade. Key note of ICDE conference, April 2005, Tokyo, Japan.

[16] Seltzer, M. I.: Beyond Relational Databases. Databases, Vol. 3, No. 3, 2005, s. 50 – 58.

[17] Stonebraker, M.: Sequoia 2000 - a reflection on first three years. Sequoia Technical Report S2K-94-58. Berkeley, CA, 1994 Dostupné na: http://epoch.cs.berkeley.edu:8000/sequoia/techreports/ s2k-93-23/.

[18] Stonebraker, M., Çetintemel, U.: "One Size Fits All" An Idea Whose Time Has Come and Gone. In: Proc. Conference ICDE, April 2005, Tokyo, Japan, s. 2-11.

[19] StreamBase Systems, Inc.: StreamBase™ 2.0, 2005. Dostupné na: http://www.streambase.com/index.html

[20] Tomasic, A. a Simon, E.: Improving Access to Environmental Data using Context Information. ACM SIGMOD Record, Volume 26, Issue 1, 1997, s. 11 – 15.

[21] Zheng, B., Lee, D.L.: Information Dissemination via Wireless Broadcast. In: Communications ACM, May 2005/Vol. 48, No. 5, s. 105-110.

Summary

Two driving forces are being mentioned in the development of today's database systems: Web and particular sciences, as physics, biology, medicine, and engineering. Data gained by sensors appears as well as monitor systems which influence a course of computerized society. Management of this data is based on traditional file techniques or it uses commercial DBMS. With a production of huge data sets and their processing in real-time applications, needs pro an effective data management have grown significantly. New ITC technologies and techniques supporting new requirements include sensor networks, stream processing, accessing uncertain and imprecise data, knowledge discovery and intelligent data analysis, and wireless broadcast a mobile computing. Both research and practice indicate that traditional universal DBMS architecture hardly satisfies these trends and new solutions are needed. Rather separate specialized engines connected into networks are beneficial. The paper discusses recent advances in database technologies and attempts to highlight them with respect to new demands.