DORP

Přednáška 1 - Úvod

- Distribuovaná databáze – kolekce propojených DB, v počítač síti
- Jedna logická DB

DDBMS

- Každý fragment má svůj DBMS
- Data rozdělené do fragmentů
- Fragmenty mohou být replikované
- Spojení po síti
- Paralelní DBMS
    - Jeden stroj, více procesorů
- Homogenní – na každém místě stejný DBMS
- Nehomog – jiný DBMS nebo i datový model

Metodologie návrhu

- Shora dolů – na zelené louce
- zdola nahoru – integrace již existujících dat
- Architektura
    - Lokální interní schéma
    - Lokální konceptuální schéma
    - Globální konceptuální schéma
    - Externí schéma

Základní problémy

- Fragmentace- Rozdělení
- Alokace – uloženo na místě s optimální distribucí
- Replikace – kopie fragmentu

Fragmentace

- Vhodné provést analýzu
    - Kvantitativní
        - Frekvence spouštění
        - Výkon
        - Místo kde app běží
    - Kvalitativní
        - Přístupy k relacím, attr, řádkům
- Proč fragmentovat
    - Data jsou v místě kde se nejčastěji používají
    - Nepotřebná data pro lokální aplikace se v místě neukládají
    - Bezpečnost
    - Paralelizmus
- Měla by být
    - Úplná – po dekomponování se každý fragment z původní relace nachází v nějakém fragmentu
    - Rekonstruovatelná – možnost složení původní relace

    - Disjunktní – prvek by se měl objevit pouze v jedno fragmentu
        - Výjimka vertikální fragmentace
- Horizontální fragmentace – řádky - selekce
- Vertikální – sloupce – natural join
- Smíšená – oba, vertikálně fragmentovaný horizontální fragment, naopak
    - Projekce a selekce
- Odvozená horizontální fr.
    - Zajišťuje že fragmenty používající se často spolu jsou uložené na jednom místě

Alokace dat

- Alternativní strategie vzhledem umístění dat
- Centralizované
- Fragmentované
- Plně replikované
- Selektivně replikované

Přednáška 2- Replikace

- Kopírování
- Celá tabulka / fragmenty – výsledky dotazu

Požadavky

- Automaticky synchronizovat kopie
- Propagovat změny
- Ochrana transakční a logické integrity dat
- Replikace z/do heterogenních serverů
- Podpora návrhu aplikací

Výhody

- Lokální aktualizované kopie
- Lokální db – minimalizace síťového provozu
- Zvýšení výkonu a dostupnosti

Problémy

- Kolik kopií udržovat ??
- Více kopií -> náročnější úprava dat ale rychlejší dotazy
- Udržení identického stavu
- Sjednocení dat po výpadku v síti

Ceny

- Změna/dotaz ze stejného místa – 1z/d
- Změna/dotaz z jiného místa – 10z/d
- Frekvence dotazů q / změn u
- Náklady propagace změn – Z=10zu1+10zu2
- Optimální 9dq + 9zu >= Z

Formy

- Dle místa kde začíná
    - Centralizované – master-slave – pouze jedna kopie je upravovaná

    - Distribuované – aktualizace může začít kdekoliv
- Dle propagace
    - Synchronní s vkládáním dat
    - Async – pokud nastane událost – metoda pull/ push
- Synchronní
    - Propagace změny v rámci jedné transakce
    - požadované vlastnosti transakcí platí na všech místech
    - nevýhoda – řádné ukončení transakce až po provedení všech změn
- Asynchronní
    - Nejprve aktualizace na jednom místě (master, primární kopie)
    - Po ukončení transakce se propaguje na další kopie
    - Transakce nečeká na ukončení propagace
    - Možná nekonzistence ale je to rychlejší

Využití

- Sjednocení dat s centrálním serverem
- Rozdělení procesu na více než jeden server
- Sdílení dat mezi více místy

Přednáška 3 – Replikace MSSQL

- Pojmy
    - Článek – základní jednotka replikace
    - Publikace – kolekce článků – různá nebo stejná pro různé odběratele
    - Přihlášení k odběru – požadavek na kopii publikace
- typy
    - Snímková replikace – ne časté změny, přijatelná nesych po nějakou dobu
    - Slučovaná replikace
        - zejména pro mobilní offline zpracování
        - upravuje vydavatel i předplatitel
        - modifikuje schéma – trigery, PKs, sys tabulky
    - Transakční replikace
        - Průběžná replikace
        - Malá prodleva
        - Předplatitel většinou read only
- Agenti
    - Distribution agent – snímková a transakční
    - Snapshot agent – všechny typy
    - Merge agent - slučovací
    - Log reader – transakční
- Typy DB
    - Odběratel - subscriber
    - Vydavatel - publisher
    - Distributor
- Správa
    - Management studio
    - Replication Programming interface

Přednáška 3 – Oracle DDBMS

- DB linky
- Distribuované dotazy, DML
- Synonyma pro table, type, view, sequence, proc, pack

Replikace

- table, index, view, pack, proc,…
- Multimaster
    - Rovnocenné uzly
    - Transakce, sync/async
    - Master group – skupina- replikují najednou
- Materializované pohledy – snapshots
    - Manuální replikace
    - Časová replikace

Přednáška 4- Integrace dat

- Proces přistupování dat z několika/různých zdrojů jako v jedné DB

Problémy

- Existující struktura jež nelze měnit
- Nekompatibilita mezi DBs
    - Lexikální – jiné pojmenování sloupců
    - Interpretace dat – v jiných jednotkách atd
    - Sémantická

Přístupy

- Federativní
    - Zdroje jsou nezávislé – transformace dotazů a odpovědí
    - Je třeba implementovat spojení mezi každými dvěma zdroji
- Datové sklady
    - kopie dat, transformovaná do unifikovaného formátu, periodicky
    - globální schéma
    - buďto periodické updaty
    - nebo znovu vytvoření ze zdrojových db
- Mediátor – jako datový sklad ale neukládá data

Adaptéry / wrapery

- Vychází z klasifikace očekávaných dotazů – vytváří šablony dotazů
- Metody zjednodušení – kombinace šablon

Uživatelsky definované funkce UDF

- Pohledy neumožní pracovat s parametry, funkce ano
- Lze na ně odkazovat ve from
- Přímé tabulkové UDF – jeden select
- Vícepříkazuvé UDF- více selectů

Přednáška 5 – Distribuované dotazy (Zpracování)

- Cena přístupu k řádku = 1
- Cena přenosu řádku = 10
- Transformace SQL dotazu do posloupnosti DB operací nad fragmenty
- Optimalizátor – nalezne nejlepší místo pro zpracování dotazu a pořadí přenesení

Dekompozice dotazu

- Vytvoření listu pro každou relaci
- Kořen – výsledek

1. Normalizace
    a. where do konjuktní/ disjunktní normální formy
    b. dle and a or
2. Zjednodušení – redundantní podmínky
    a. Využití pravidel boolovské algebry
3. Vyjádří se jako algebraický dotaz
    a. Operátory redukující počet by měli být provedeny co nejdříve
    b. A zároveň od nejjednoduššího k nejtěžšímu
    c. Vyhnout se crossjoinu
    d. Semijoin redukce
    e. Pravidla
        i. Kumulativa
        ii. Asociativa
        iii. Konjunktivní selekce lze transformovat na kaskádu selekcí

Lokalizace dat

- Bere v úvahu distribuci dat
- Nahrazení globální relace na listech stromu jejich vyjádřením pomocí fragmentů
- Redukce
    - Pro primární horiz fragmentaci
    - Vertikální frag
    - Odvozenou frag
    - Hybridní frag

Přednáška 7 – Objektově relační DB

Klady relačních DB

- Dotazovací schopnosti
- Rozsáhlé soubory s jednoduchou strukturou – tabulky
- Většina HW platforem
- Brány mezi jednotlivými DB systémy

Nedostatky relačních DB

- Nevhodné pro bohatost datových typů

OO datový model

- Dlouho názory že vytlačí relační Db
- Bohatost typů objektů
- Větší složitost z hlediska struktury a vztahů

Požadavky na OO DBS

- Podpora perzistence, paralelizmus, zotavení, kladení dotazů
- Podpora
    - Indentifikace objektů
    - Zapouzření
    - Dědičnost
    - Typy, třídy
    - Polymorfizmus

Objektově relační DBS

- Kompromis
- Možnost pozvolné migrace
- Doplnění relačního modelu o práci s dat strukturami z programovacích jazyků
- Možnost ukládání objektů do relační databáze
- Bin objekty (audio, video), prostorová data
- Rozšiřitelnost
    - Možnost přidání dat. Typů pro efektivní vyhledávání

Rozšíření relačního modelu

- Strukturované typy atributů – hodnotou může být celá relace
- Reference – sdílení řádku mezi tabulkami
- Metody
- Identifikátory řádků
- Array, muiltiset, list, set
- Typ ROW
- UDT
    - Typ tabulky
    - Nebo typ atributu v tabulce

Přednáška 9 – Návrh datového skladu

- Pro podporu rozhodování
- Data mining
- OLAP
- Reporting, querying
- ETL

Zpracování dat v datovém skladu

- Warehouse manager
    - Zajištění konzistence dat
    - Vytvoření indexů, agregace
    - Zálohy dat
- Query manager – zpracování dotazů
- Metadata management
    - Info o struktuře
    - Info o původních strukturách a jejich transformace
    - O agregacích

Metodologie návrhu

- Corporate Information Factory - Inmon
    - Začíná vytvořením dat modelu pro celý podnik a pak vytvoření datamartů
    - Potencionální konzistentní a úplný pohled na data
- Business Dimensional lifecycle
    - Začíná identifikací analytických témat a příslušných business procesů
    - Tržiště se integrují do skladu
    - Použití nové techniky – dimenzionální modelování
    - Rozdělení na etapy
    - První tržiště v plánovaném čase

Dimenzionální modelování

- Hvězdicové schéma
- Střed má složený klíč – tabulka faktů
- Okolní tabulky tvoří dimenze
- Kroky
    - Vybrat business process
    - Vybrat granularitu v tabulce faktů
    - Vybrat dimenze
    - Identifikuje fakta
    - Identifikuje atributy dimenzí
    -