Tietokantakyselyjen optimointi relaatiotietokannassa

Olli Rissanen

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 24. helmikuuta 2013

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Olli Rissanen

Työn nimi — Arbetets titel — Title

Tietokantakyselyjen optimointi relaatiotietokannassa

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Kandidaatintutkielma	24. helmikuuta 2013	4

Tiivistelmä — Referat — Abstract

Tutkielmassa tutustutaan tietokantakyselyjen optimointiin relaatiotietokantojen hallintajärjestelmien osalta sekä optimoinnin vaikutukseen kyselyjen suorituskyvyssä. Tärkeimmät suorituskykymittarit ovat prosessorin ja muistin käyttö.

Avainsanat — Nyckelord — Keywords
Information systems Query optimization

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto

Tietokantojen suorituskyky on yhä tärkeämpää tiedon määrän kasvaessa. Optimoimalla tietokantakyselyjen suoritusta voidaan helpottaa käyttäjien tiedonhakua sekä kasvattaa tietokannan suorituskykyä. Kyselyn optimointi on toteutettu automaattisena toimenpiteenä tietokannan hallintojärjestelmien sisällä, ja se on ydintekijä erityisesti relaatiomalliin pohjautuvien hallintajärjestelmien menestyksessä.

Tietokannan hallintajärjestelmä on kokoelma ohjelmia tiedon tallentamiseen, muokkaamiseen, analysointiin ja keräämiseen tietokannasta. Hallintajärjestelmää käytetään tietokantakyselyillä, ja tutkielman oletuskyselykielenä on SQL. Hallintajärjestelmän sisältämän kyselyoptimoijan tehtävänä on löytää kyselylle suorituskykyisin kyselysuunnitelma mahdollisimman nopeasti. Optimoinnilla voidaan saavuttaa merkittäviä säästöjä prosessorin ja muistin käytössä.

Kyselyoptimoijan tavoitteena on minimoida itse optimointiin käytetty aika ja maksimoida optimoinnista saatu hyöty. Kyselyoptimoija toimii etsien kyselyä vastaavat mahdolliset kyselysuunnitelmat ja valitsemalla niistä tehokkaimman. Kyselysuunnitelma sisältää sarjan algebrallisia operaatioita tietokannan relaatioille jotka tuottavat tulokseksi halutun vastauksen. Tietokantakyselyä vastaavia kyselysuunnitelmia voi olla useita, sillä kyselyiden algebralliset esitykset voidaan usein esittää monena loogisesti vastaavana esityksenä. Algebrallista operaatiota kohden voi myös löytyä useita toteutuksia, kuten join-operaatiota toteuttavat merge join ja hash join. Toisiaan vastaavat esitykset voivat olla suorituskyvyltään jopa eri asteikolla. Haasteeksi nousee kyselysuunnitelman luominen ja kyselysuunnitelmien suorituskyvyn ennustaminen. Optimointi on vaikea hakuongelma, jossa hakualue voi nousta erittäin suureksi kyselyn ollessa monimutkainen. Optimoijan tulee valita pienin mahdollinen hakualue, joka pitää sisällään halvimmat suunnitelmat. Suorituskyvyn ennustamisen ja hakualueen rajauksen lisäksi optimoija tarvitsee tehokkaan algoritmin koko hakualueen läpikäymiseen.

Tutkielman rakenteesta

2 workname: Taustaluku

Tietokannan hallintajärjestelmien jako: query optimizer ja query execution engine.

Tietokantakyselyiden optimoinnilla viitataan tietokantakyselyn suorittamiseen mahdollisimman tehokkaasti. Optimoinnin tavoitteena on joko maksimoida suorituskyky annetuilla resursseilla tai minimodia resurssien käyttö. Mitattavia resursseja ovat suorittimen ja muistin käyttö sekä kommunikointikustannukset. Muistin käyttö jakautuu tallennuskustannukseen sekä ulkomuistiin pääsyn kustannukseen. Tallennuskustannuksella tarkoitetaan

ulkomuistin sekä puskurimuistin käyttöä, ja se tulee aiheelliseksi kun muistin käyttö aiheutuu pullonkaulaksi.

Resurssin merkitys riippuu tietokantatyypistä. Hajautetuissa tietokannoissa hitailla yhteysväylillä kommunikointikustannukset hallitsevat kustannuksia. Paikallisesti hajautetuissa tietokannoissa kaikilla resursseilla on sama painoarvo. Keskitetyissä tietokannoissa ulkomuistiin pääsyn kustannus ja prosessorin käyttö ovat oleellisia. Tämän tutkielman aihepiiriin kuuluu vain keskitettyjen tietokantojen optimointi.

todo: liitoskohta

Relaatiotietokanta on relaatiomalliin perustuva tietokanta. Relaatiomallin keskeinen piirre on kaiken datan esittäminen n-paikkaisen karteesisen tulon osajoukkona, ja se tarjoaa deklaratiivisen menetelmän datan ja kyselyjen määrittämiseen. Relaatiomalli koostuu attribuuteista, monikoista ja relaatioista. Matemaattisessa määritelmässä attribuutti on pari joka sisältää attribuutin nimen ja tyypin sekä jokaiseen attribuuttiin liittyy sen arvojoukko. Monikko on järjestetty joukko attribuuttien arvoja. Relaatio koostuu otsakkeesta ja sisällöstä(body?), jossa otsake on joukko attribuutteja ja keho on joukko monikkoja. Relaation otsake on myös jokaisen monikon otsake. Visuaalisessa esityksissä relaatio on taulukko ja monikko taulukon rivi.

todo: SQL ja relaatiomalli

todo: relaatiotietokanta vs no-sql

todo: optimoi menee

[1]

3 menetelmä 1

4 menetelmä 2

5 menetelmä n

6 menetelmien vertailu

7 case study?

8 yhteenveto

Lähteet

[1] Chaudhuri, Surajit: *An overview of query optimization in relational systems.* Teoksessa *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, sivut 34–43. ACM, 1998.