**Database Management System**

**Course Code : CSE 3102**

Assignment on

**Query Processing and Query Optimization**

**Submitted to :**

**Md Hasan Hafirzur Rahman**

**Assistant Professor , Dept. of CSE**

**Comilla University**

**Submitted by :**

**Name : Md. Moin Uddin**

**Id : 11808011**

**Session : 2017-20118**

**Submission Date : 06 September 2021**

**Query Processing (প্রশ্ন প্রক্রিয়াকরণ)**

***কুয়েরি প্রসেসিং হচ্ছে ডাটাবেস থেকে ডেটা বের করার কাজ। ক্যোয়ারী প্রক্রিয়াকরণে, এটি ডাটাবেস থেকে ডেটা আনার জন্য বিভিন্ন পদক্ষেপ নেয়। জড়িত পদক্ষেপগুলি হল :***

***-পার্সিং এবং অনুবাদ ( Parsing and Translation )***

***-অপ্টিমাইজেশন ( Optimization )***

***-মূল্যায়ন ( Evaluation )***

প্রশ্ন প্রক্রিয়াকরণ বলতে বোঝায়, SQL অথবা উচ্চতর লেভেলের প্রশ্নকে নিম্নতর এক্সপ্রেশন যেমন রিলেশনাল এলজেবরা এক্সপ্রেশনে প্রকাশ করা যাতে ফলাফল পাওয়ার জন্য ঐ এক্সপ্রেশন ফিজিক্যাল লেভেলের উপর ব্যবহার হতে পারে,প্রশ্নকে অপটিমাইজ করার জন্য ব্যবহার হতে পারে এবং প্রকৃত বাস্তবায়নে ব্যবহার হতে পারে।এই সমস্ত প্রক্রিয়াকেই প্রশ্ন প্রকিয়াকরণ বলা হয়।

**কোয়েরি অপটিমাইজেশন:**

মূলত, একই ফলাফল পাওয়ার জন্য n সংখ্যক কোয়েরি থাকে। কোয়েরি অপটিমাইজেশনের ক্ষেত্রে, যতগুলো কোয়েরি আছে প্রথমে সব কোয়েরির এক্সিকিউশন প্ল্যান খুঁজে বের করা হয় এবং সবগুলো প্ল্যান এর মধ্যে সবচেয়ে কার্জকর প্ল্যানকে এক্সিকিউশনের জন্য নির্ধারন করা হয়ে থাকে, এই সমস্ত প্রক্রিয়াকে বলা হয় কোয়েরি অপটিমাইজেশন।

**Query processing steps (প্রশ্ন প্রক্রিয়াকরণের ধাপসমূহ):**

প্রশ্ন প্রক্রিয়াকরণের ধাপসমূহ হচ্ছে :-

* বিশ্লেষন এবং অনুবাদ (Parsing and Translation)
* অপটিমাইজেশন
* মূল্যায়ন (Evaluation)

বিশ্লেষণ এবং অনুবাদ

রিলেশনাল এলজেবরা

প্রশ্ন

চিত্র: প্রশ্ন প্রক্রিয়াকরণের ধাপ সমূহ

ডাটার পরিসংখ্যান

সম্পাদন পরকিল্পনা

ফলাফল

অপটিমাইজার

ডাটা

ডাটা

মূল্যায়ন

* একটি প্রশ্নের (query) উত্তর গণনা করার জন্য বিভিন্ন পদ্ধতি রয়েছে। যেমন ‍SQL এ দেখা যায় যে একটি query বিভিন্নভাবে প্রকাশ করা যায়।

অন্যদিকে, একটি query এর রিলেশনাল বীজগাণিতিক উপস্থাপনা কিভাবে একটি query কে মুল্যয়ন করবে তা শুধুমাত্র আংশিকভাবে উল্লেখ করে। বিভিন্ন উপায় আছে রিলেশনাল বীজগাণিতিক এক্সপ্রেশন মুল্যায়ন করার। উদাহরণ হিসেবে বলা যেতে পারে যে,

select balance

from account

where balance < 2500

এই query টিকে দুই ধরণের রিলেশনাল বীজগাণিতিক উপস্থাপনায় অনুবাদ করা যেতে পারে,

* σ balance <2500 (π balance (account))
* π balance (σ balance < 2500(account))

একটি রিলেশনাল বীজগাণিতিক অপারেশন কিভাবে মুল্যায়ন করতে হবে তার নির্দেশনাবলীকে বলা হয় আদিম মুল্যায়ন (primitive evaluation)। একটি query মুল্যায়নের জন্য ব্যবহার করা যেতে পারে এমন আদিম ক্রিয়াকলাপের একটি ক্রম হলো query মুল্যায়ন পরিকল্পনা (Query evaluation plan)।

π balance

σ balance < 2500: use index

account

Fig: A Query Evaluation Plan

Query এক্সিকিউশন ইঞ্জিন সাধারণত query পরিচ্লনার পরিকল্পনা করে, সেই পরিকল্পনার বাস্তবায়ন করে এবং প্রশ্নের উত্তর প্রদান করে।

**Measure of Query Cost (প্রশ্নের খরচের পরিমাপ):**

আমরা ইতোপূর্বে দেখেছি যে, হাই লেভেল কোয়োরি কে চাহিদা অনুযায়ী রিলেশনাল এলজেবরা এক্সপ্রেশনে অনুবাদের মাধ্যমে কিংবা অন্যান্য প্রক্রিয়ায় লো লেভেল কোয়েরিতে কনভার্ট করতে হয়।কোনো কোয়েরিকে চাহিদাসম্পন্ন কোয়েরিতে কনভার্ট করার জন্য কিছু পরিমাপের প্রয়োজন হয়।

কোয়েরি খরচের কিছু মৌলিক পরিমাপ হচ্ছে:-

* ডিস্ক এক্সেস
* CPU cycles
* Transit time in network

মুলত CPU cost গণনা একটু জটিল।তাই কোয়েরি খরচ গণনার ক্ষেত্রে সবচেয়ে প্রজনীয় জিনিস হচ্ছে ডিস্ক এক্সেস।CPU cost বলতে বোঝায় যে প্রশ্ন মুল্যায়নে কতগুলো CPU cycle খরচ হয়। CPU স্পিড ডিস্ক স্পিড এর তুলনায় দ্রুততর যার মানে হচ্ছে CPU access কোয়েরি কে ডিস্ক এর তুলনায় দ্রুত পরিচালনা করতে সক্ষম আর তাই CPU cost ডিস্ক cost এর তুলনায় কম।নেটওয়ার্কের ট্রানজিট সময়কে প্রাথমিকভাবে প্যারালাল সিস্টেম এর সাথে তুলনা করা হয়।নেটওয়ার্কের ট্রানজিট সময় তাই কোয়েরি খরচ পরিমাপের ক্ষেত্রে খুব বেশি প্রয়োজন হয় না।

* **Disk Access cost**: ডিস্ক এর খরচ গণনার ক্ষেত্রে প্রয়োজনীয় পরিমাপগুলো হচ্ছে:-
* খোঁজার সংখ্যা(Number of seeks)
* ব্লক রিড এর সংখ্যা (Number of blocks read)
* ব্লক রাইটের সংখ্যা (Number of blocks write)

ডিস্ক এক্সেস সবচেয়ে উপযুক্ত query টি খুঁজে পাওয়ার জন্য গৃহীত পদক্ষেপ সংখ্যাই হচ্ছে খোঁজার সংখ্যা । এটি রেনডম I/O ‌এবং সিকুয়েনশিয়াল I/O এর ক্ষেত্রে ভিন্ন হয়।

সাধারণত,

(রাইটিং) এর খরচ > (রিডিং) এর খরচ

* খোঁজার সংখ্যা (N), খরচ = N\*খোঁজার গড় সময়
* ব্লক রিড এর সংখ্যা, খরচ = N\* ব্লক রিডের গড় খরচ
* ব্লক রাইটের সংখ্যা , খরচ = N \* ব্লক রাইটের গড় খরচ
* খরচের পরিমাপ হিসেবে ডিস্ক থেকে ব্লক ট্রান্সফার এর সংখ্যা এবং খোঁজার সংখ্যা,

tT = ১টি ব্লক ট্রান্সফারের জন্য প্রয়োজনীয় সময়

tS = ১ বার খোঁজার জন্য প্রয়োজনীয় সময়

b blocks এবং s seeks এর মোট খরচ =(b\*tT+ s\*TS)

এটিই কোয়েরি খরচের পরিমাপের সুত্র।

**Selection operation (নির্বাচন অপারেটর):**

প্রশ্ন প্রক্রিয়াকরণে ফাইল স্ক্যান হচ্ছে ডাটা এক্সেস করার জন্য সর্বনিম্ন স্তরের অপারেটর। ফাইল স্ক্যান হচ্ছে সার্চ এলগরিদম যা একটি রেকর্ড শনাক্ত করে এবং পুনরুদ্ধার করে যা নির্বাচন শর্ত পূরণ করে।

রিলেশনাল সিস্টেমে , একটি ফাইল স্ক্যান সেই ক্ষেত্রে একটি সম্পুর্ণ সম্পর্ক পড়তে দেয় যেখানে সম্পর্কটি একটি একক ডেডিকেটেড ফাইলে সংরক্ষন করা হয়।

নির্বাচন অপারেশন বাস্তবায়নের জন্য দুটি স্ক্যান এলগরিদম হলো-

1. লিনিয়ার সার্চ
2. বাইনারি সার্চ

**লিনিয়ার সার্চ:**

লিনিয়ার সার্চ এর ক্ষেত্রে সিস্টেম প্রতিটি ফাইলব্লক স্ক্যান করে এবং সমস্ত রেকর্ড পরীক্ষা করে এটা দেখার জন্য যে তারা নির্বাচন শর্ত পূরণ করে কিনা।ইনপুট আউটপুটের সংখ্যার উপর ভিত্তি করে লিনিয়ার সার্চ এর খরচ হচ্ছে br.যেখানে br ফাইলের ব্লক সংখ্যা নির্দেশ করে।

**বাইনারি সার্চ:**

যদি ফাইলটি একাটি এট্রিবিউটে অর্ডার করা থাকে এবং সিলেকশন কন্ডিশনটি এট্রিবিউটের তুলনায় সমান হয়, তখন আমরা বাইনারি সার্চ ব্যবহার করতে পারি রেকর্ড গুলোকে শনাক্ত করার জন্য যা সিলেকশনকে সন্তুষ্ট করে।সিস্টেমে ফাইল ব্লকে বাইনারি সার্চ করে । একটি ব্লক খোঁজার জন্য যে সংখ্যক ব্লক পরীক্ষা করতে হবে তার রেকর্ড হচ্ছে = log2(br)

যেখানে br ফাইলের ব্লক সংখ্যা নির্দেশ করে।

* নির্বাচন অপারেশনে মুলত রিলেশন থেকে একটি নির্দিষ্ট টাপল অথবা সারি নির্বাচন করা হয় যা নির্দিষ্ট শর্ত পূরণ করে।সিলেক্ট অপারেটরকে সিগমা দ্বারা প্রকাশ করা হয়। যেমন, σθ(r)

এটি দ্বারা বোঝায় যে, যেকোনো রিলেশন r থেকে কোনো কিছু সিলেক্ট করা যা θ কন্ডিশনকে স্যাটিসফাই করে এখানে, .

r = রিলেশন, θ = কন্ডিশন, এবং σ = সিলেক্ট অপারেটর নির্দেশ করে।

নির্বাচন অপারেশন commutative এবং নির্বাচন অপারেটর এর শর্ত composite হতে পারে।

**কোয়েরি প্রসেসিং এ সর্টিং (Sorting in Query Processing):**

সর্টিং এর প্রয়োজন হয়:-

* যখন কোনো ডাটাবেজকে সর্টেড অর্ডারে রাখার প্রয়োজন হয় এবং
* কোয়েরি প্রসেসিংকে আরো কর্মক্ষম করে তোলার জন্য ।
* সর্টিং রিলেশন যা মেমোরিতে জায়গা হয় না তাকে এক্সটার্নাল সর্টিং বলে।
* ধরা যাক M মেইন মেমোরি বাফার এর ব্লক সংখ্য নির্দেশ করে যা সর্টিং এর জন্য সহজলভ্য।

**প্রথম ধাপ:**

* কিছু সংখ্যক সর্টেট রান সৃষ্টি হয়।
* প্রত্যেক রান সর্টেট থাকে কিন্তু ম্ত্র কিছু সংখ্যক রান রিলেশনের রেকর্ড ধারণ করে।

i=0

repeat

read M blocks of the relation, or the rest of the relation, whichever is smaller sort the in-memory part of the relation

write the sorted data to run file Ri

i= i+1

until the end of the relation.

**দ্বিতীয় ধাপ:**

* রানসমুহ মার্জ আকারে থাকে।
* মোট রান সংখ্যাN, M থেকে ছোট হয়ে থাকে যাতে আমরা প্রতিটি রানের জন্য একটি ব্লক বরাদ্দ রাখতে পারি।

read

one block of each of the N files Ri into a buffer block in the memory:

repeat

choose the first tuple among all buffer blocks;

মার্জ সর্টের আউটপুটই হচ্ছে সর্টেট রিলেশন।ডিস্ক রাইট অপারেশনের সংখ্যা কমাতে আউটপুট ফাইলটি বাফার করা হয়।যেহেতু এটি N সংখ্যক মার্জকে রান করে তাই একে N-way মার্জ বলা হয়।