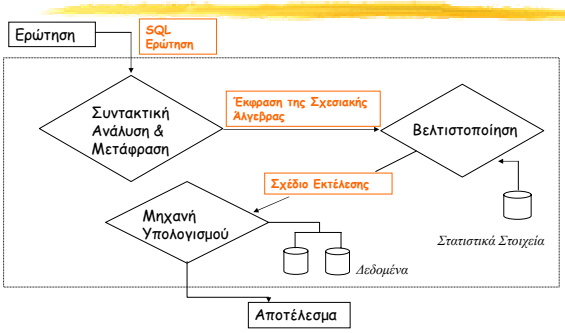
Επεξεργασία Ερωτήσεων: Επανάληψη και Ασκήσεις



Τα βασικά βήματα στην επεξεργασία μιας ερώτησης είναι

1. Συντακτική Ανάλυση & Μετάφραση
2. Βελτιστοποίηση
3. Υπολογισμός

1. Συντακτική Ανάλυση (Parsing) & Μετάφραση

Η SQL ερώτηση μεταφράζεται σε μια εσωτερική μορφή αφού γίνει ο απαραίτητος συντακτικός και σημασιολογικός έλεγχος (π.χ., τα ονόματα που αναφέρονται είναι ονόματα σχέσεων που υπάρχουν)

Αντικατάσταση των όψεων από τον ορισμό τους

Σε ποια εσωτερική μορφή; Έκφραση της σχεσιακής άλγεβρας

select $A_1, A_2, \ldots, A_n$
from $R_1, R_2, \ldots, R_m$
where P

$\pi_{A1, A2, \ldots, An} (\sigma_P (R_1 \times R_2 \times \ldots \times R_m))$

2. Βελτιστοποίηση

Μια SQL ερώτηση μπορεί να μεταφραστεί σε διαφορετικές (ισοδύναμες) εκφράσεις της σχεσιακής άλγεβρας

select balance
from account
where balance < 25000

- $\sigma_{balance < 2500} (\pi_{balance}(account))$
- $\pi_{balance} (\sigma_{balance < 2500} (account))$

Δέντρο ερώτησης ή τελεστών

Φύλλα: σχέσεις εισόδου

Εσωτερικοί κόμβοι: πράξεις της σχεσιακής άλγεβρας

Εκτέλεση δέντρου ερώτησης

Ευριστικοί Κανόνες

1. Διάσπαση των πράξεων επιλογής με συζευκτικές συνθήκες σε ακολουθίες πράξεων επιλογής

2. Μετατοπίζουμε την πράξη επιλογής όσο πιο κάτω επιτρέπεται από τα γνωρίσματα που περιλαμβάνονται στη συνθήκη

3. Επαναδιευθέτηση των φύλλων ώστε να εκτελούνται πρώτα οι σχέσεις που έχουν τις πιο περιοριστικές πράξεις επιλογής

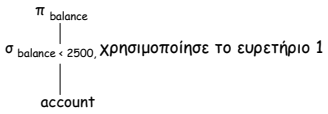
4. Συνδυασμός μιας πράξης καρτεσιανού γινομένου με μια πράξη επιλογής που ακολουθεί

5. Διάσπαση και μετακίνηση των λιστών προβολής όσο πιο κάτω γίνεται στο δέντρο

6. Εντοπισμός υποδέντρων με ομάδες πράξεων που μπορεί να εκτελεστούν με κοινό αλγόριθμο

βασικές (primitive) πράξεις: πράξη + αλγόριθμος

Σχέδιο εκτέλεσης (execution plan): μια ακολουθία από βασικές πράξεις



- Τα διαφορετικά σχέδια εκτέλεσης έχουν και διαφορικό κόστος
- Βελτιστοποίηση: η διαδικασία επιλογής του σχεδίου εκτέλεσης που έχει το μικρότερο κόστος
- Εκτίμηση του κόστους (συνήθως χρήση στατιστικών στοιχείων)

3. Εκτέλεση

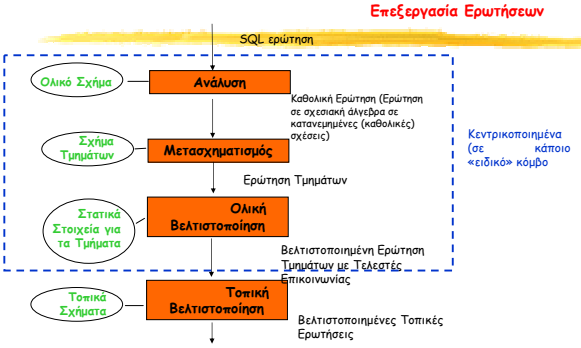
Μηχανή εκτέλεσης που εκτελεί τις βασικές πράξεις

Τι θα συζητήσουμε:

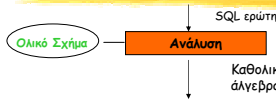
1. Μετασχηματισμός ερωτήσεων που απευθύνονται σε καθολικές σχέσεις σε ερωτήσεις που απευθύνονται σε συγκεκριμένα τμήματα της κατανεμημένης βάσης.

2. Μεθόδους που βελτιστοποιούν την αποτίμηση μιας ερώτησης.

Επεξεργασία Ερωτήσεων

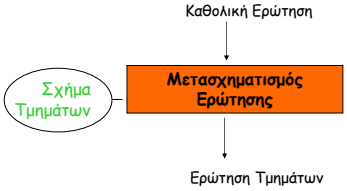


1ο Βήμα: Ανάλυση Ερωτήσεων



1. Λεξική και συντακτική ανάλυση
2. Παραγωγή του δέντρου ερώτησης (Σημείωση: τα φύλλα αντιστοιχούν σε καθολικές σχέσεις)
3. Πιθανές βελτιστοποιήσεις

2ο Βήμα: Μετασχηματισμός Ερώτησης



1. Καθορισμός των τμημάτων που συμμετέχουν στην ερώτηση
2. Αντικατάσταση της σχέσης από τα τμήματά της
3. Βελτιστοποίηση

Παράδειγμα (οριζόντια τμηματοποίηση)

EMP

EMP1 = $\sigma_{ENO \leq E3}$ (EMP)

EMP2 = $\sigma_{E3 < ENO \leq E6}$ (EMP)

EMP3 = $\sigma_{ENO > E6}$ (EMP)

select ENAME

from EMP, ASG, PROJ

where EMP.ENO= ASG.ENO and ASG.PNO = ROJ.PNO and ENAME <> "J. DOE" and PNAME = "CAD/CAM" and (DUR = 12 OR DUR = 24)

ASG

ASG1 = $\sigma_{ENO \leq E3}$ (ASG)

ASG2 = $\sigma_{ENO > E3}$ (ASG)

- Αντικατάσταση του EMP με το (EMP1 ∪ EMP2 ∪ EMP3)
- Αντικατάσταση του ASG με το (ASG1 ∪ ASG2)

Νέο δέντρο ερώτησης

Παράδειγμα

EMP

EMP1 = $\sigma_{ENO \leq E3}$ (EMP)

EMP2 = $\sigma_{E3 < ENO \leq E6}$ (EMP)

EMP3 = $\sigma_{ENO > E6}$ (EMP)

ASG

ASG1 = $\sigma_{ENO \leq E3}$ (ASG)

ASG2 = $\sigma_{ENO > E3}$ (ASG)

Παραλληλισμός

Χρήση της προσεταιριστικότητας μεταξύ ένωσης και συνένωσης

(R1 ∪ R2) J (S1 ∪ S2) =
(R1 J S1) ∪ (R1 J S2) ∪ (R2 J S1) ∪ (R2 J S2)

Νέο δέντρο ερώτησης

Απλοποιήσεις;

Ελαχιστοποιήσεις για Οριζόντια Τμηματοποίηση

Γενικά η ελαχιστοποίηση για οριζόντια τμήματα αφορά (μετά την αναδόμηση των δέντρων ερώτησης) τον καθορισμό των υποδέντρων που παράγουν άδειες σχέσεις και την απαλοιφή τους

1. Ελαχιστοποίηση με επιλογή
2. Ελαχιστοποίηση με συνένωση

Ελαχιστοποίηση με Επιλογή

Αν η συνθήκη της επιλογής αντιβαίνει τη συνθήκη επιλογής της κατάτμησης

Έστω τμήμα R_j: $R_j = \sigma_{Pj}(R)$

$\sigma_{Pi}(R_j) = \emptyset, \forall x \in R: \neg(Pi(x) \wedge Pj(x))$

EMP

EMP1 = $\sigma_{ENO \leq E3}$ (EMP)

EMP2 = $\sigma_{E3 < ENO \leq E6}$ (EMP)

EMP3 = $\sigma_{ENO > E6}$ (EMP)

Παράδειγμα

select *

from EMP

where ENO = "E5"

Ελαχιστοποίηση με Συνένωση

Δυνατή αν η συνθήκη για την κατάτμηση είναι με βάση το γνώρισμα της συνένωσης. Σε αυτήν την περίπτωση, καταμερισμός των συνενώσεων και της ένωσης και απαλοιφή των άχρηστων συνενώσεων

EMP

EMP1 = $\pi_{ENO, ENAME}$ (EMP)

EMP2 = $\pi_{ENO, TITLE}$ (EMP)

select ENAME

from EMP

Αντικατάσταση του EMP με

EMP1 J EMP2

Απλοποίηση

Γενικά η ελαχιστοποίηση για κάθετα τμήματα αφορά (μετά την αναδόμηση των δέντρων ερώτησης) τον καθορισμό των υποδέντρων που παράγουν άχρηστες ενδιάμεσες σχέσεις και την απαλοιφή τους

1. Ελαχιστοποίηση με προβολή

Ελαχιστοποίηση με Προβολή

Άχρηστη να τα γνωρίσματα της προβολής δεν ανήκουν στο τμήμα!

Οι κανόνες για την κεντρικοποιημένη περίπτωση συν:

1. Απαλοιφή άδειων σχέσεων που προκύπτουν από μη συμβατές επιλογές σε οριζόντια τμήματα
2. Απαλοιφή άχρηστων σχέσεων που προκύπτουν από προβολές σε κάθετα τμήματα
3. Κατανομή των συνενώσεων πάνω από τις ενώσεις και απαλοιφή των άχρηστων συνενώσεων

EMP

EMP1 = $\sigma_{ENO \leq 4}$ ($\pi_{ENO, ENAME}$ (EMP))

EMP2 = $\sigma_{ENO > 4}$ ($\pi_{ENO, ENAME}$(EMP))

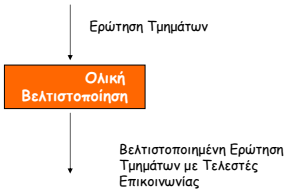
EMP3 = $\pi_{ENO, TITLE}$ (EMP)

select ENAME

from EMP

where ENO = "E5"

Ολική Βελτιστοποίηση



Ολική Βελτιστοποίηση Ερωτήσεων

Εύρεση ενός καλού (όχι απαραίτητα βέλτιστου) ολικού πλάνου

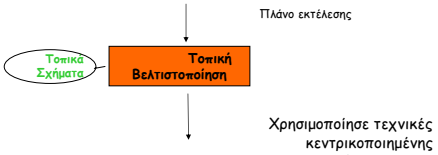
1. Ελαχιστοποίηση κάποιας συνάρτησης κόστους
2. Κατανεμημένη επεξεργασία συνένωσης
   - Bushy/linear δέντρα
   - Ποιες σχέσεις να μεταφέρουμε που
   - Μεταφορά όλης της σχέσης ή όταν χρειάζεται
3. Χρήση ή όχι της ημι-συνένωσης
4. Μέθοδοι υπολογισμού συνένωσης

Ελαχιστοποίηση Κόστους

- **Ολικό Κόστος**
- **Χρόνος απόκρισης** για μια ερώτηση: χρόνος που περνά από την υποβολή της ερώτησης μέχρι την ολοκλήρωση της

Τοπική Βελτιστοποίηση

Σε κάθε κόμβο



Χρησιμοποίησε τεχνικές κεντρικοποιημένης βελτιστοποίησης για την εύρεση του καλύτερου (τοπικά) τρόπου εκτέλεσης

Επεξεργασία Ερωτήσεων

Περίληψη Ευριστικών Κανόνων

Κανόνας 1

Βασιζόμενοι στην επιμεριστικότητα των μοναδιαίων τελεστών σε σχέση με τους δυαδικούς, σπρώχνουμε επιλογές και προβολές όσο πιο χαμηλά γίνεται στο δέντρο (με αποτέλεσμα να μειώνουμε γρήγορα και τοπικά το μέγεθος της εμπλεκόμενης πληροφορίας).

Για παράδειγμα, αντί για να κάνουμε την επιλογή $\sigma_{CITY='RHO'}$ (TASK J DEPT_TASK), προτιμήθηκε να σπρώξουμε την επιλογή μόνο στα τμήματα TASK1 και TASK2, ώστε να μειώσουμε το κόστος μεταφοράς δεδομένων.

Επεξεργασία Ερωτήσεων

Κανόνας 2

Βασιζόμενοι στην αντιμεταθετικότητα και τη μοναδική ποσότητα των μοναδιαίων τελεστών, παράγουμε συνδυασμούς από επιλογές και προβολές σε κάθε εμπλεκόμενη σχέση.

Στο προηγούμενο παράδειγμα, αντί να εμπλέξουμε όλο το TASK1, προτιμήθηκε να χρησιμοποιηθεί η έκφραση π_{TNUM} ($\sigma_{CITY='RHO'}$(TASK1)), η οποία σαφώς μειώνει το μέγεθος της αποτιμούμενης σχέσης.

Κανόνας 3

Μπορούμε να απαλείφουμε φύλλα του δέντρου (τμήματα δηλαδή των καθολικών σχέσεων), αν ο συνδυασμός των αλγεβρικών εκφράσεων που τα ορίζουν με τις αλγεβρικές εκφράσεις που τους επιβάλλονται έρχεται σε σύγκρουση.

Για παράδειγμα, η έκφραση $\sigma_{CITY='RHO'}$ (TASK2) ισοδυναμεί με την έκφραση $\sigma_{CITY='RHO'}\sigma_{CITY='KAST'}$ (TASK)
η οποία προφανώς εμπεριέχει αντίφαση και δεν έχει νόημα να αποτιμηθεί.

(Αντίστοιχα, σε μια περίπτωση κάθετης κατάτμησης, θα μπορούσε να συμβαίνει το ίδιο με μια προβολή).

Κανόνας 4

Μπορούμε να εκτελούμε τις συνενώσεις πριν από τις ενώσεις (ανεβάζουμε τις ενώσεις όσο πιο ψηλά μπορούμε στο δέντρο).

Με τον τρόπο αυτό, εκτελούμε τις συνενώσεις τοπικά, με αποτέλεσμα να μειώνουμε νωρίς την πληροφορία που διακινούμε στο δίκτυο.

Κανόνας 5

Όπου έχουμε αυξημένη μετα-πληροφορία, μπορούμε να τη χρησιμοποιούμε για να αποφεύγουμε άσκοπες συνδέσεις.

Για παράδειγμα, αν γνωρίζουμε ότι DEPTNUM < 10 $\Rightarrow$ AREA = NORTH, έστω και αν αυτό δε φαίνεται στον ορισμό των τμημάτων, μπορούμε να αποφύγουμε τη συνένωση ενός τμήματος με AREA = SOUTH με κάποιο τμήμα DEPT_TASK με συνθήκη DEPTNUM < 10.

Κανόνας 6

Χρησιμοποιούμε όσο το δυνατό πιο συχνά semi-joins αντί για joins, με σκοπό να μειώσουμε τη διακινούμενη πληροφορία.

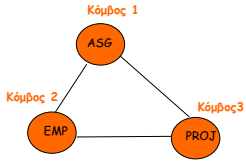
PROJ

PROJ1 = $\sigma_{PNO \leq P2}$ (PROJ)

PROJ2 = $\sigma_{PNO > P2}$ (PROJ)

EMP(ENO, ENAME, TITLE)
PROJ(PNO, PNAME, BUDGET)
ASG(ENO, PNO, RESP, DUR)

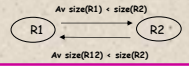
select BUDGET
from ASG, PROJ
where PROJ.PNO = ASG.PNO
and ASG.PNO =P4



size(EMP) = 700
size(ASG) = 100
size(PROJ) = 300
size (EMP J ASG) = 300
size(ASG J PROJ) = 200
size(EMP J PROJ) = 400

Αποτέλεσμα στον κόμβο 1
size(result) = 500

Διαφορετικοί τρόποι εκτέλεσης (χωρίς semijoin)



Υπόθεση: μεταφέρουμε πάντα την μικρότερη σχέση

Άσκηση 3

Employees(eid, did, sal)

Departments(did, mgrid, budget)

mgrid ξένο κλειδί (eid)

Κάθε πλειάδα 20 bytes, μέγεθος eid 10 bytes, μέγεθος mgrid 10 bytes, sal, budget τιμές ομοιόμορφα κατανεμημένες 0 – 1.000.000. Μέγεθος σελίδας: 4000 bytes

Employees: 100.000 σελίδες Departments: 5.000 σελίδες

I/O t_d επικοινωνία t_s

Άσκηση 3(a)

```
select *
from Employees E, Departments D
where E.eid = D.mgrid
```

1% employees είναι managers. Σχέση Emp στη Νάπολη, σχέση Dept στο Βερολίνο, αποτέλεσμα στο Νέο Δελχί. Κόστος των παρακάτω. Κόστος join R1 J R2: 3(size(R1) + size(R2))

1. Υπολογισμός στην N, μεταφορά Dept στο N, μεταφορά αποτελέσματος στο D
2. Υπολογισμός στο B, μεταφορά Emp στο B, μεταφορά αποτελέσματος στο D
3. Υπολογισμός της ερώτησης στη N, με semijoin, μεταφορά αποτελέσματος στο D
4. Υπολογισμός της ερώτησης στο B, με semijoin, μεταφορά αποτελέσματος στο D

Άσκηση 3

Employees(eid, did, sal)

Departments(did, mgrid, budget)

mgrid ξένο κλειδί (eid)

Κάθε πλειάδα 20 bytes, μέγεθος eid 10 bytes, μέγεθος mgrid 10 bytes, sal, budget τιμές ομοιόμορφα κατανεμημένες 0 – 1.000.000. Μέγεθος σελίδας: 4000 bytes

Employees: 100.000 σελίδες Departments: 5.000 σελίδες

I/O t_d επικοινωνία t_s

Άσκηση 3(b)

Αποθηκευμένη σε 10 κόμβους. Dept οριζόντια κατάτμηση και στους 10 κόμβους με βάση το did, με τον ίδιο αριθμό πλειάδων σε κάθε κόμβο. Emp οριζόντια κατάτμηση με βάση την τιμή στο πεδίο sal, sal ≤ 100.000 στον 1ο κόμβο, 100.000 < sal ≤ 200.000 στον 2ο κόμβο κοκ. Ειδικά το sal ≤ 100.000 έχει αντίγραφα σε όλους τους κόμβους. Δώστε το καλύτερο πλάνο για τα παρακάτω:

1. Υπολογισμός φυσικής συνένωσης με βάση την υπόθεση ότι μεταφέρουμε τη μικρότερη σχέση
2. Εύρεση του πιο υψηλόμισθου υπαλλήλου
3. Εύρεση του πιο υψηλόμισθου υπαλλήλου με μισθό μικρότερο των 100.000