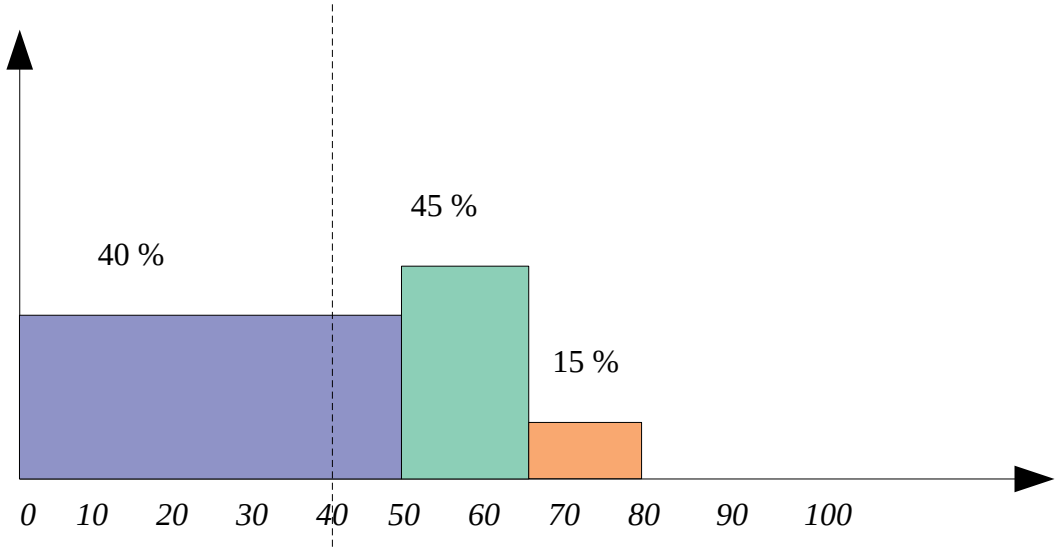
Προχωρημένα Θέματα Τεχνολογίας & Εφαρμογών Βάσεων Δεδομένων
Άσκηση 3

Όθωνας Γκαβαρδίνας ΑΜ: 2620

ΘΕΜΑ 1
(α)



-Απ' το ιστόγραμμα και την ερώτηση ισχύει:
Ποσοστό = (50 – 40 / 50 – 0) * 40 % + 45% + 15 % =
= 1/5 * 40% + 60%
= 8% + 60% = 68%

Κάθε πλειάδα της R αποτελεί το 1/30 μιας σελίδας.
Κάθε πλειάδα της S αποτελεί το 1/20 μιας σελίδας.
Συνόλικα μια πλειάδα που διαθέτει τα στοιχεία και των δύο αποτελεί:
1/30 + 1/20 = (20 + 30) / 600 = 5 / 60 = 1/12 μιας σελίδας στο δίσκο

1.
Το B είναι κλειδί, οπότε συμμετέχουν 100 διαφορετικές τιμές της S.
Το A είναι ξένο κλειδί στο B, οπότε, η R θα έχει μόνο τιμές του S.
Στο αποτέλεσμα, θα περιλαμβάνονται όλες οι πλειάδες της R.
Συνεπώς 150.000 * 100 = 15.000.000 πλειάδες.
Δεν υπάρχουν εδώ ελάχιστα και μέγιστα. Μόνο μία περίπτωση.

Στην περίπτωση αυτή:
Μέγεθος αποτελέσματος:
15.000.000 * 68% = 10200000 πλειάδες.

Χώρος στο δίσκο:
10200000 * 1/12 = 850000 σελίδες.

2.
Το Β είναι κλειδί, οπότε συμμετέχουν 100 διαφορετικές τιμές της S.
-Ελάχιστο: Η R, στο Α μπορεί να μην περιλαμβάνει καμία από αυτές τις τιμές.
Τότε το αποτέλεσμα θα έχει 0 πλειάδες και καθόλου χώρο στο δίσκο.

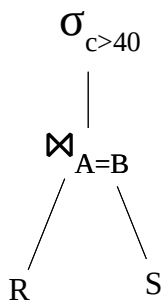
-Μέγιστο: Η R στο Α μπορεί να περιλαμβάνει στη χειρώτερη περίπτωση τις 100 διαφορετικές τιμές διασκορπισμένες σε όλες τις πλειάδες της. Τότε τα ταιριάσματα που θα γίνουν θα φέρουν ως αποτέλεσμα, 150.000 * 100 = 15.000.000 πλειάδες.
Τότε το αποτέλεσμα θα είναι αυτό της 1. περίπτωσης.

3.
Ελάχιστο: Το Α και το Β μπορεί να μην έχουν κανένα κοινό.
Τότε το αποτέλεσμα θα έχει 0 πλειάδες και καθόλου χώρο στο δίσκο.

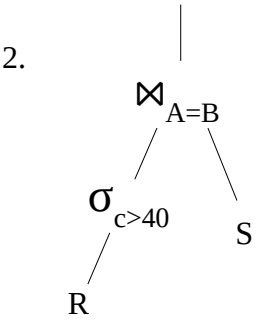
Μέγιστο: Το Α και το Β μπορεί να έχουν όλες τις τιμές κοινές. Τότε 150.000 * 100 = 15.000.000 πλειάδες.
Τότε το αποτέλεσμα θα είναι αυτό της 1. περίπτωσης.

(β)

1.



2.



Στο πλάνο 1, μπορεί να γίνει pipelining, επειδή είναι αριστεροβαθές δένδρο, και οι πλειάδες μπορούν να εξαχθούν και από την R, και από την S ταυτόχρονα.
Στο πλάνο 2, δεν μπορεί να γίνει άμεσα pipelining, μπορεί όμως να εφαρμοστεί πρώτα το φίλτρο της επιλογής σε αυτό και στη συνέχεια να γίνει pipelining στα αποτελέσματα του φίλτρου με τις πλειάδες του S.

Επιλέγω το πλάνο 2.

(γ)
Με βάση το (α) ερώτημα, αν υποθέσουμε πως η γενική περίπτωση της άσκησης είναι το Β να είναι πρωτεύον κλειδί και το Α ξένο κλειδί, έχουμε range query με:

<u>Selectivity</u>: 15.000.000 * 68% = 10200000 πλειάδες.

<u>Κόστος I/O</u>:
Για κάθε μία περίπτωση εφαρμόζω την αντίστοιχη φόρμουλα.
M: σελίδες R (150000 / 30 = 5000)
m: πλειάδες R (150000)
N: σελίδες S (100 / 20 = 5)
n: πλειάδες S (100)

1. Κόστος = M + M*N = 5000 + 5000*100 = 505000 I/O
2. Κόστος = M + ⌐M/B¬ *N = 5000 + ⌐150000/52¬ *100 = 5000 + 288500 = 293500 I/O
3. θεωρείται τυπικά, Κόστος = M + N = 5000 + 5 = 5005 I/O
ενώ μπορεί να φτάσει έως M * N = 5000 * 5 = 25000 I/O
4. Κόστος = 3 * (M+N) = 3 * 5005 = 15015 I/O

(δ)
Πρέπει να ισχύουν οι εξής συνθήκες:
1. Πρέπει να σπρωχθούν προς τα φύλλα του δέντρου του πλάνου εκτέλεσης οι πράξεις επιλογής.
2. Πρέπει να τοποθετηθούν οι πράξεις προβολής από το βελτιστοποιητή σε κάθε τελεστή, έτσι ώστε να περνούν μόνο τα γνωρίσματα που χρειάζονται.
3. Δεν πρέπει να γίνεται χρήση καρτεσιανού γινομένου.
4. Το πλάνο να είναι ένα αριστεροβαθές δένδρο.

Το πλάνο εκτέλεσης που έχω επιλέξει στο (β) ερώτημα ικανοποιεί αυτές τις συνθήκες, και επομένως διαθέτει το ελάχιστο δυνατό κόστος, που έχει υπολογισθεί στο ερώτημα (γ).

ΘΕΜΑ 2

1.

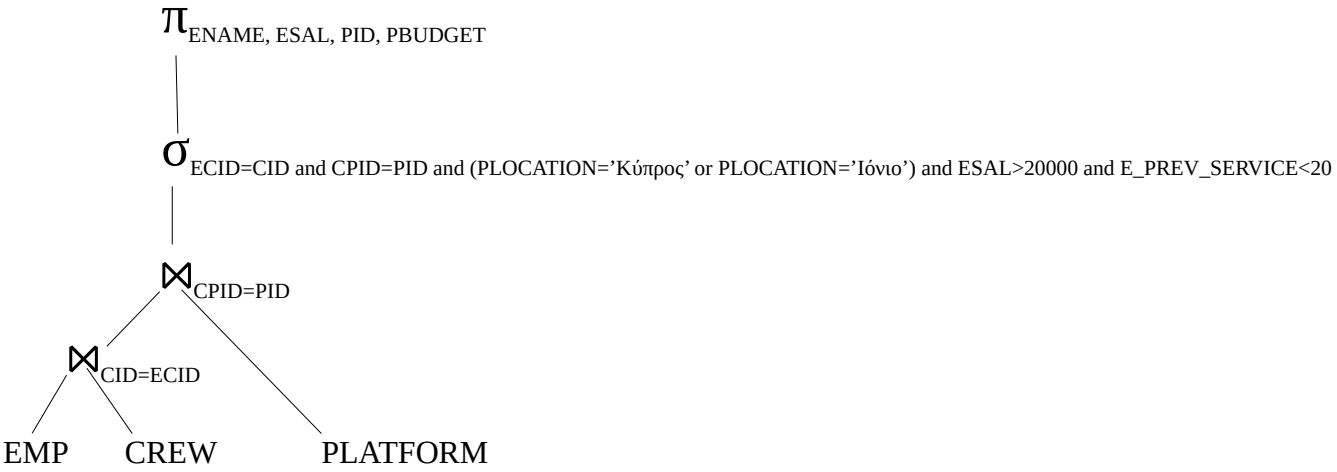
```
SELECT ENAME, ESAL, PID, PBUDGET
FROM EMP, CREW, PLATFORM
WHERE ECID=CID and CPID=PID
and (PLOCATION='Κύπρος' or PLOCATION='Ιόνιο')
and ESAL>20000 and E_PREV_SERVICE<20;
```

$\pi_{\text{ENAME, ESAL, PID, PBUDGET}}$ ($\sigma_{\text{PLOCATION='Κύπρος' or PLOCATION='Ιόνιο'}}$ (PLATFORM) ..

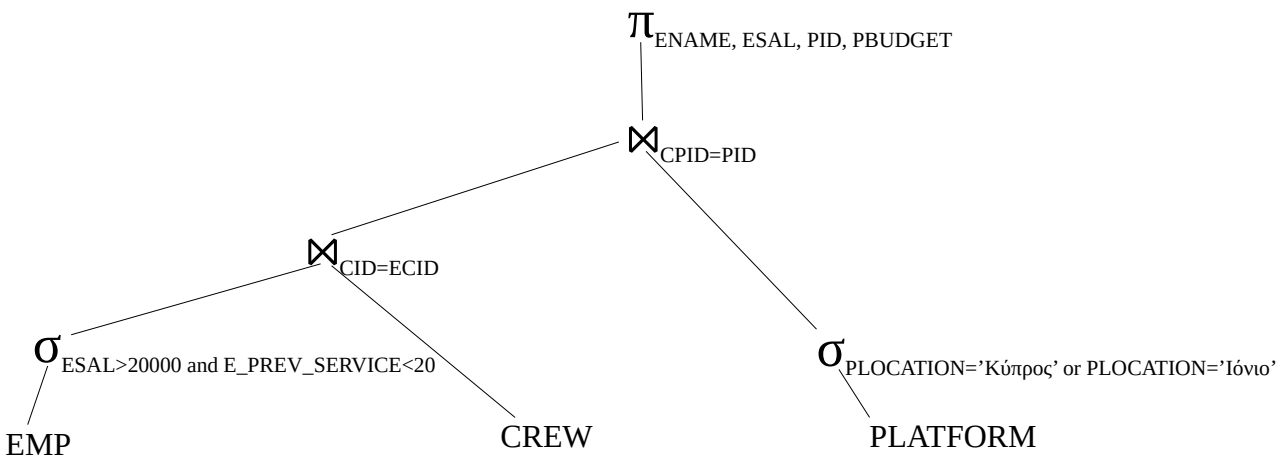
.. $\bowtie_{\text{PID=CPID}}$ (CREW) $\bowtie_{\text{CID=ECID}}$ $\sigma_{\text{ESAL>20000 and E\_PREV\_SERVICE<20}}$ (EMP))

2.

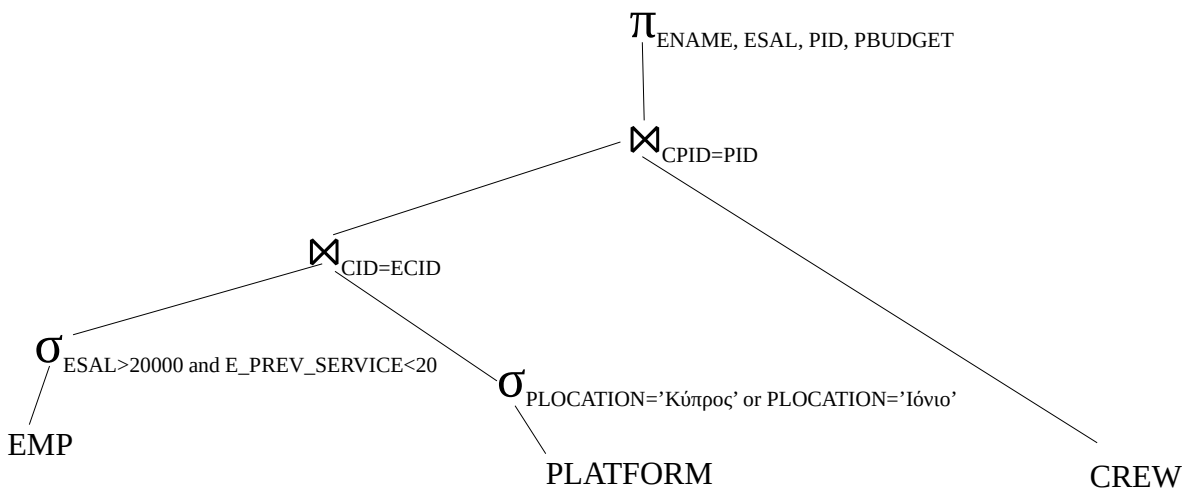
Τυφλή μετάφραση από SQL:



1ο εναλλακτικό πλάνο:



2ο εναλλακτικό πλάνο:



Στα δύο παραπάνω πλάνα πληρούνται οι εξής προυποθέσεις:
-Τα selections έχουν τοποθετηθεί όσο πιο κοντά στα φύλλα γίνεται.
-Για κάθε τελεστή γίνονται projections με βάση τα πεδία που χρειάζεται ο DBMS.
-Δεν περιλαμβάνονται καρτεσιανά γινόμενα.
-Τα πλάνα αποτελούν αριστεροβαθή δέντρα, δηλαδή, σε κάθε join πράξη, εκτός της πρώτης, αριστερά βρίσκεται το output της προηγούμενης πράξης και δεξιά βρίσκεται πίνακας.

3.

Θα επιλέξω το 1ο εναλλακτικό πλάνο εκτέλεσης.

Γενικά, για nr εγγραφές της σχέσης r, και την πράξη $\sigma_{A<=v}(r)$, με v σταθερά και A το γνώρισμα που με ενδιαφέρει ισχύει:

εγγραφές = nr * (v - min(A, r)) / (max(A, r) – min(A, r))

και όταν, στην επιλογή υπάρχει σύζευξη, ισχύει ο τύπος:
nr * (Σi πιθi) / nr^n

$\sigma_{ESAL>20000 \text{ and } E\_PREV\_SERVICE<20}$

nr = 1000
min(EMP, ESAL) = 20.000
max(EMP, ESAL) = 600000
min(EMP, E_PREV_SERVICE) = 0
max(EMP, E_PREV_SERVICE) = 39

Επειδή το γνώρισμα E_PREV_SERVICE ξεκινά απ’ το 0, προσθέτω +1 και +2, για να προκύψουν συνολικά 40 πλειάδες. (έστω 1 με 41)

πιθ1 = 1000 * (60000 – 20001) / (60000 – 20000) = 1000 * 39999 / 40000 =
39999000 / 40000 = 999,975

πιθ2 = 1000 * (21 – 1 - 1) / (41 – 1) = 950

εγγραφές = 1000 * πιθ1 * πιθ2 / 1000 * 1000 = 949976250 / 1000000 = 949,97625
(~950)

$\sigma_{PLOCATION='Κύπρος' \text{ or } PLOCATION='Ιόνιο'}$

Ο τύπος για ισότητα είναι: selectivity = nr / V(A, r)
όπου V(A, r) είναι ο αριθμός των ξεχωριστών τιμών εμφάνησης της σχέσης r, για την ιδιότητα A.

και για τη διάζευξη ισχύει ο τύπος selectivity = 1 - (1 – selectivity1/nr) * (1 – selectivity2/nr) *...
και έτσι παίρνω μια πιθανότητα η οποία ονομάζεται

nr = 20
V(PLOCATION, PLATFORM) = 5

πιθ1 = 20/5 = 4

πιθ2 = 20/5 = 4

selectivity = 1 – (1 – 4/20) * (1 – 4/20) = 1 – 4/5*4/5 = 1 – 0.64 = 0.36

και
εγγραφές = nr * selectivity = 20 * 0.36 = 7,2 (~7)

$\bowtie_{CID=ECID}$

Εγγραφές που έρχονται από αριστερά: 950
Εγγραφές που έρχονται από δεξιά: Δεν γνωρίζω

Αφού γίνεται συνένωση με γνώρισμα το κλειδί CID της σχέσης CREW, το οποίο έχει ξένο κλειδί το ECID, της σχέσης EMP, οι εγγραφές που θα προκύψουν θα είναι αυτές του αριστερού μέλους.

Επομένως: Εγγραφές = 950

$\bowtie_{CPID=PID}$

Εγγραφές που έρχονται από αριστερά: 950
Εγγραφές που έρχονται από δεξιά: 7

Αφού γίνεται συνένωση με γνώρισμα το κλειδί PID της σχέσης PLATFORM, το οποίο έχει ξένο κλειδί το CPID, της σχέσης CREW, οι εγγραφές που θα προκύψουν θα είναι αυτές του αριστερού μέλους.

Επομένως: Εγγραφές = 950

$\pi_{ENAME, ESAL, PID, PBUDGET}$

Με την προβολή από χώρο με αρκετές διαστάσεις θα μετατρέψω το μέγεθος των εγγραφών σε χώρο λιγότερων διαστάσεων.