



Προχωρημένα Θέματα Τεχνολογίας και Εφαρμογών Βάσεων Δεδομένων

Αποθήκες Δεδομένων (Data Warehouses)

Πάνος Βασιλειάδης
pvassil@cs.uoi.gr

Σεπτέμβρης 2016

www.cs.uoi.gr/~pvassil/courses/db_III/

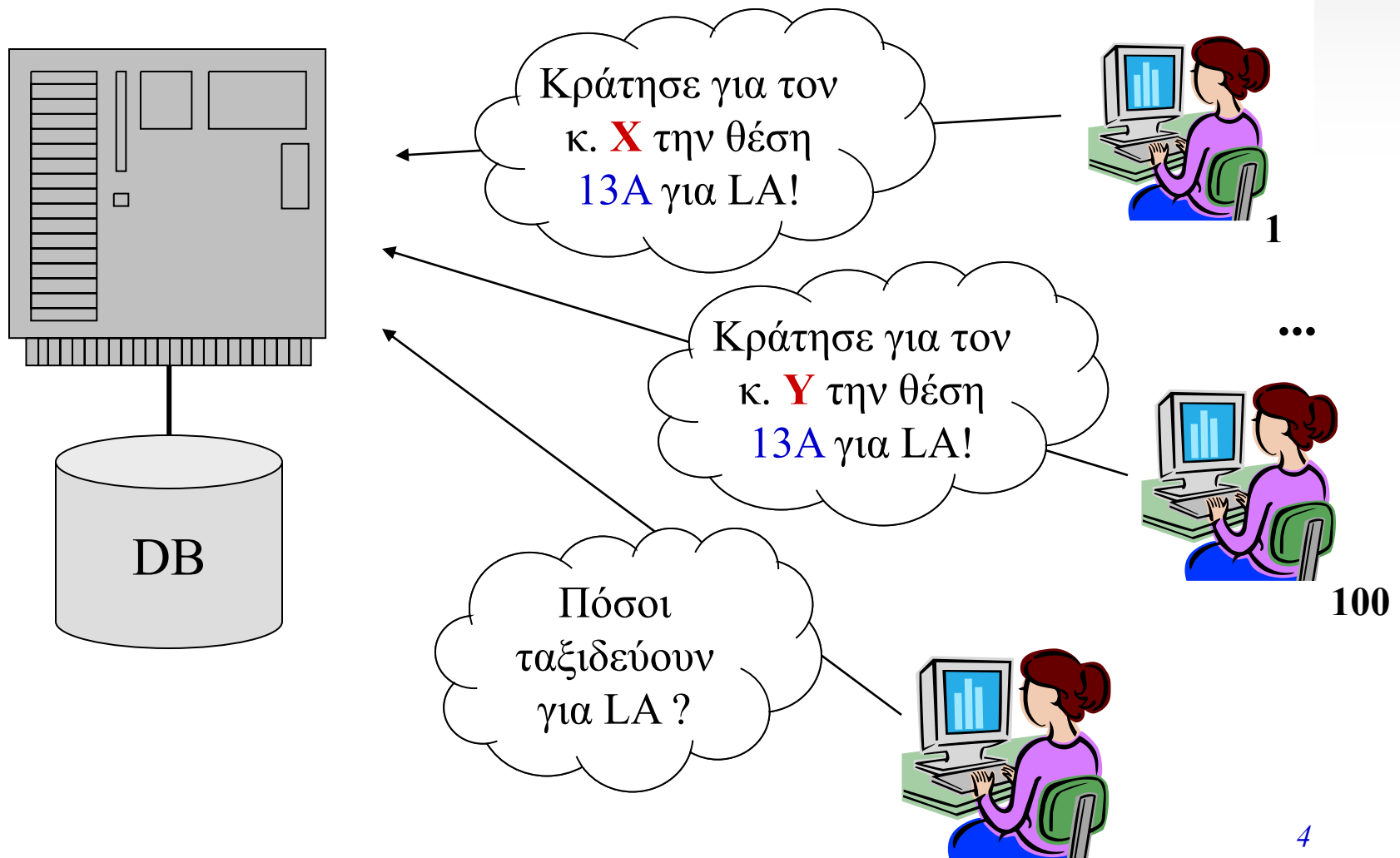
Θεματολόγιο

- Τι είναι οι Αποθήκες Δεδομένων
- Αρχιτεκτονική και σχήμα της Αποθήκης Δεδομένων
- Λειτουργικές διαδικασίες της Αποθήκης Δεδομένων
- Συστήματα επερώτησης της Αποθήκης Δεδομένων
- Λεξικό Μεταπληροφορίας
- Ανακεφαλαίωση

On-Line Transaction Processing (OLTP)

- **Σύστημα Επεξεργασίας Συναλλαγών (OLTP):** Ένα πλήρες σύστημα που περιέχει εργαλεία για **προγραμματισμό** εφαρμογών, **εκτέλεση** και **διαχείριση** των συναλλαγών
- Μια τέτοια εφαρμογή πρέπει να δουλεύει συνεχώς, να αντεπεξέρχεται αποτυχιών, εξελίσσεται συνεχώς, είναι συνήθως κατανεμημένη και περιλαμβάνει:
 - Βάση Δεδομένων
 - Δίκτυο
 - Προγράμματα για την εφαρμογή
- Εξαιρετικά κρίσιμη για τη λειτουργία κάθε οργανισμού

OLTP – αεροπορική εταιρεία...



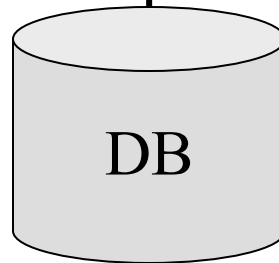
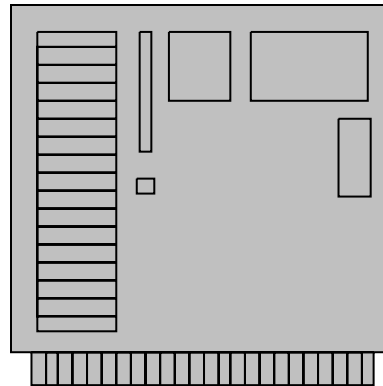
OLTP – τράπεζα...



Γκισέ



ATM



Δάνεια



Πιστωτικές
κάρτες



On-Line Transaction Processing (OLTP) – Χαρακτηριστικά συναλλαγών

- Ελάχιστος χρόνος διαθέσιμος για την εκτέλεση μιας συναλλαγής
- Λιγότερες από 10 προσβάσεις δίσκου
- Περιορισμένος αριθμός υπολογισμών
- Κάτω όριο λειτουργικών απαιτήσεων:
 - ✦ 100 on-line Transactions Per Second (TPS) σε μια ΒΔ της τάξης του 1 GB
- Άνω όριο λειτουργικών απαιτήσεων:
 - ✦ 50000 TPS σε μια ΒΔ μεγαλύτερη του 1 TB

Αναλυτική επεξεργασία των δεδομένων

- Συστήματα Στήριξης Αποφάσεων – Decision Support Systems (DSS)
 - Υποβοήθηση λήψης αποφάσεων με πληροφορίες και αναφορές
- On-Line Analytical Processing (OLAP)
 - Ευέλικτη, υψηλής απόδοσης πρόσβαση και ανάλυση μεγάλου όγκου σύνθετων δεδομένων από διαφορετικές εφαρμογές
 - Ειδικού τύπου ερωτήσεις
 - Οπτικοποίηση / στατιστική ανάλυση / πολυδιάστατη ανάλυση
- Εξόρυξη Γνώσης (Knowledge Discovery / Data Mining)
 - Εξεύρεση patterns σε τεράστιες βάσεις δεδομένων
 - OLAP + Data Mining => On-line Analytical Mining

Παραδείγματα ερωτήσεων OLAP

- Ποιος ήταν ο όγκος πωλήσεων ανά περιοχή και κατηγορία προϊόντος την περασμένη χρονιά;
- Πόσο σχετίζονται οι αυξήσεις τιμών των υπολογιστών με τα κερδών των πωλήσεων τα 10 τελευταία χρόνια;
- Ποια ήταν τα δέκα πρώτα καταστήματα σε πωλήσεις CD;
- Πόσους δίσκους πουλήσαμε στην Δυτική Περιφέρεια το τελευταίο τέταρτο της περσινής χρονιάς σε καταστήματα με κατανάλωση μεγαλύτερη από 100 δίσκους μηνιαίως, και ποιο το κέρδος μας από αυτές τις πωλήσεις;
- Τι ποσοστό από τους πελάτες που αγοράζουν αναψυκτικά, αγοράζουν και πατατάκια;

Λειτουργικά χαρακτηριστικά απαιτήσεων OLAP

- Πρόσβαση σε **μεγάλο όγκο** δεδομένων
- Συμμετοχή αθροιστικών και ιστορικών δεδομένων σε **πολύπλοκες** ερωτήσεις
- Μεταβολή της «οπτικής γωνίας» παρουσίασης των δεδομένων (π.χ., από πωλήσεις ανά περιοχή -> πωλήσεις ανά τμήμα κλπ.)
- Συμμετοχή **πολύπλοκων υπολογισμών** (π.χ. στατιστικές συναρτήσεις)
- **Γρήγορη** απάντηση σε οποιαδήποτε χρονική στιγμή τεθεί ένα ερώτημα (εξ' ου και "On-Line").

Δύο προβλήματα σαν παράδειγμα

✦ Απόδοση

- ✦ Αν μια πολύπλοκη OLAP ερώτηση χρειαστεί να κλειδώσει ένα ολόκληρο πίνακα, τότε **ΟΛΕΣ** οι OLTP συναλλαγές την **περιμένουν** μέχρι να τελειώσει :-(
Φανταστείτε το σε μια τράπεζα...

✦ Εννοιολογική διαφορά και ετερογένεια

- ✦ Αν στην Oracle ΒΔ του **marketing** ο πελάτης είναι **EMP (AT, Name, Surname...)** και στην COBOL ΒΔ των **πωλήσεων** “είναι” **ΑΦΜ, FullName, ...** η επερώτηση δεν είναι πάντα εύκολη...

Η λύση: Αποθήκες Δεδομένων (Data Warehouses)

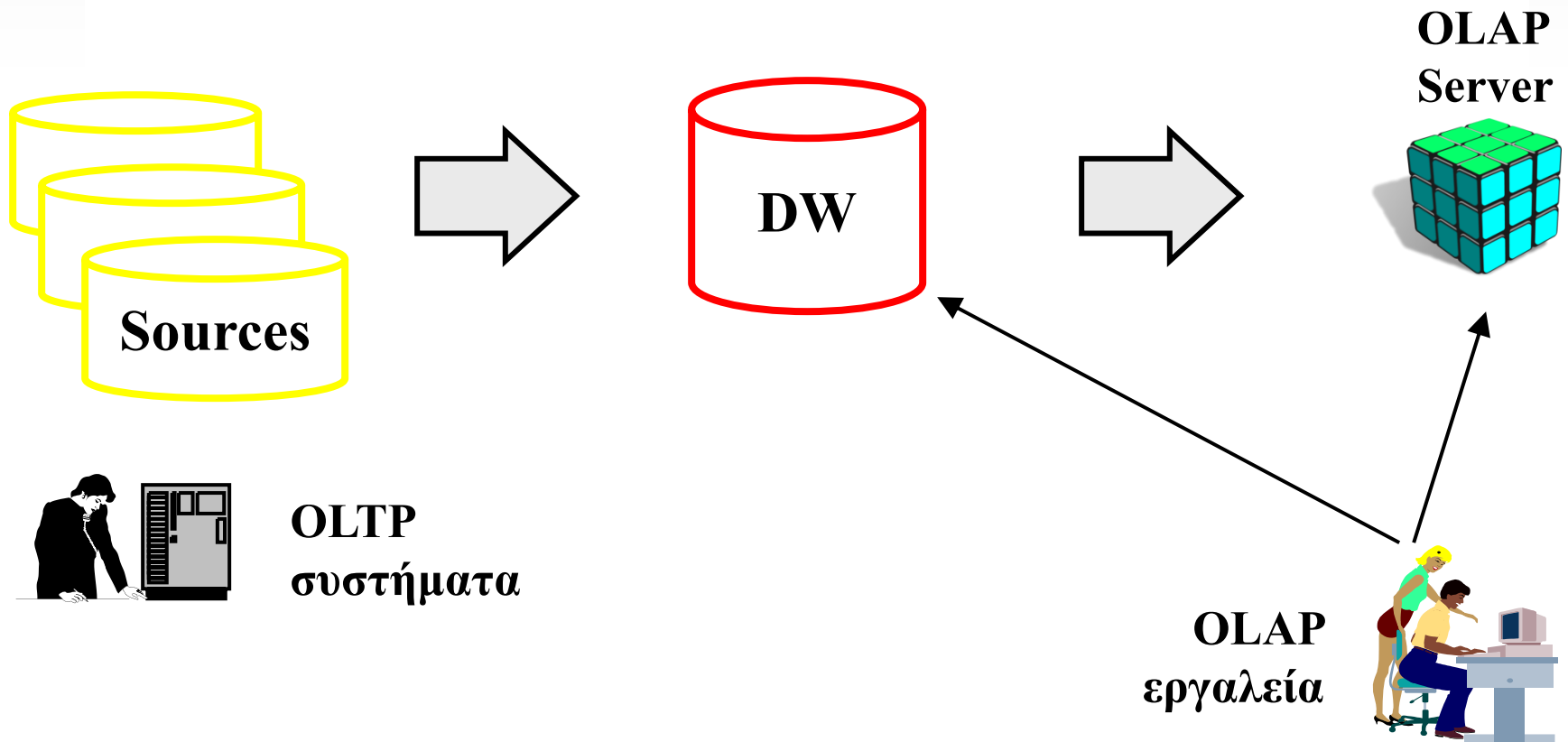
- ✦ Μια **κεντρική ΒΔ** με σκοπό:
 - ✦ την **ολοκλήρωση (integration)** ετερογενών πηγών πληροφοριών (data sources) => συνάθροιση **όλης** της ενδιαφέρουσας πληροφορίας σε **μία** τοποθεσία
 - ✦ την **αποφυγή της σύγκρουσης** μεταξύ OLTP και OLAP (DSS) **συστημάτων** => **απόδοση** εφαρμογών και **διαθεσιμότητα** του συστήματος
- ✦ Συμπληρώνεται ενίοτε και από εξειδικευμένα θεματικά υποσύνολα (**Data Marts**) για περαιτέρω απόδοση των OLAP εφαρμογών

Εναλλακτικά: *Συγκεντρωτικές Αποθήκες Δεδομένων*

Ορισμός. Μια Αποθήκη Δεδομένων είναι ...

- Μια ΒΔ υποστήριξης αποφάσεων, που συντηρείται χωριστά από την ΒΔ παραγωγής (operational database) ενός οργανισμού.
 - S. Chaudhuri, U. Dayal, VLDB'96 tutorial
- Μια συλλογή δεδομένων που χρησιμοποιείται κυρίως για την λήψη αποφάσεων σε ένα οργανισμό, και είναι θεματικά προσανατολισμένη, έχουσα «ολοκληρωμένα» δεδομένα, τα οποία κρατώνται σε βάθος χρόνου χωρίς να διαγράφονται.
 - W.H. Inmon, Building the Data Warehouse, 1992 (ο εφευρέτης του όρου)

OLTP – DW - OLAP



Αιτίες και ιδιότητες της Αποθήκης Δεδομένων

✦ Εννοιολογική εναρμόνιση

- ✦ Οι διαφορετικές πηγές δεδομένων του ίδιου οργανισμού, μοντελοποιούν τις **ίδιες** οντότητες με **διαφορετικούς** τρόπους
- ✦ **Η Αποθήκη Δεδομένων περιλαμβάνει το σύνολο αυτών των δεδομένων κάτω από ένα «εναρμονισμένο» σχήμα βάσης**
- ✦ Επιπλέον, κρατάμε και την ιστορία των δεδομένων που φορτώνουμε κάθε φορά

Εννοιολογική εναρμόνιση

Source 1:
Personnel
(Cobol)

<u>EMP ID</u>	Name	DoB	Salary	Total Income	DeptID
110	Kostas	1/1/72	1500	1200	132
...	...				...

Source 2: Accounting
(DB2)

<u>EMP ID</u>	<u>IL ID</u>	Amount
110	10	1500
110	30	300

**EMP
INCOME**

<u>EMP ID</u>	Name	Age
110	Kostas	30
120	Mitsos	48
130	Roula	29

EMP

<u>IL ID</u>	Descr
10	Μισθός
20	Επίδομα Τέκνων
30	Φόρος
...	...

**Income
Lookup**

Αιτίες και ιδιότητες της Αποθήκης Δεδομένων

✦ Απόδοση

- ✦ Οι εφαρμογές OLAP επιταχύνονται αν τα δεδομένα οργανωθούν με **μη παραδοσιακούς τρόπους** (π.χ., αποκανονικοποιημένα)
- ✦ Οι σύνθετες **OLAP** ερωτήσεις **θα συγκρούονταν** με τις παραδοσιακές **OLTP συναλλαγές**, με αποτέλεσμα την υπερφόρτωση του συστήματος
- ✦ Αν και τεχνικά είναι πλέον εφικτό να ρωτάμε **ετερογενείς πηγές δεδομένων**, **πρακτικά, η Αποθήκη Δεδομένων είναι ο μόνος τρόπος να κάνουμε ερωτήσεις σε αποδεκτούς χρόνους**

Αιτίες και ιδιότητες της Αποθήκης Δεδομένων

✦ Ποιότητα Δεδομένων

- ✦ Η ποιότητα των δεδομένων στις πηγές είναι, σχεδόν πάντα, προβληματική (τα δεδομένα μπορεί να μην είναι πλήρη, να έχουν ασυνέπειες, να είναι παλιά, να παραβιάζουν τους λογικούς και δομικούς κανόνες αξιοπιστίας, ...)
- ✦ Έχει βρεθεί ότι τουλάχιστο 10% των δεδομένων είναι προβληματικά στις πηγές, με αποτέλεσμα οικονομικές απώλειες του 25-40%
- ✦ **Η Αποθήκη Δεδομένων λειτουργεί και ως ένα «buffer» στον οποίο καθαρίζουμε τα δεδομένα**

Αιτίες και ιδιότητες της Αποθήκης Δεδομένων

✦ Διαθεσιμότητα

- ✦ Όσο περισσότερα «αντίγραφα» των δεδομένων, τόσο πιο πολύ το σύστημα είναι διαθέσιμο, αφενός στην Αποθήκη Δεδομένων και αφετέρου στις πηγές

Διαθεσιμότητα: το ποσοστό του χρόνου που το σύστημα είναι σε λειτουργία και προσβάσιμο στις εφαρμογές.

24x7: Οι OLTP εφαρμογές, σε πολλούς οργανισμούς πρέπει να είναι διαθέσιμες 24 ώρες X 7 μέρες τη βδομάδα (π.χ., τράπεζες, αεροπορικές εταιρείες,...)

OLAP - OLTP Ομοιότητες / Διαφορές

	<u>OLTP</u>	<u>OLAP</u>
Δομή	Files/DBMS's	RDBMS
Πρόσβαση	SQL/COBOL/...	SQL + επεκτάσεις
Ανάγκες που καλύπτουν	Αυτοματισμός καθημερινών εργασιών	Άντληση και επεξεργασία πληροφορ. για χάραξη στρατηγικής
Τύπος Δεδομένων	Λεπτομερή Λειτουργικά	Συνοπτικά, Αθροιστικά
Όγκος Δεδομένων	~ 100 GB	~ 1 TB
Φύση Δεδομένων	Δυναμικά, Τρέχοντα	Στατικά, Ιστορικά

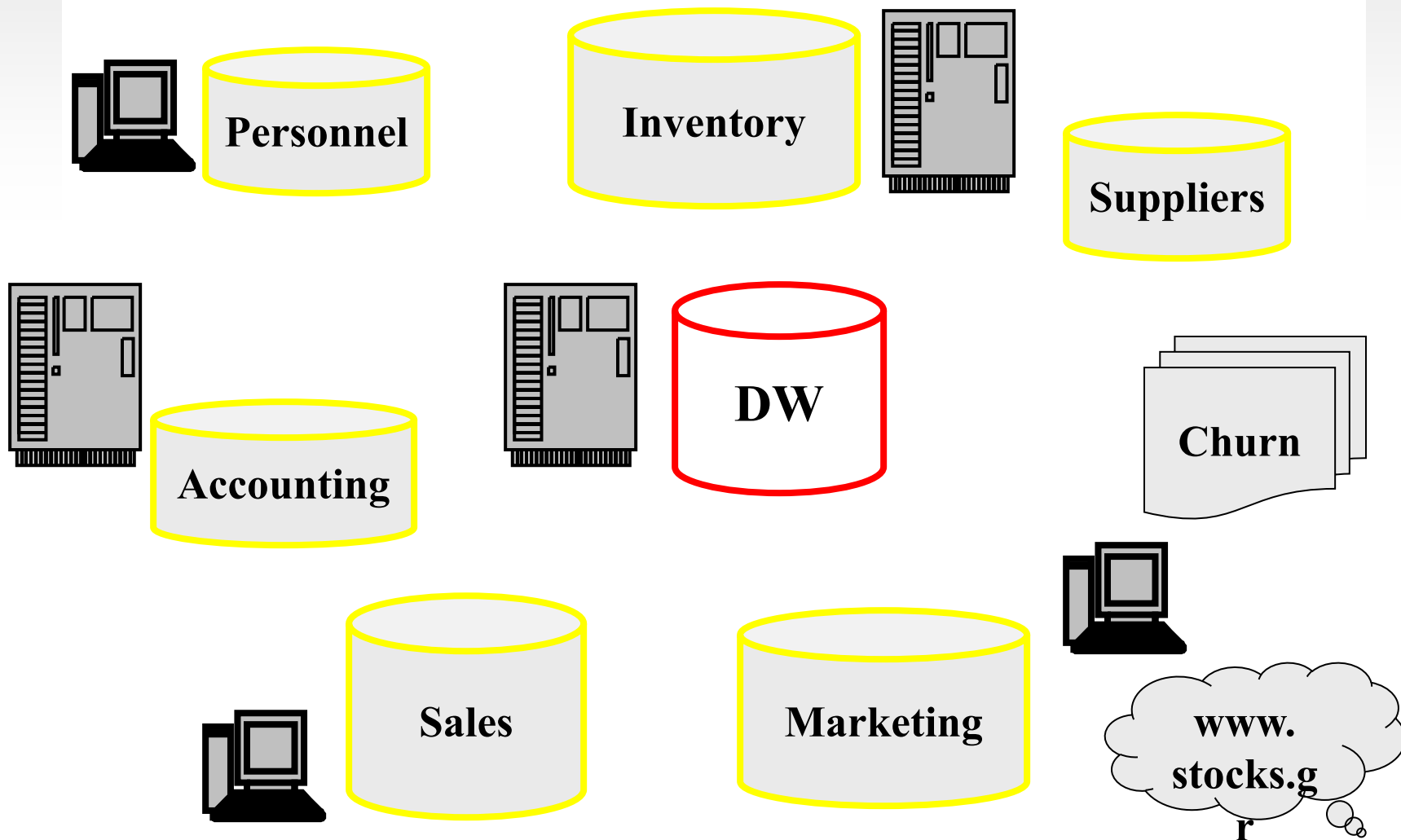
OLAP - OLTP Ομοιότητες / Διαφορές

	<u>OLTP</u>	<u>OLAP</u>
I/O Τύποι	Περιορισμένο I/O Συχνά disk seeks	Εκτεταμένο I/Os disk scans
Τροποποιήσεις	Συνεχείς	Περιοδικές Ενημερώσεις
Μέτρηση Απόδοσης	Throughput	Χρόνος Απόκρισης
Φόρτος	Συναλλαγές με πρόσβαση λίγων εγγραφών	Ερωτήσεις που σαρώνουν εκατομμύρια εγγραφών
Σχεδίαση ΒΔ	Κατευθυνόμενη από Εφαρμογή	Κατευθυνόμενη από Περιεχόμενο

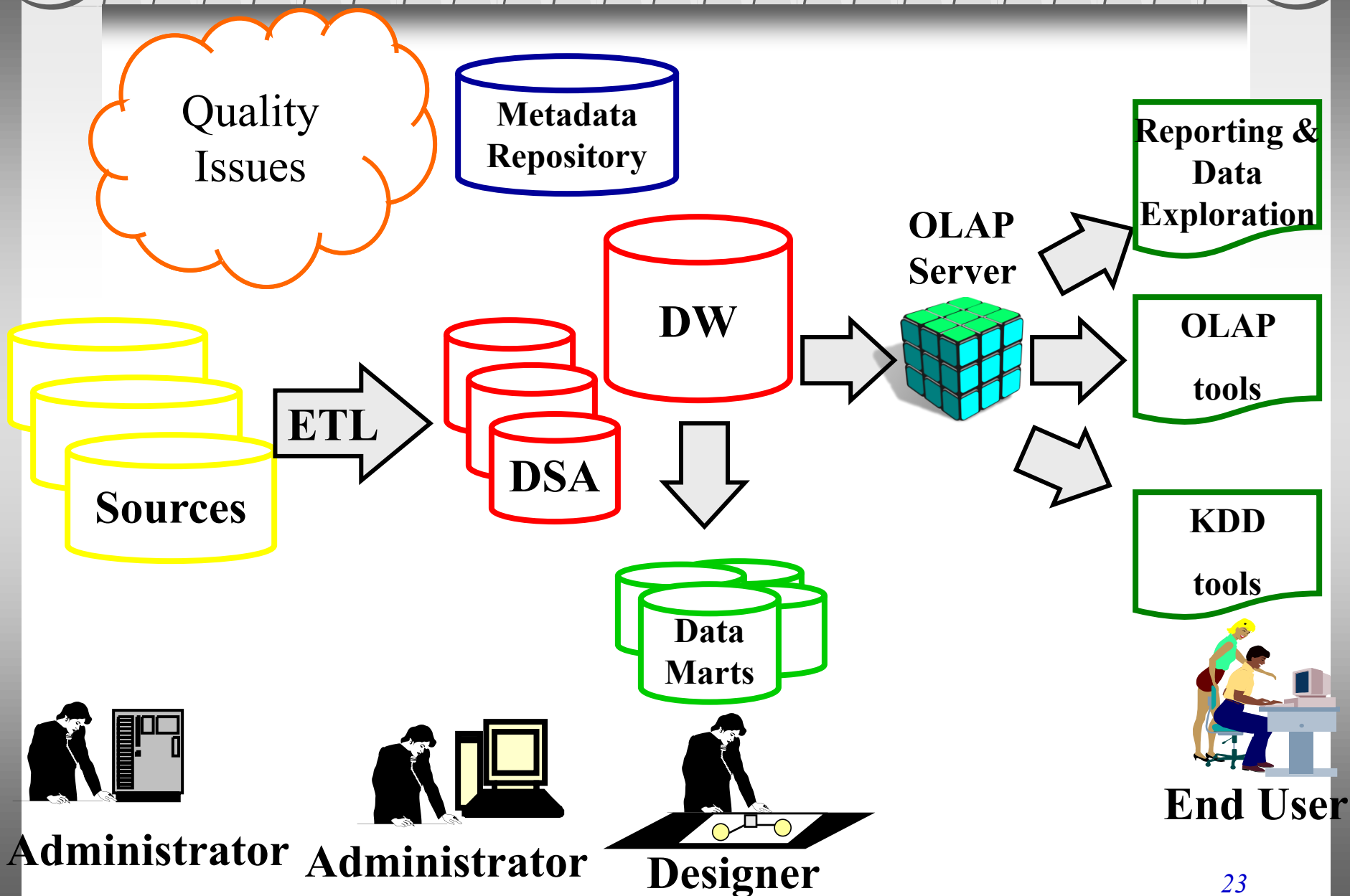
OLAP - OLTP Ομοιότητες / Διαφορές

	<u>OLTP</u>	<u>OLAP</u>
Τυπικοί Χρήστες	Χαμηλόβαθμοι Υπ.	Υψηλόβαθμοι Υπ.
Χρήση	Μέσω προκατασκευασμένων φορμών	Ad-hoc
Αριθμός Χρηστών	Χιλιάδες	Δεκάδες
Εστίαση	Εισαγωγή Δεδομένων	Εξαγωγή Πληροφοριών

Παράδειγμα Αποθήκης Δεδομένων – Telecom Co.



Αρχιτεκτονική της Αποθήκης Δεδομένων



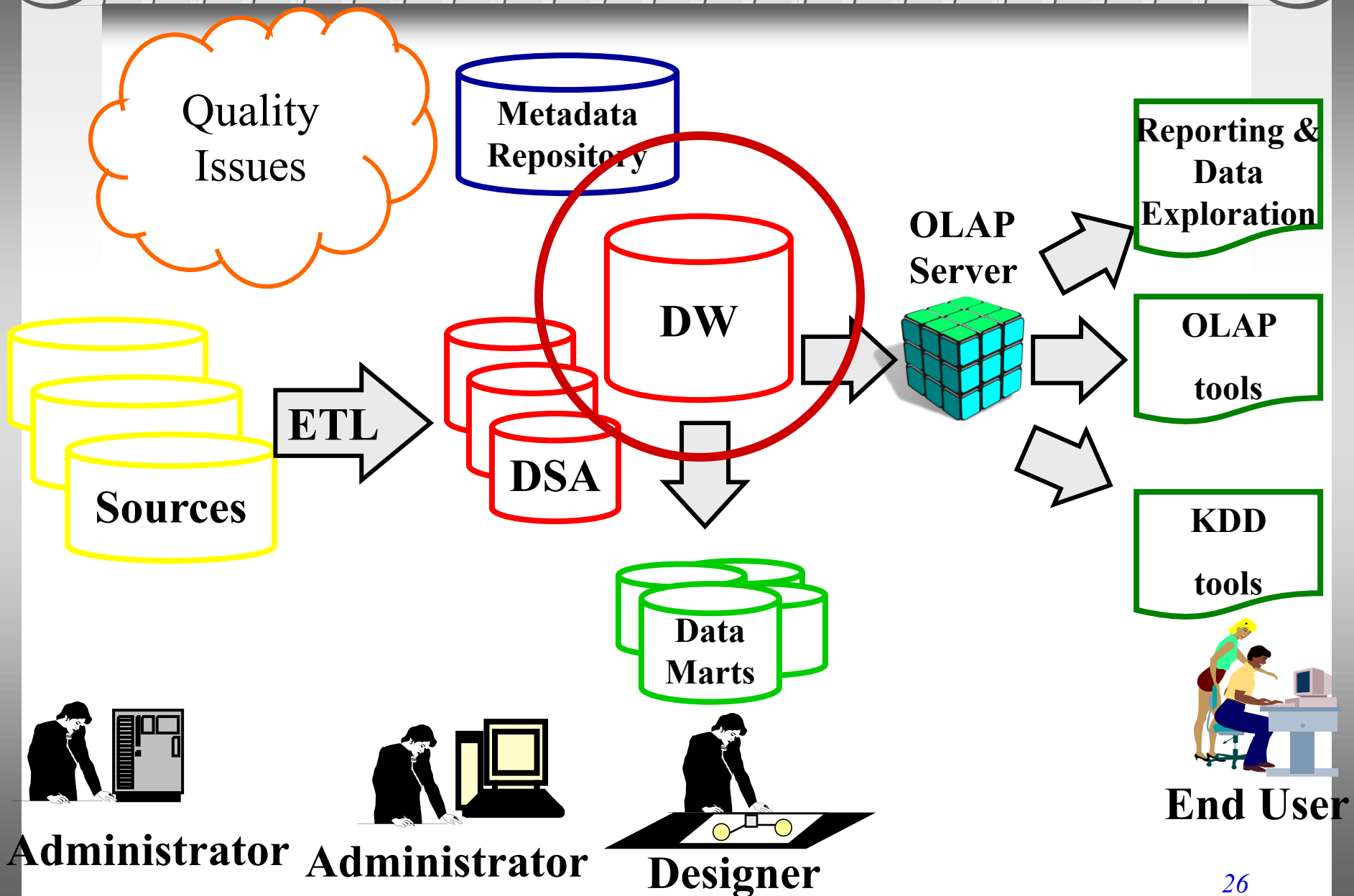
Αρχιτεκτονική της Αποθήκης Δεδομένων

- **Sources (Πηγές):** Κάθε πηγή από την οποία η Αποθήκη Δεδομένων αντλεί δεδομένα.
- **Data Staging Area:** Μια ΒΔ στην οποία εκτελούνται οι μετασχηματισμοί και ο καθαρισμός των δεδομένων πριν την φόρτωση στην Αποθήκη Δεδομένων
- **Αποθήκη Δεδομένων (DW), Συλλογές Δεδομένων (Data Marts):** Τα συστήματα που αποθηκεύονται τα δεδομένα που παρέχονται προς τους χρήστες.
- **Βάση Μετα-Δεδομένων (Metadata Repository):** Το υποσύστημα αποθήκευσης πληροφορίας σχετικά με τη δομή και λειτουργία όλου του συστήματος.

Αρχιτεκτονική της Αποθήκης Δεδομένων

- **ETL (Extract-Transform-Load) εφαρμογές:**
Εφαρμογές που εκτελούν τις διαδικασίες εξαγωγής, μεταφοράς, μετασχηματισμού, καθαρισμού και φόρτωσης των δεδομένων από τις πηγές στην Αποθήκη Δεδομένων.
- **Εφαρμογές Ανάλυσης:** Εφαρμογές παραγωγής αναφορών, OLAP , DSS, Data Mining

Αποθήκη Δεδομένων



Ζητήματα Ενδιαφέροντος

- **Servers** – Τεχνολογικές λύσεις
- Τα σχήματα της Αποθήκης Δεδομένων (εννοιολογικό σχήμα / λογικό σχήμα)
- Σχεδίαση της Αποθήκης Δεδομένων (από την πλευρά της Τεχνολογίας Λογισμικού)
- Ειδικά θέματα σε επίπεδο φυσικού σχήματος (indexing, επεξεργασία ερωτήσεων, παραλληλία, ...) δε θα μας απασχολήσουν εδώ

Servers & Τεχνολογικές λύσεις

- **DW**: Σχεσιακά και επεκτεταμένα σχεσιακά DBMS
- **OLAP**:
 - ✦ Relational OLAP (ROLAP)
 - ✦ Multidimensional OLAP (MOLAP)

Relational DBMS's & Αποθήκες Δεδομένων

- Εξειδικευμένες τεχνικές δεικτοδότησης (**indexing**)
- Εξειδικευμένες τεχνικές συνένωσης (**join**)
- Διαμοίραση των δεδομένων (**data partitioning**) και χρήση παράλληλων τεχνικών
- Εξειδικευμένες τεχνικές αποθήκευσης και επεξεργασίας ερωτήσεων για συναθροίσεις δεδομένων (**aggregates**)
- **Επεκτάσεις της SQL** και της επεξεργασίας των σχετικών ερωτήσεων

ROLAP Servers

- Βασική ιδέα: **χρήση ενός RDBMS ως μέσου αποθήκευσης και επερώτησης** (με όλα τα σχετικά πλεονεκτήματα)
- Επιπλέον λειτουργικότητα των client εργαλείων:
 - Δυνατότητα επαναχρησιμοποίησης συναθροίσεων
 - Χρήση multi statement SQL
 - Βελτιστοποίηση των ερωτήσεων ανά RDBMS
- Αργά ως συστήματα (μέχρι στιγμής τουλάχιστον)
- + Δυνατότητα υποβολής οποιασδήποτε ερώτησης
- + Εύκολη χρήση από τους administrators που γνώριζαν τη σχεσιακή τεχνολογία

Πλάνο και στατιστικά από ένα ROLAP εργαλείο

```

select a3.EKSAM_FOIT_CODE EKSAM_FOIT_CODE,
       max(a3.DESCR) DESCR,
       a2.SEX SEX,
       (SUM(a1.FOO1)) M0000000
from   FACT1 a1,
       FOITITIS a2,
       EKSAM_FOIT a3
where  a2.FOITITIS_CODE = a1.FOITITIS_CODE
and    a1.EKSAM_FOIT_CODE = a3.EKSAM_FOIT_CODE
and    ((((((a2.SEX = '1'))
and      ((EXISTS (select *
                  from    EKSAM_FOIT m1
                  where    m1.EKSAM_FOIT_CODE = a3.EKSAM_FOIT_CODE
                  and      m1.CATEGORY = 'EAPINO')))))
or      (((a2.SEX = '2'))
and      ((EXISTS (select *
                  from    EKSAM_FOIT m1
                  where    m1.EKSAM_FOIT_CODE = a3.EKSAM_FOIT_CODE
                  and      m1.CATEGORY = 'EAPINO')))))
or      (((a2.SEX = '1'))
and      ((EXISTS (select *
                  from    EKSAM_FOIT m1
                  where    m1.EKSAM_FOIT_CODE = a3.EKSAM_FOIT_CODE
                  and      m1.CATEGORY = 'XEIMEPINO')))))
or      (((a2.SEX = '2'))
and      ((EXISTS (select *
                  from    EKSAM_FOIT m1
                  where    m1.EKSAM_FOIT_CODE = a3.EKSAM_FOIT_CODE
                  and      m1.CATEGORY = 'XEIMEPINO'))))))))
group by a3.EKSAM_FOIT_CODE, a2.SEX

```

PERFORMANCE METRICS (Seconds)

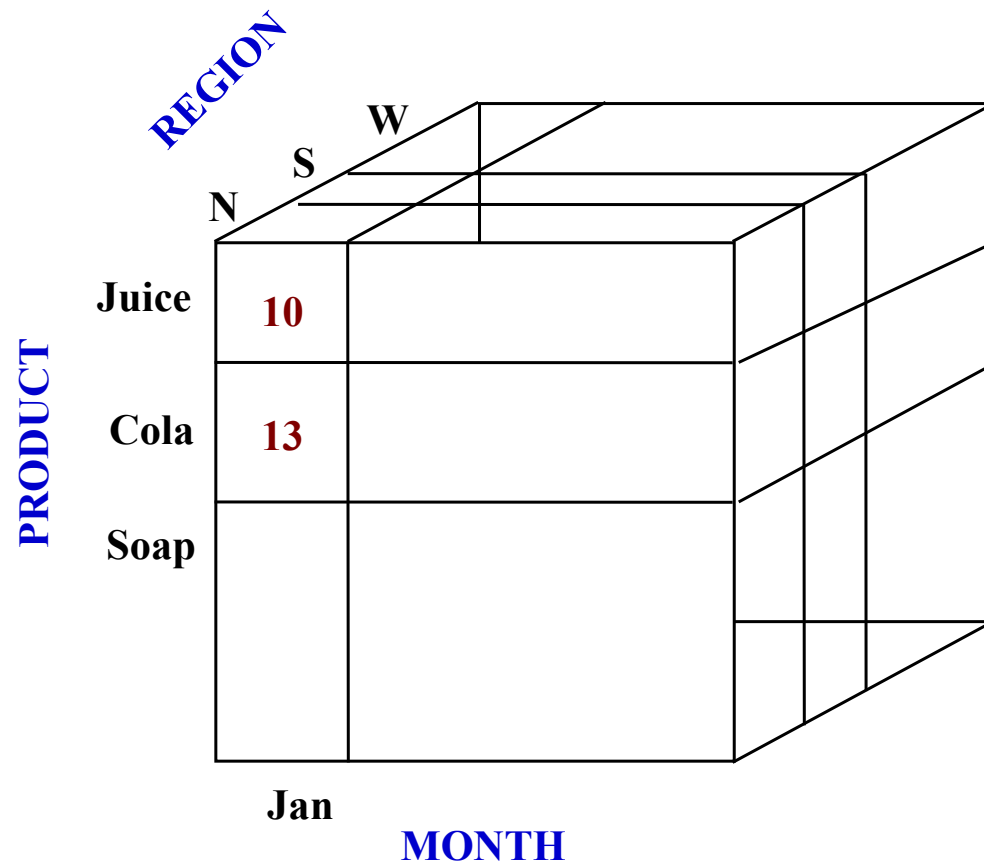
Loading Parameters:	0,0
SQL Generation:	0,4
Executing Query:	0,3
Results Processing:	0,8

Total Machine Time:	1,5
Rows returned from Database :	24

MOLAP Servers

- Η αποθήκευση γίνεται σε **πολυδιάστατους πίνακες** (multi-dimensional arrays)
 - «πίνακες» με την έννοια της άλγεβρας / γλωσσών προγραμματισμού /..., και όχι του σχεσιακού μοντέλου
- Χρήση τεχνικών συμπίεσως (οι πίνακες είναι αραιοί σε βαθμό ως και 80%)
- Στις αρχές του 2002 είχαν το 98% της αγοράς στο πεδίο των client tools
- + Πολύ γρήγοροι υπολογισμοί των λειτουργιών OLAP
- Κανονικά απαιτούν τον **προϋπολογισμό των απαραίτητων συναθροίσεων**

Πολυδιάστατοι πίνακες



Structure of a Multidimensional Data Store

Array header

Array name, Number of dimensions, Number of cells, comments
Dimension 1; Number of positions Pos1 name Pos2 name • • • Fast offset calculation data
Dimension #n

Array data

True
21.43
Denver
4
\$87.65
16
Robinson
3.14

Υλοποίηση
πολυδιάστατων
πινάκων

Μοντέλο Διαστάσεων για OLAP

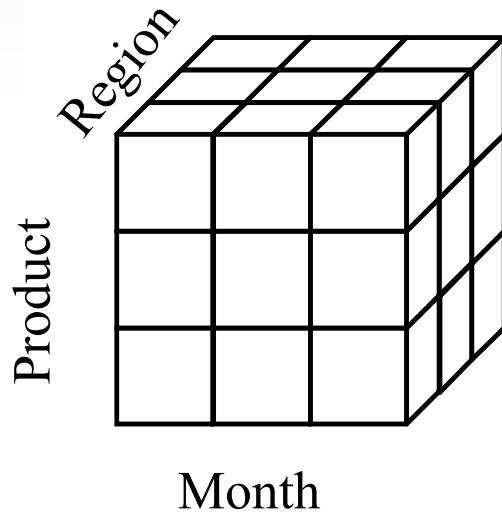
- **Μοντέλο Διαστάσεων** (Dimensional Modeling): τεχνική ειδικά για Αποθήκες Δεδομένων
- Βασίζεται στην θεώρηση των δεδομένων μέσω ενός **πολυδιάστατου** μοντέλου δεδομένων με βασικά στοιχεία πίνακες πολυδιάστατων δεδομένων, και πίνακες διαστάσεων.

Πολυδιάστατο μοντέλο δεδομένων

- Αφορά την ανάλυση κάποιων μετρήσιμων μεγεθών (μέτρων)
 - πωλήσεις, απόθεμα, κέρδος,...
- Διαστάσεις: παράμετροι που καθορίζουν το περιβάλλον (context) των μέτρων
 - ημερομηνία, προϊόν, τοποθεσία, πωλητής, ...
 - διαφορετικά **επίπεδα** λεπτομέρειας ανά διάσταση: μέρα, μήνας, χρόνος,...
- **Κύβοι**: συνδυασμοί διαστάσεων που καθορίζουν κάποια μέτρα
 - Ο κύβος καθορίζει ένα πολυδιάστατο χώρο διαστάσεων, με τα μέτρα να είναι σημεία του χώρου αυτού

Πολυδιάστατο μοντέλο δεδομένων

Sales volume



Διαστάσεις: Product, Region, Date

Ιεραρχίες διαστάσεων:

Industry
|
Category
|
Product

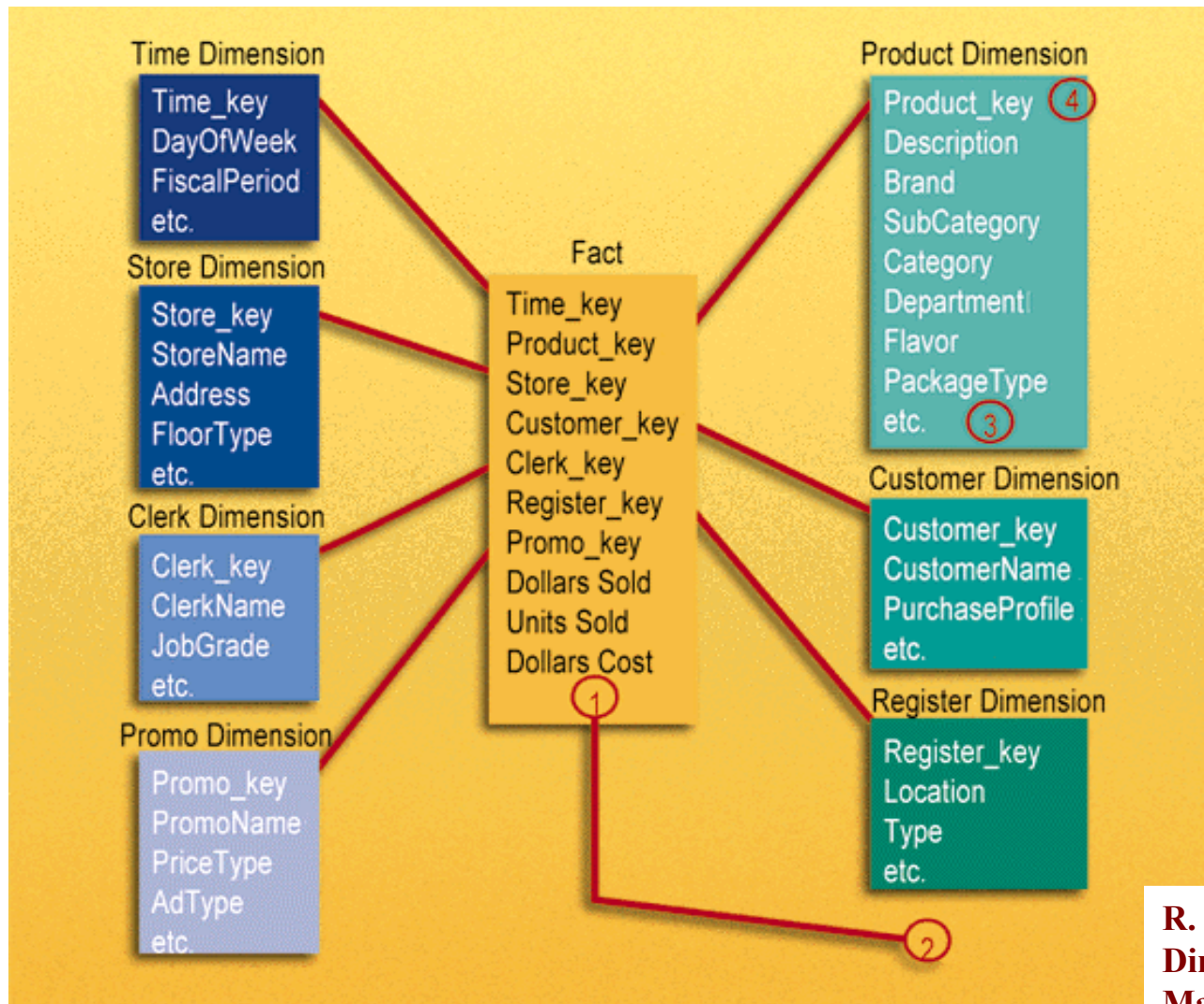
Country
|
Region
|
City
|
Store

Year
|
Quarter
/ \
Month Week
\< /
Day

Μοντέλο Διαστάσεων

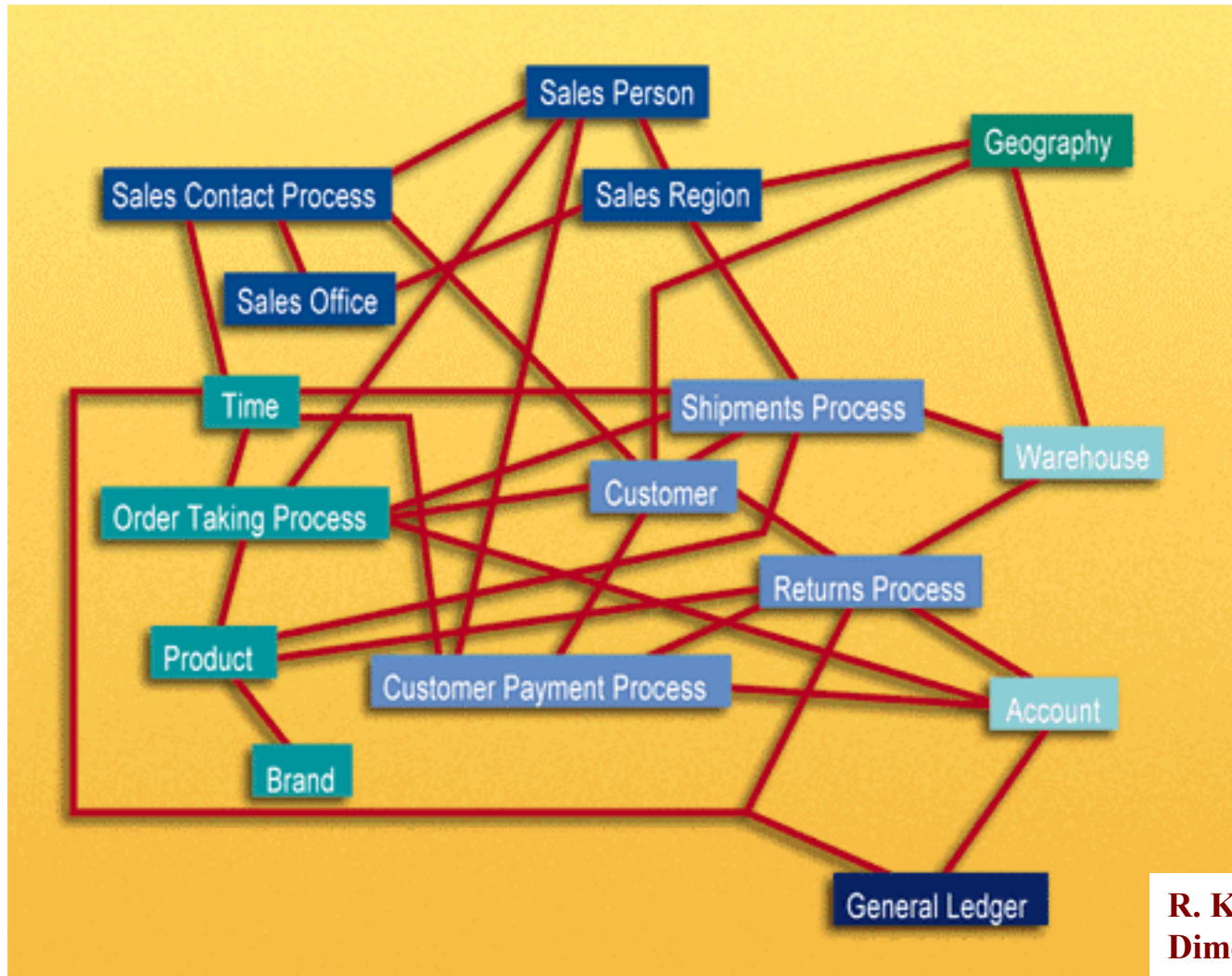
- **Πίνακας Διάστασης (Dimension table)** : Ο πίνακας που περιέχει πληροφορία σχετική με μια διάσταση. Περιέχει ένα τεχνητό κλειδί και από μία στήλη για κάθε επίπεδο της διάστασης (πιθανόν και κάποιες στήλες με επιπλέον χαρακτηριστικά των επιπέδων).
- **Πίνακας Δεδομένων (Fact table)** : Ο πίνακας που υλοποιεί τον υπό μοντελοποίηση κύβο.
 - Κάθε εγγραφή του πίνακα αντιστοιχεί σε ένα κελί του κύβου
 - Για κάθε διάσταση έχει και ένα εξωτερικό κλειδί στον αντίστοιχο πίνακα και για κάθε μέτρο και μια αντίστοιχη στήλη.
 - Το πρωτεύον κλειδί είναι ο συνδυασμός των κλειδιών των διαστάσεων (οι συντεταγμένες του κελιού).

Παράδειγμα σχήματος στο Μοντέλο Διαστάσεων



**R. Kimball, A
Dimensional Modeling
Manifesto, DBMS
Magazine, Aug. 1997**

Παράδειγμα του αντίστοιχου ER σχήματος



**R. Kimball, A
Dimensional Modeling
Manifesto, DBMS
Magazine, Aug. 1997**

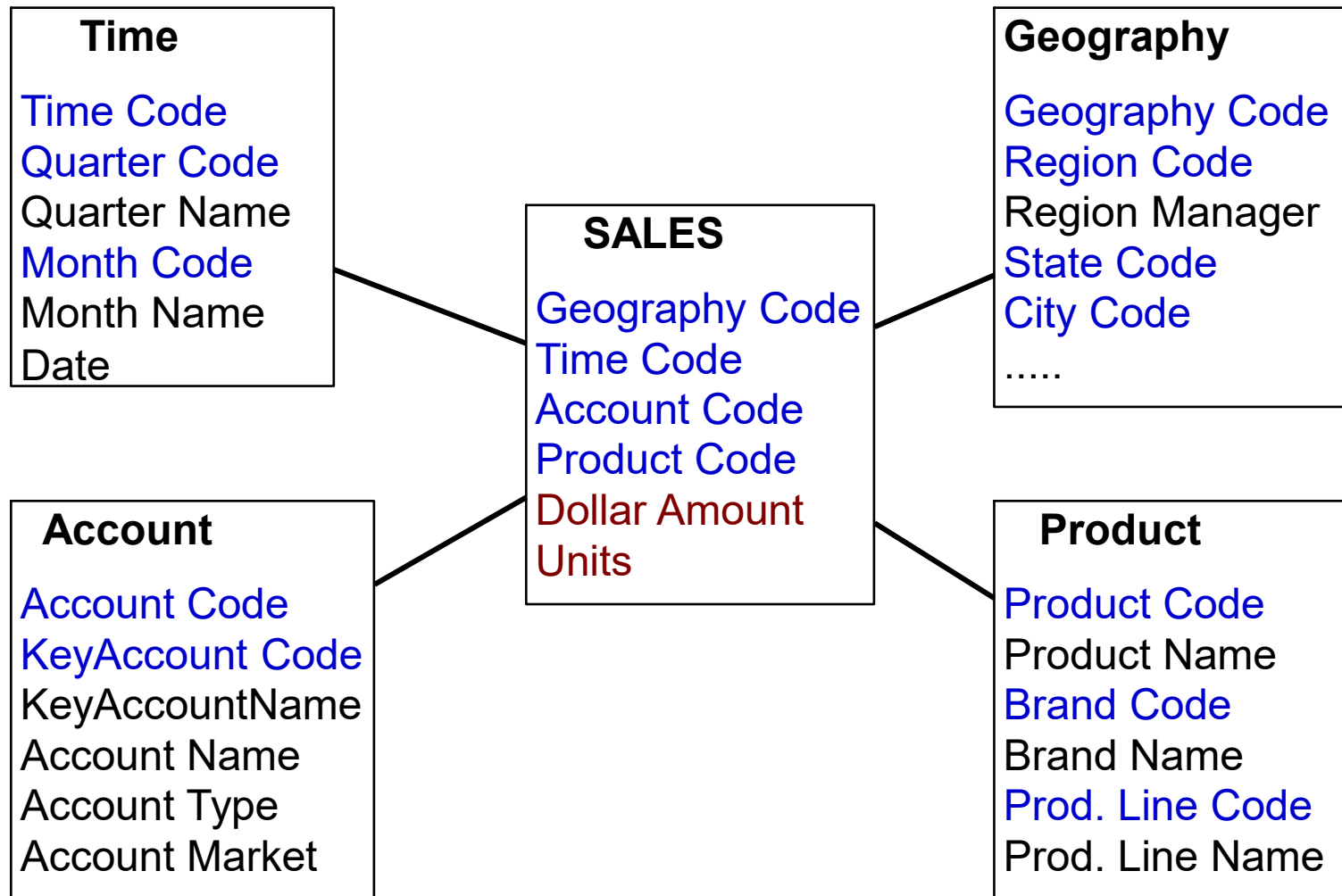
Το λογικό σχήμα της Αποθήκης Δεδομένων

- Υπάρχουν 2 βασικές κατηγορίες σχημάτων για Αποθήκες Δεδομένων:
 - ✦ Σχήμα **Αστέρα** (Star Schema)
 - ✦ Σχήμα **Χιονονιφάδας** (Snowflake Schema)
- Επιπλέον τεχνικές
 - ✦ **Αστερισμοί σχημάτων** (Fact Constellation)
 - ✦ Διαχείριση **Συναθροισμένων Δεδομένων** (Aggregated data)

Σχήμα Αστέρα - Star Schema

- Ένα σχήμα αστέρα αποτελείται από ένα κεντρικό **πίνακα δεδομένων** (**fact** table) και κάποιους αποκανονικοποιημένους **πίνακες διαστάσεων** (**dimension** tables).
- Τα **μέτρα** (**measures**) είναι τα ενδιαφέροντα μεγέθη υπό μέτρηση (π.χ., **Dollar Amount**, **Units** στον πίνακα **SALES**).
- Για κάθε διάσταση του μοντέλου, εισάγουμε και ένα πίνακα (π.χ., **Geography**, **Product**, **Time**, **Account**), ο οποίος περιέχει όλα τα **επίπεδα συνάθροισης** (**levels** of aggregation) καθώς τις σχετικές τους ιδιότητες.

Star Schema



Star Schema

- Αποκανονικοποίηση (Denormalization): παραβίαση των γνωστών 4 κανονικών μορφών => επανάληψη της ίδιας πληροφορίας
- Είναι πρόβλημα? Σε εφαρμογές OLAP είναι αναγκαίο κακό ...
- **ΑΠΑΓΟΡΕΥΕΤΑΙ ΣΕ OLTP ΣΥΣΤΗΜΑΤΑ!!!**

Code	Quarter	Quarter Name	Month	Month Name	Date	Day Name
1	1	1ο Τέταρτο	1	Ιαν.	1/1/02	Δευτέρα
2	1	1ο Τέταρτο	1	Ιαν.	2/1/02	Τρίτη
3	1	1ο Τέταρτο	1	Ιαν.	3/1/02	Τετάρτη

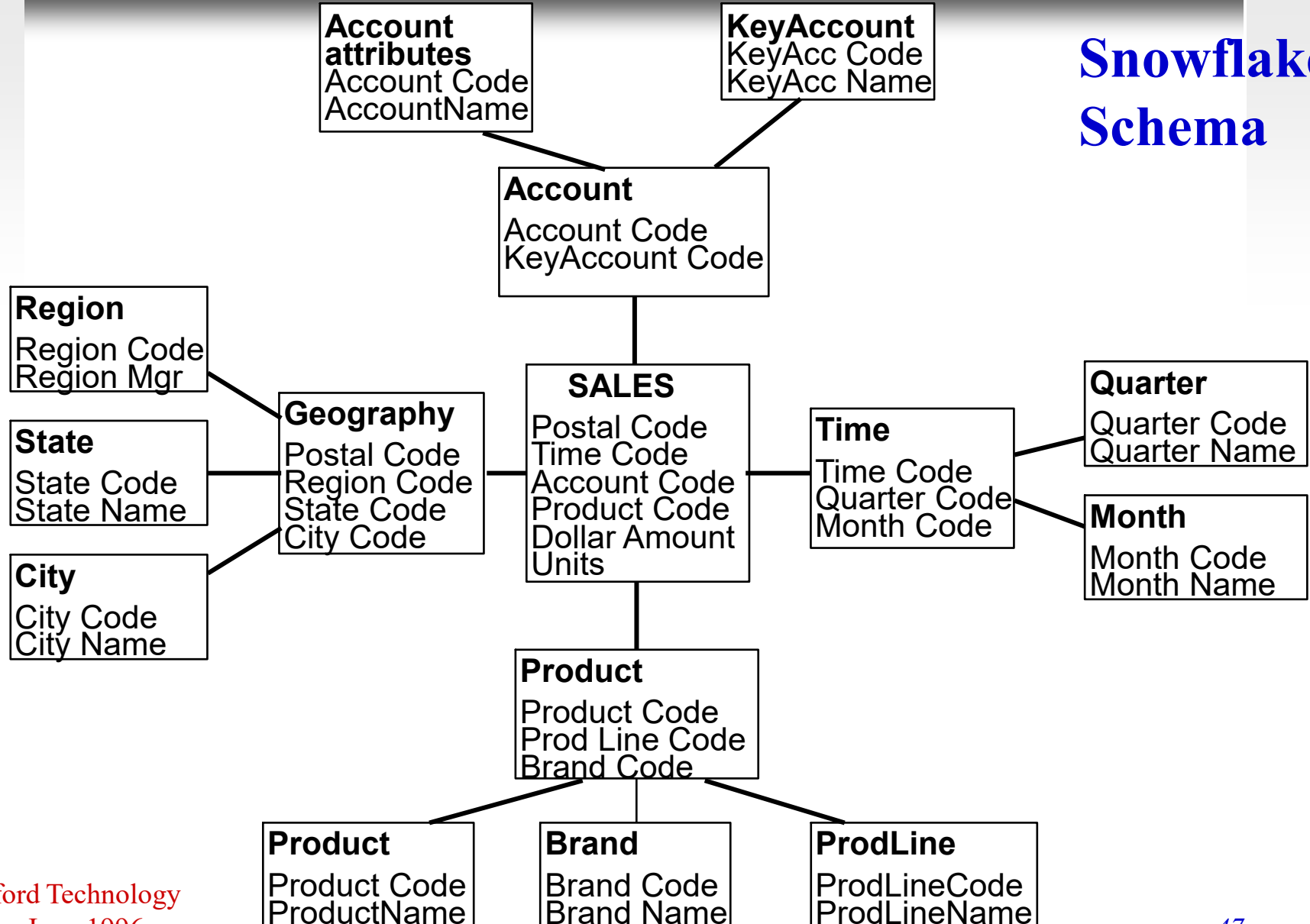
Star Schema

- Απευθείας υλοποίηση του μοντέλου διαστάσεων
- Προσοχή στα κλειδιά! (βλ. και παρακάτω)

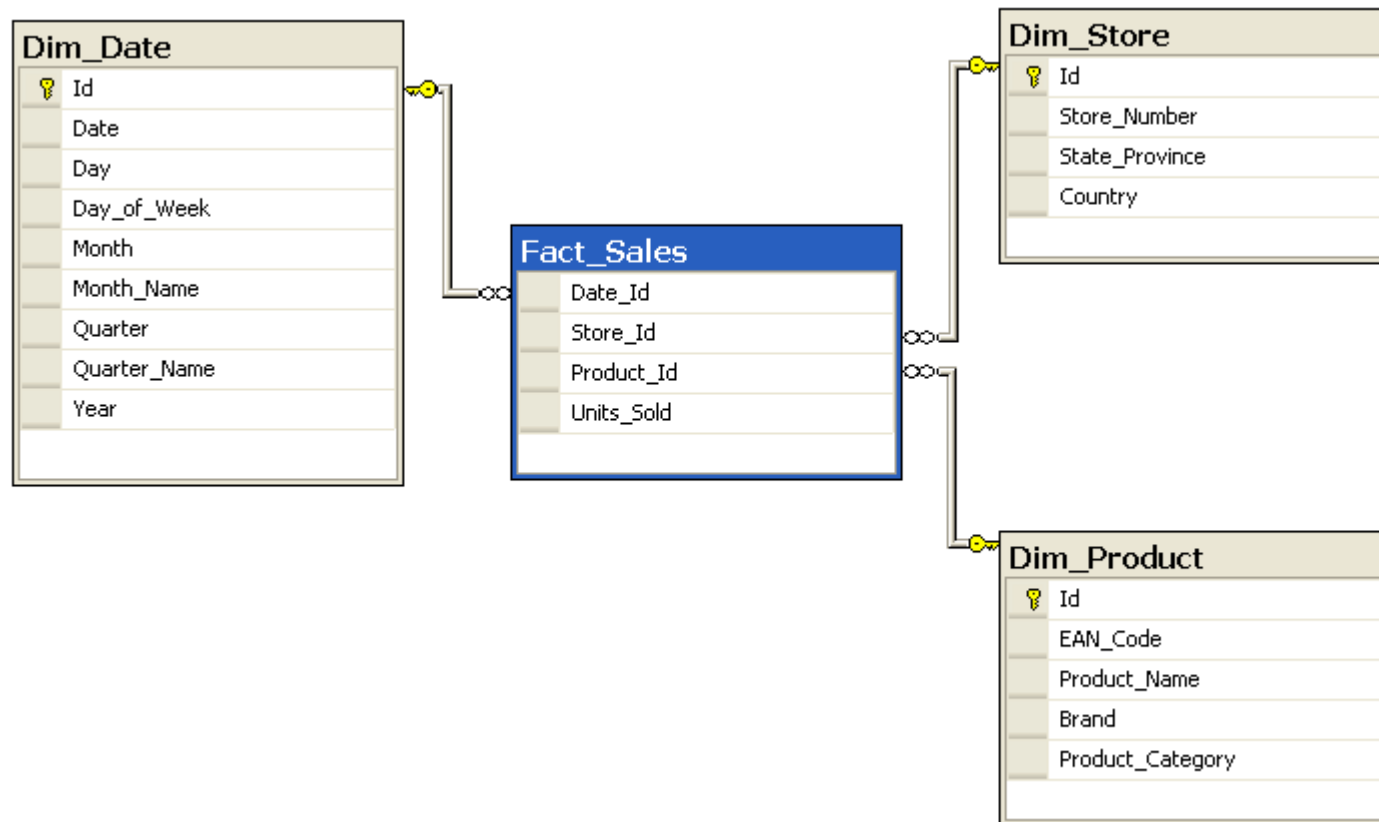
Σχήμα Χιονονιφάδας - Snowflake Schema

- Η κάπως πιο κανονικοποιημένη εκδοχή του σχήματος αστέρα
 - Ο πίνακας γεγονότων έχει ξένα κλειδιά σε πίνακες διαστάσεων πάλι
 - Οι πίνακες διαστάσεων έχουν μόνο ξένα κλειδιά σε πίνακες επιπέδων
 - Για κάθε επίπεδο της ιεραρχίας των διαστάσεων εισάγουμε και ένα δικό του πίνακα
- Πιο τυπικό, εγγυάται την ακεραιότητα των δεδομένων (όπως όλα τα κανονικοποιημένα σχήματα), αλλά πιο αργό στις απαντήσεις των ερωτήσεων

Snowflake Schema

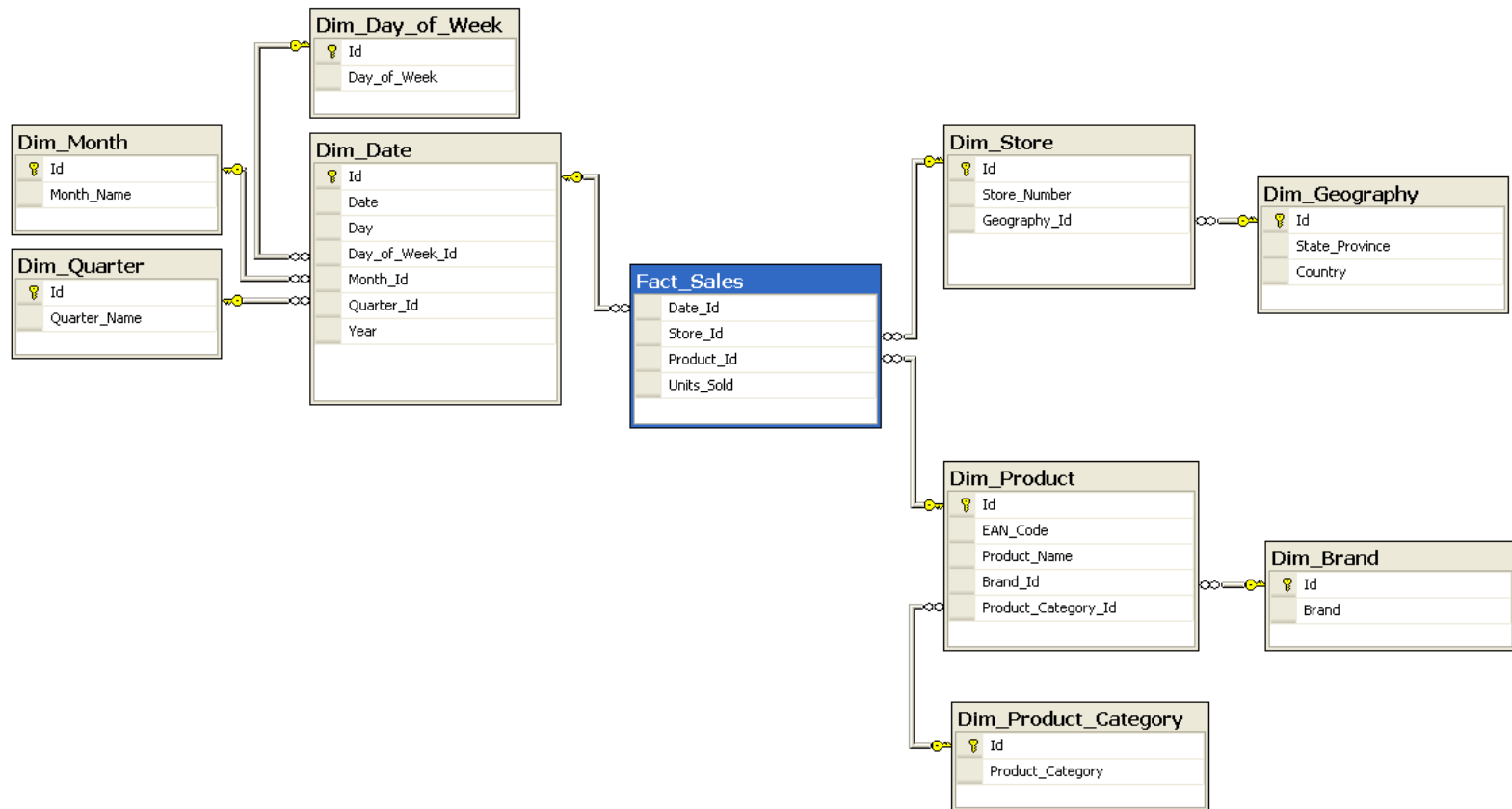


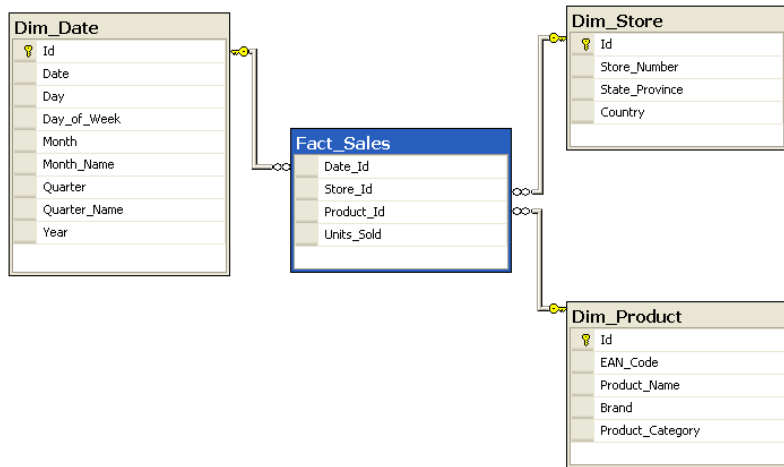
https://en.wikipedia.org/wiki/Star_schema



Another example to compare the two kinds of schemata

https://en.wikipedia.org/wiki/Snowflake_schema

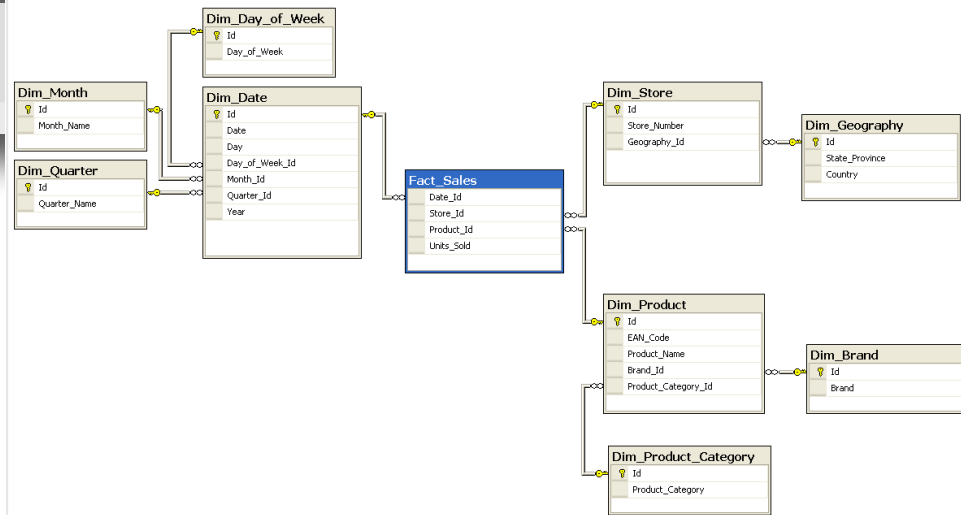




```

SELECT P.Brand, S.Country,
SUM(F.Units_Sold)
FROM Fact_Sales F
    INNER JOIN Dim_Date D ON (F.Date_Id =
        D.Id)
    INNER JOIN Dim_Store S ON (F.Store_Id =
        S.Id)
    INNER JOIN Dim_Product P ON
        (F.Product_Id = P.Id)
WHERE D.Year = 1997 AND
        P.Product_Category = 'tv'
GROUP BY P.Brand, S.Country
  
```

The same query over the two alternatives



```

SELECT B.Brand, G.Country,
SUM(F.Units_Sold)
FROM Fact_Sales F
    INNER JOIN Dim_Date D ON F.Date_Id =
        D.Id
    INNER JOIN Dim_Store S ON F.Store_Id =
        S.Id
    INNER JOIN Dim_Geography G ON
        S.Geography_Id = G.Id
    INNER JOIN Dim_Product P ON
        F.Product_Id = P.Id
    INNER JOIN Dim_Brand B ON P.Brand_Id
        = B.Id
    INNER JOIN Dim_Product_Category C ON
        P.Product_Category_Id = C.Id
WHERE D.Year = 1997 AND
        C.Product_Category = 'tv'
GROUP BY B.Brand, G.Country
  
```

Snowflake schema

- ✦ Υπάρχει και η ακόμα πιο κανονικοποιημένη εκδοχή, στην οποία ο πίνακας γεγονότων δείχνει απ' ευθείας στον πίνακα του πιο χαμηλού επιπέδου της ιεραρχίας
 - ✦ Λείπει, δηλ., ο συγκεντρωτικός πίνακας κάθε διάστασης με τα ξένα κλειδιά στους πίνακες των επιπέδων
- ✦ Αυτή η σχεδίαση είναι 3NF

Αστερισμός Αστέρων - Fact Constellation

- Όταν χρειάζεται να έχω πολλούς fact tables οι οποίοι να μοιράζονται πίνακες διαστάσεων (dimension tables)
 - ✦ Συχνό φαινόμενο στα DW, πιο σπάνιο στα data marts

Συναθροίσεις Δεδομένων - Aggregations

- ✦ Εκτός από τις λεπτομερείς πληροφορίες των fact tables, μπορεί να υπολογίσουμε και συναθροίσεις των δεδομένων για καλύτερους χρόνους απόκρισης.
- ✦ Για παράδειγμα, αν ο fact table είναι

**SALES (GeographyCode, ProductCode, TimeCode,
AccountCode, Amount, Unit)**

μπορούμε να υπολογίσουμε

- ✦ **AVG(Sales)** ανά Region, Product, Quarter
- ✦ **MAX(Sales)** ανά Brand, Month, με Region = Europe
- ✦ **SUM(Sales)** ανά City

Συναθροίσεις Δεδομένων - Τεχνικές

- ✦ Ενσωμάτωση των αθροιστικών εγγραφών στον βασικό (base / basic) fact table + μια **επιπλέον στήλη** που να εξηγεί το επίπεδο συνάθροισης
 - ✦ **ΠΑΡΑΞΕΝΟ MODELING!** (με κάποια πλεονεκτήματα, πάντως)
- ✦ Χωριστός πίνακας
 - ✦ Εσχάτως, τα RDBMS's υποστηρίζουν **υλοποιημένες όψεις** (materialized views)

Γνωστό ερευνητικό πρόβλημα: ποιες από όλες τις πιθανές συναθροίσεις πρέπει να υλοποιήσω?

Συναθροίσεις Δεδομένων

- Χωριστός πίνακας/όψη αθροισμάτων

Sales table

RID	City	...	Amount
1	Athens	...	\$100
2	N.Y.	...	\$300
3	Rome	...	\$120
4	Athens	...	\$250
5	Rome	...	\$180
6	Rome	...	\$65
7	N.Y.	...	\$450

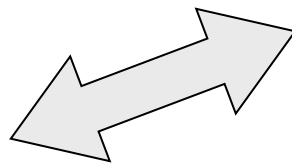
**City-dimension
sum table**

City	Amount
Athens	\$350
N.Y.	\$750
Rome	\$365

- Επέκταση του υπάρχοντος βασικού πίνακα

Extended Sales table

RID	City	...	Amount	Level
1	Athens	...	\$100	NULL
2	N.Y.	...	\$300	NULL
3	Rome	...	\$120	NULL
4	Athens	...	\$250	NULL
5	Rome	...	\$180	NULL
6	Rome	...	\$65	NULL
7	N.Y.	...	\$450	NULL
8	Athens	...	\$350	City
9	N.Y.	...	\$750	City
10	Rome	...	\$365	City



Σχεδίαση της Αποθήκης Δεδομένων

✦ Top – down

- ✦ Κατασκευή ενός **κεντρικού μοντέλου αναφοράς** (enterprise model)
- ✦ Περιγραφή των πηγών (ενίοτε και των data marts) σε σχέση με το κεντρικό μοντέλο αναφοράς
 - Χρονοβόρα μέθοδος, απαιτητική σε ανθρωπομήνες, **καθυστέρηση στην παρουσίαση των πρώτων αποτελεσμάτων**
 - Αυξάνει το ρίσκο του έργου, λόγω των παραπάνω
 - + Παρέχει μία **συνεπή συνολική εικόνα** των δεδομένων (και των επί μέρους συστημάτων του οργανισμού)

Σχεδίαση της Αποθήκης Δεδομένων

✦ Bottom – up

- ✦ Αρχικά, κατασκευή μικρότερων data marts
- ✦ Προοδευτικός συνδυασμός των κατασκευαζόμενων data marts
 - Πολύ δύσκολα θα καταλήξει σε μια συνολική εικόνα του συστήματος
 - Το ρίσκο βρίσκεται στην πιθανή αδυναμία ενοποίησης αργά στο έργο
- + Γρήγορη παρουσίαση αποτελεσμάτων
- + Πιο φτηνό σε χρόνο και πόρους

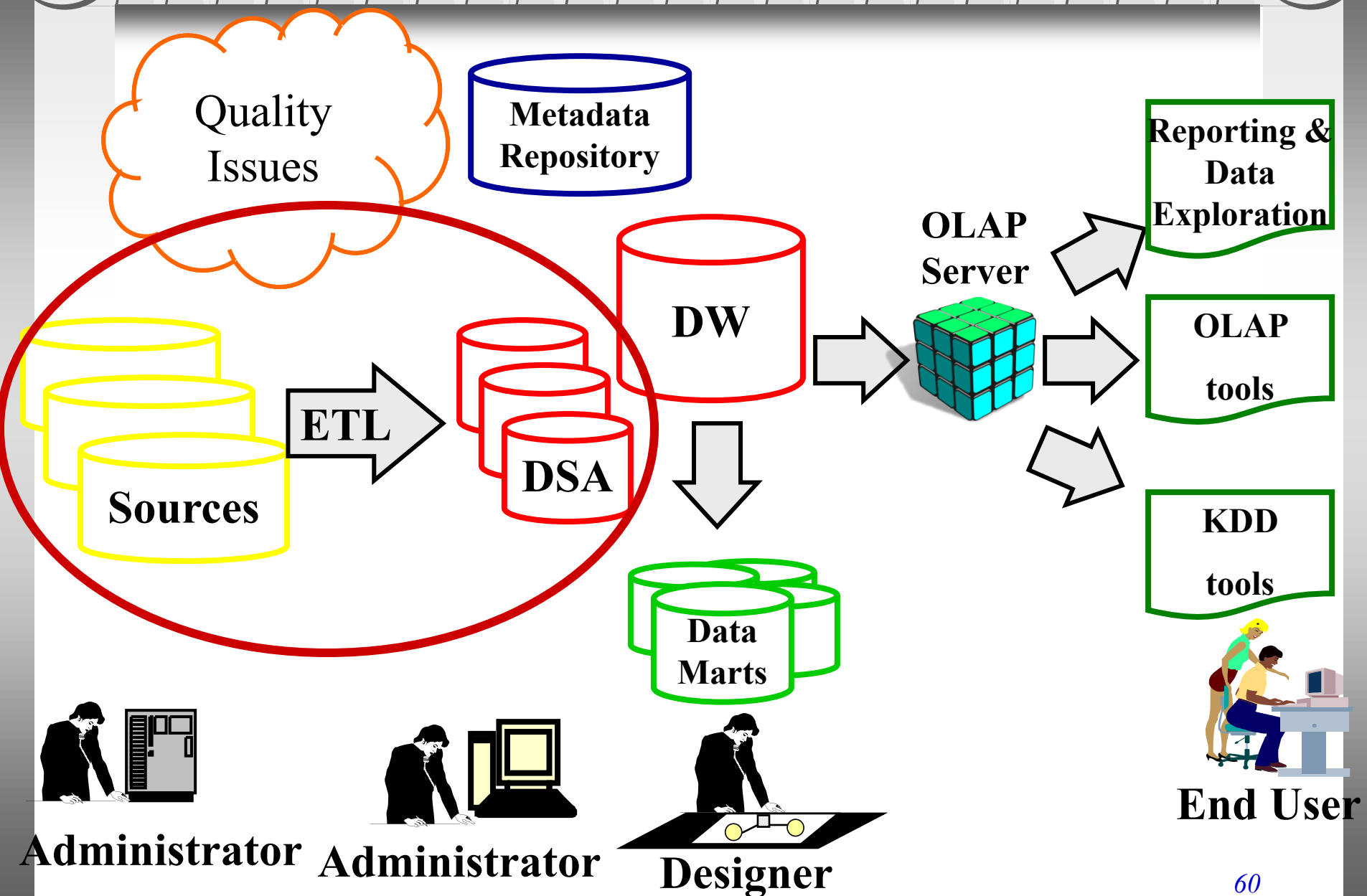
Βήματα ομογενοποίησης σχημάτων & δεδομένων

- **Επιλογή (Preintegration).** Ποια σχήματα θα «ενοποιήσω» και με ποια σειρά
- **Σύγκριση (Schema Comparison).** Ποιες οι συσχετίσεις ανάμεσα στα διαφορετικά σχήματα και ποιες οι πιθανές συγκρούσεις (π.χ., μεταξύ ονομάτων πεδίων, λογικών κανόνων, τύπων πεδίων ...)
- **Επίλυση (Schema Conforming).** Εξεύρεση κανόνων για την επίλυση των συγκρούσεων σε επίπεδο πεδίων και δεδομένων
- **Υλοποίηση (Schema Merging and Restructuring).** Υλοποίηση των εξευρεθέντων λύσεων και παραγωγή ενός ενιαίου σχήματος

Υλοποίηση σε φυσικό επίπεδο

- Επιλογή του κατάλληλου εξοπλισμού σε μηχανήματα, συστήματα Βάσεων Δεδομένων και εργαλείων λογισμικού.
- Εγκατάσταση επικοινωνίας μεταξύ των servers και των εργαλείων ανάλυσης
- Δημιουργία της φυσικής οργάνωσης της Αποθήκης Δεδομένων, υλοποίηση των σχετικών δομών και των μεθόδων πρόσβασης στην Αποθήκη.
- Εγκατάσταση των λειτουργικών και σύνδεση με τις πηγές δεδομένων.

Ενημέρωση της Αποθήκης Δεδομένων



Λειτουργικές Διεργασίες της Αποθήκης Δεδομένων

- Ενημέρωση = **εξαγωγή** + **μεταφορά** + **μετασχηματισμός** + **καθαρισμός** + **φόρτωση** των δεδομένων (data extraction, transform & load – ETL) από τις πηγές στην Αποθήκη Δεδομένων.
- Υπολογισμός συναθροίσεων και αποθηκευμένων όψεων
- Εξυπηρέτηση **επερωτήσεων**
- Επίβλεψη (**monitoring**) της λειτουργία της Αποθήκης Δεδομένων.

Ενημέρωση της Αποθήκης Δεδομένων

- Η ενημέρωση της Αποθήκης Δεδομένων μεταφέρει τις αλλαγές που συμβαίνουν στα δεδομένα των πηγών εκτελώντας αντίστοιχες αλλαγές στα δεδομένα της Αποθήκης.
- Η διαδικασία αυτή ακολουθεί όλα τα βήματα (εξαγωγή, μεταφορά, μετασχηματισμός, καθαρισμός και εισαγωγή).
- Η πολιτική ενημέρωσης καθορίζεται από το διαχειριστή της Αποθήκης Δεδομένων με βάση τις ανάγκες των εφαρμογών ανάλυσης, τη διαθεσιμότητα των πηγών και τη κατάσταση του δικτύου που συνδέει την Αποθήκη με τις πηγές.

Ενημέρωση της Αποθήκης Δεδομένων

- **Κάθε πότε** στέλνω τις αλλαγές που συνέβησαν στις πηγές, στην Αποθήκη Δεδομένων ?
 - ✦ σε κάθε αλλαγή (on update)?
 - ✦ σε κάθε επερώτηση προς την Αποθήκη Δεδομένων (on demand)?
 - ✦ περιοδικά (**batch**)? [default]
- **Χρονικός περιορισμός**: η ενημέρωση έχει πάντα περιορισμένο χρονικό όριο (**time window**) για να ολοκληρωθεί

Εξαγωγή των δεδομένων

- **Στόχος:** να βρω ποιες ήταν οι αλλαγές στις πηγές, ήτοι ποιες εγγραφές εισήχθησαν, διεγράφησαν ή ανανεώθηκαν (new/deleted/updated)
- **Βασικές τεχνικές:** Είτε εξάγω πλήρη στιγμιότυπα (**snapshot**) ή διαφορικά (**differentials**) των πηγαίων δεδομένων
- **Περιορισμοί:**
 - Το διαθέσιμο χρονικό παράθυρο είναι πεπερασμένο
 - Ελαχιστοποίηση της λειτουργικής επιβάρυνσης των OLTP συστημάτων των πηγών
 - Ελαχιστοποίηση των αλλαγών στο software configuration των OLTP συστημάτων των πηγών

Εξαγωγή των δεδομένων

- Πού και πώς συγκρίνω τα **πλήρη στιγμιότυπα** (snapshots)?
 - Στις πηγές ή στην DSA?
 - Διαφορετικές τεχνικές: συγκρίνοντας τμηματικά τα αρχεία, ή με κάποιου είδος hashing, ...
- **Διαφορικές τεχνικές** εντοπισμού των αλλαγών:
 - Σύγκριση των log files, όταν είναι διαθέσιμα [γρήγορο]
 - Αλλαγή των data entry προγραμμάτων [επικίνδυνο και δαπανηρό]
 - Τεχνικές triggering [σπάνιο]
- Συνήθως η όλη διαδικασία γίνεται με **αρχεία** (γρήγορο) μέσω εργαλείων **ETL** ή **scripts** (φτηνό, εύκολο?) ...

Μεταφορά των δεδομένων

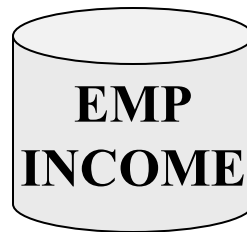
- Λόγω δικτυακής ταχύτητας / σταθερότητας / ασφάλειας, ενίοτε χρησιμοποιούνται τεχνικές συμπίεσης και κρυπτογράφησης

Μετασχηματισμός των δεδομένων

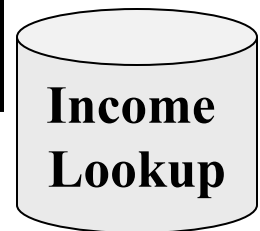
- **Αλλαγή σχήματος:** πιθανότατα το σχήμα της Αποθήκης Δεδομένων είναι διαφορετικό από το σχήμα της πηγής
- **Αλλαγή / Υπολογισμός τιμής:** είναι πιθανόν οι εγγραφές στην πηγή να έχουν άλλο format, τιμή, ..., απ' ότι στην Αποθήκη Δεδομένων, ή να χρειαστεί να υπολογισθούν νέες τιμές
 - Integer -> Real
 - Euro -> Dollar
 - Date of Birth -> Age

Αποκανονικοποίηση - Denormalization

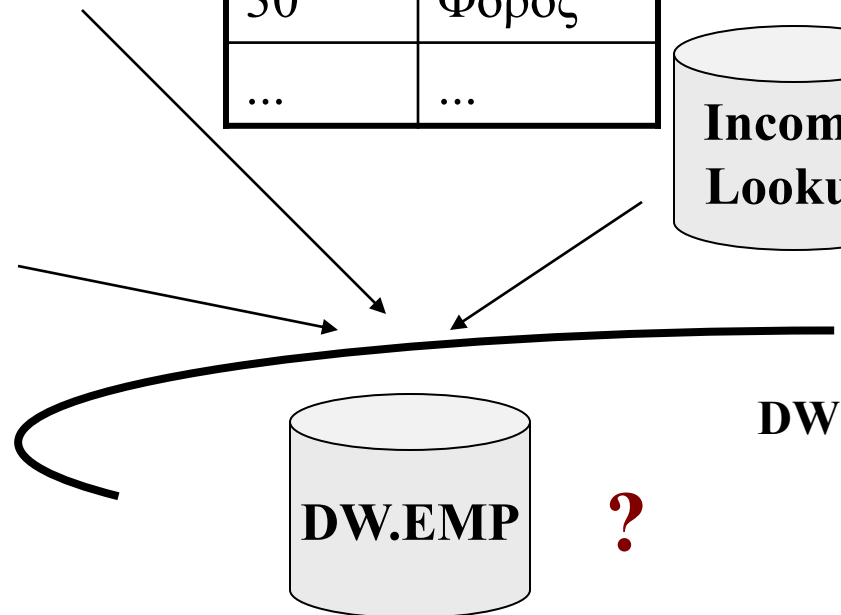
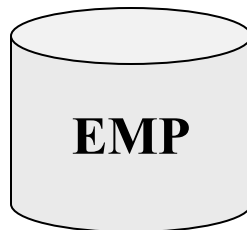
<u>EMP ID</u>	<u>IL ID</u>	Amount
110	10	1500
110	30	300



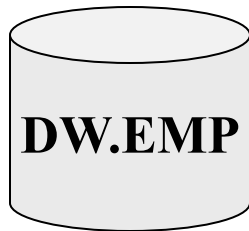
<u>IL ID</u>	Descr
10	Μισθός
20	Επίδομα Τέκνων
30	Φόρος
...	...



<u>EMP ID</u>	Name	Age
110	Kostas	30
120	Mitsos	48
130	Roula	29



Αποκανονικοποίηση - Denormalization



<u>EMP ID</u>	Name	Age	Salary	Tax	Child Bonus
110	Kostas	30	1500	300	NULL
...	...				...

- **Κακή σχεδίαση για OLTP:** αν προστεθεί νέα κατηγορία χρηματικού ποσού => νέα κολώνα στον πίνακα
- **Αποδεκτή για DW:** για μια απλή ερώτηση, γλιτώνουμε 2 joins ...
- Πιθανόν να χρειαστεί και το αντίστροφο (**κανονικοποίηση** - normalization)...

Καθαρισμός των δεδομένων

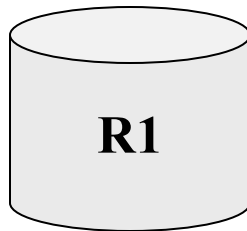
- **Δεδομένα που παραβιάζουν τους κανόνες της βάσης:**
διπλοεγγραφές, παραβιάσεις πρωτεύοντος ή ξένου κλειδιού, τιμές εκτός ορίων, παραβιάσεις λογικών κανόνων, ...
- **Συνώνυμα και συγκρούσεις**
- **Ελλιπή δεδομένα**
- **Καλλωπισμός (Renicing):** στα strings, π.χ., διευθύνσεις

Καθαρισμός των δεδομένων -- Παραβίαση λογικών κανόνων

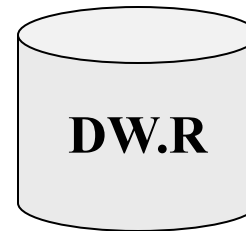
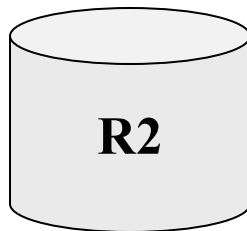
✦ **IF (SEX='F' AND ILLNESS='PROSTATE')
THEN (ALERT ERROR MESSAGE) !**

Καθαρισμός των δεδομένων – Ομογενοποίηση κλειδιών

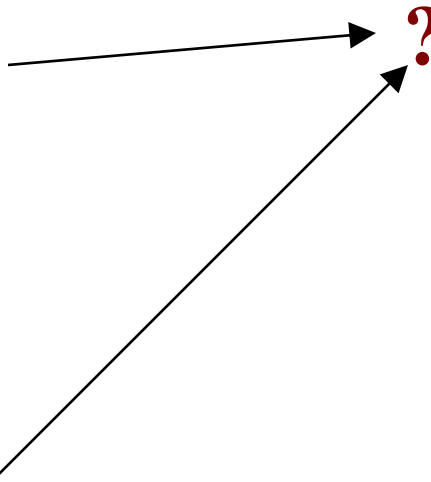
<u>ID</u>	Descr
10	Coca
20	Pepsi



<u>ID</u>	Descr
10	Pepsi
20	HBH

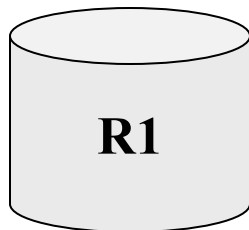


<u>ID</u>	Descr
??	??
??	??

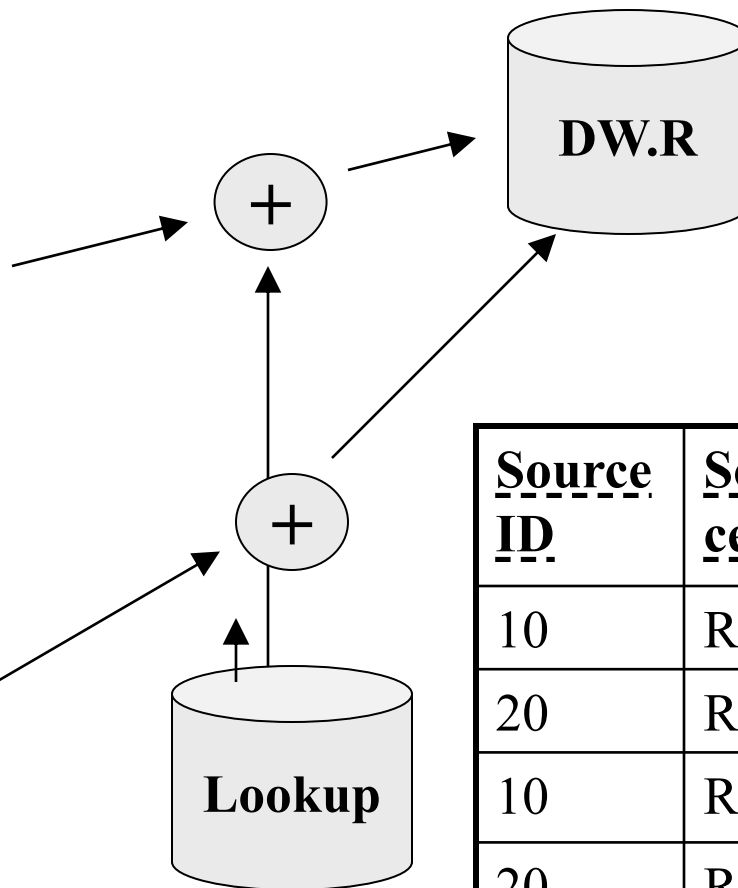
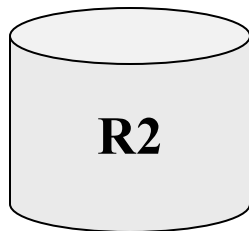


Καθαρισμός των δεδομένων – Ομογενοποίηση κλειδιών

<u>ID</u>	Descr
10	Coca
20	Pepsi



<u>ID</u>	Descr
10	Pepsi
20	HBH



<u>ID</u>	Descr
100	Coca
110	Pepsi
120	HBH

<u>Source ID</u>	<u>Source</u>	<u>Surrogate Key</u>
10	R1	100
20	R1	110
10	R2	110
20	R2	120

Καθαρισμός των δεδομένων – Ομογενοποίηση κλειδιών

- Τα κλειδιά των πηγών συνήθως ονομάζονται **κλειδιά παραγωγής (production keys)** και τα νέα, ομογενοποιημένα κλειδιά, **κλειδιά αντικαταστάτες (surrogate keys)**
- Υπάρχουν ειδικές τεχνικές για το πώς αλλάζουν τα κλειδιά στην Αποθήκη Δεδομένων, αν αλλάξει ένα κλειδί σε μια πηγή...

Καθαρισμός των δεδομένων -- Πίνακας Συνωνύμων

Source Value	DW value
HP	HP
H.P.	HP
H-P	HP
Hewlett-Packard	HP
Hioulet-Pakard	HP
DEC	DEC
Digital Co.	DEC
...	...

Πολύ συχνό για
διευθύνσεις (π.χ.,
«Νομός Θεσ/νίκης»)
και για ονόματα
(π.χ., «κ. Κώστας
Ιωαννου»)

Πίνακας συνωνύμων
στην καλή
περίπτωση, εν γένει
χρειαζόμαστε
regular expressions
(π.χ., με Perl)!!!

Φόρτωση των δεδομένων

✦ Τα ζητήματα:

- ✦ Τεράστιος όγκος δεδομένων σε μικρό παράθυρο χρόνου
- ✦ Εκτός από τους κανονικούς πίνακες, υπάρχουν και indexes και συναθροίσεις που πρέπει να ενημερωθούν
- ✦ Σε περίπτωση αποτυχίας, η ακεραιότητα των δεδομένων πρέπει να διατηρηθεί

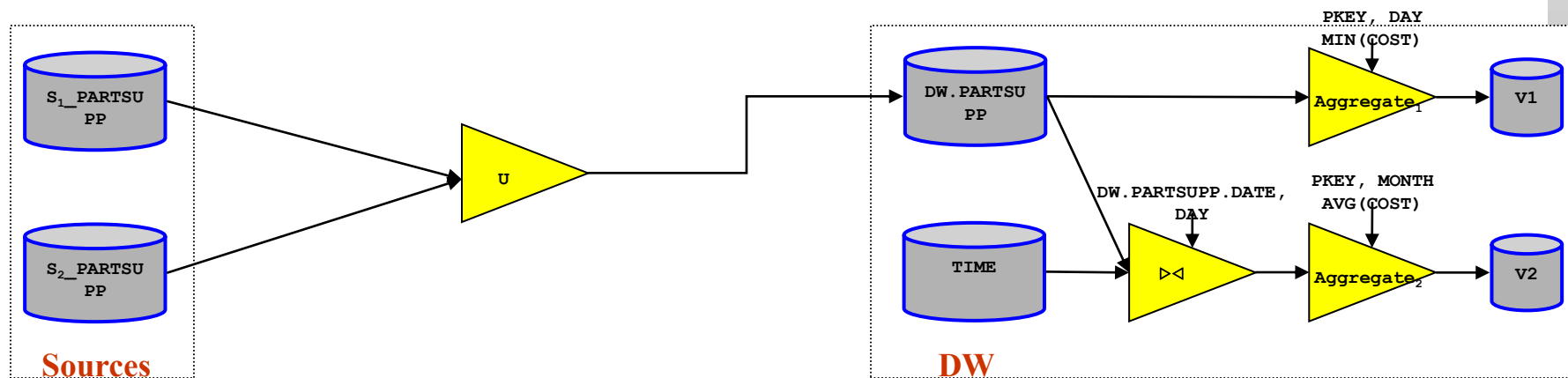
Τεχνικές φόρτωσης των δεδομένων

- Η ενημέρωση / εισαγωγή δεδομένων ΔΕΝ γίνεται μέσω SQL πρακτικά:
 - ✦ Record-at-a-time είναι αργό
 - ✦ Ακόμα πιο αργό λόγω random disc I/O
 - ✦ Μπορεί να κάνει το rollback segment ή το log file να γεμίσει => η διεργασία γίνεται zombie
- Μερική ρύθμιση: συνήθως απενεργοποιούμε τα logging & locking
 - ✦ Επικίνδυνο σε περίπτωση αποτυχίας κατά τη φόρτωση

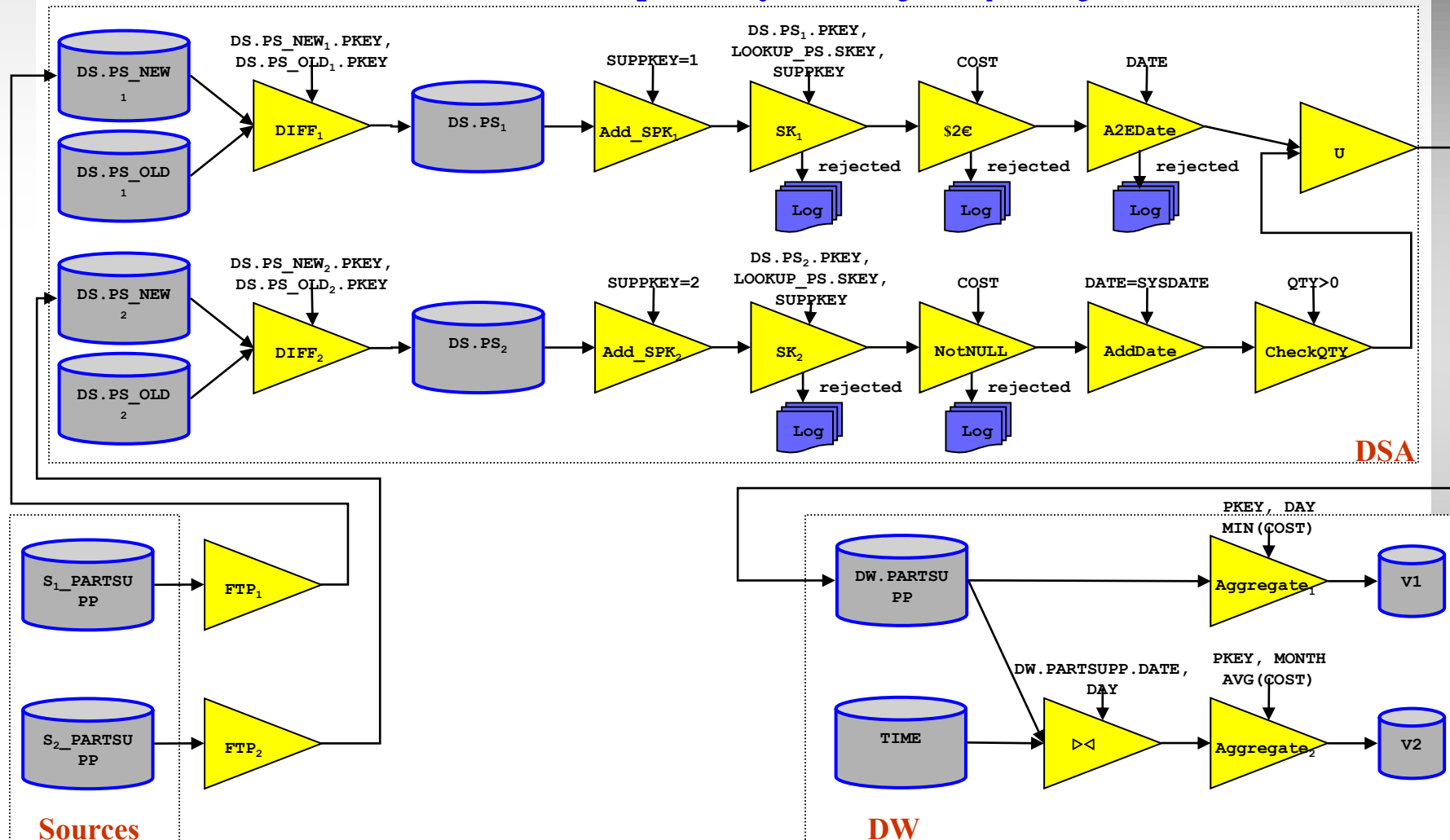
Τεχνικές φόρτωσης των δεδομένων

- Η καλύτερη λύση: τα εργαλεία **batch loading** που προσφέρουν όλα τα RDBMS's
- Επιπλέον τεχνικές:
 - ✦ Ταξινόμηση των εγγραφών με βάση κάποιο clustering key (ιδίως αν έχουμε τον αντίστοιχο πίνακα clustered)
 - ✦ Το sequential I/O είναι 100 φορές γρηγορότερο του random I/O
 - ✦ Κατασκευή των indexes σε αντιστοιχία με τους πίνακες
 - ✦ Όσο πιο παράλληλο, τόσο πιο γρήγορο!

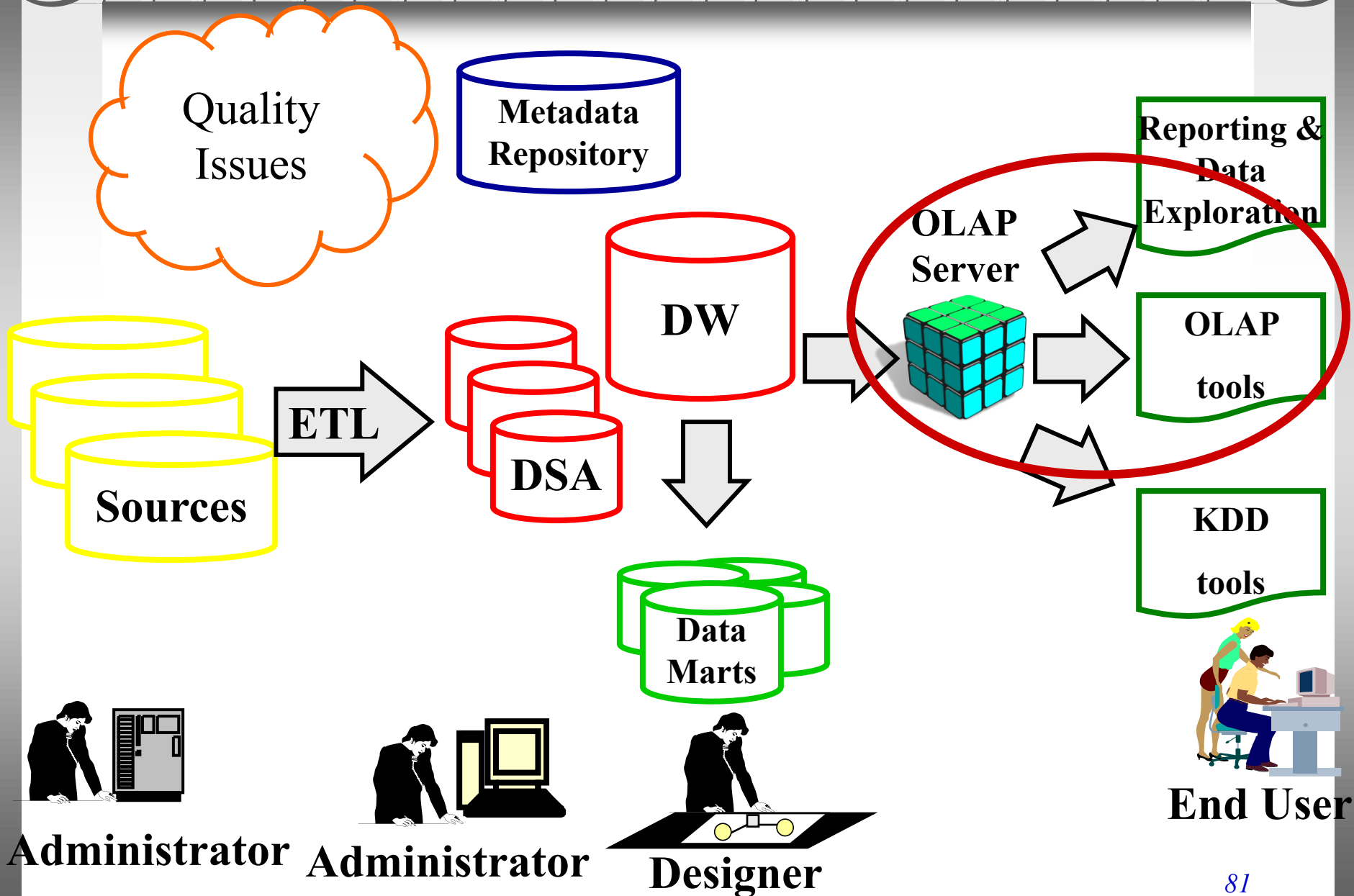
DW = Αποθηκευμένες όψεις?



DW ≠ Αποθηκευμένες όψεις!

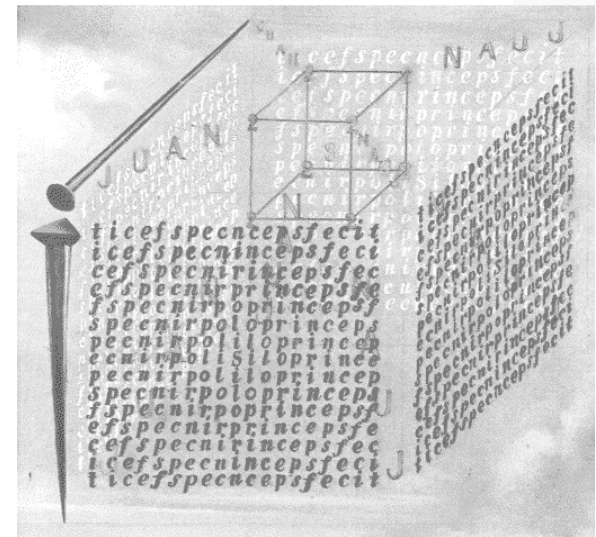


On Line Analytical Processing (OLAP)



Εργαλεία για υποστήριξη αποφάσεων

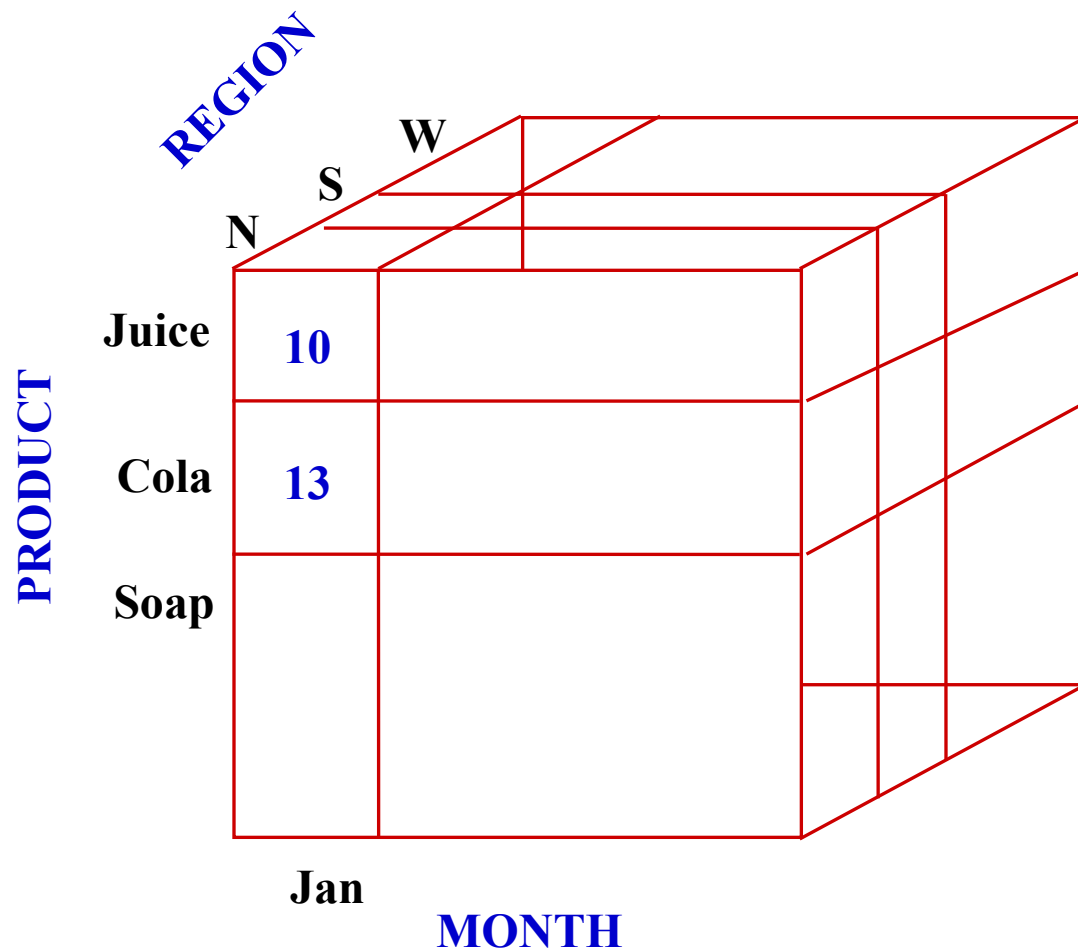
- Ad hoc ερωτήσεις και αναφορές
 - Π.χ.,: MS Excel, Oracle Forms, ...
- **OLAP**
 - pivot tables, drill down, roll up, slice, dice
- Data Mining



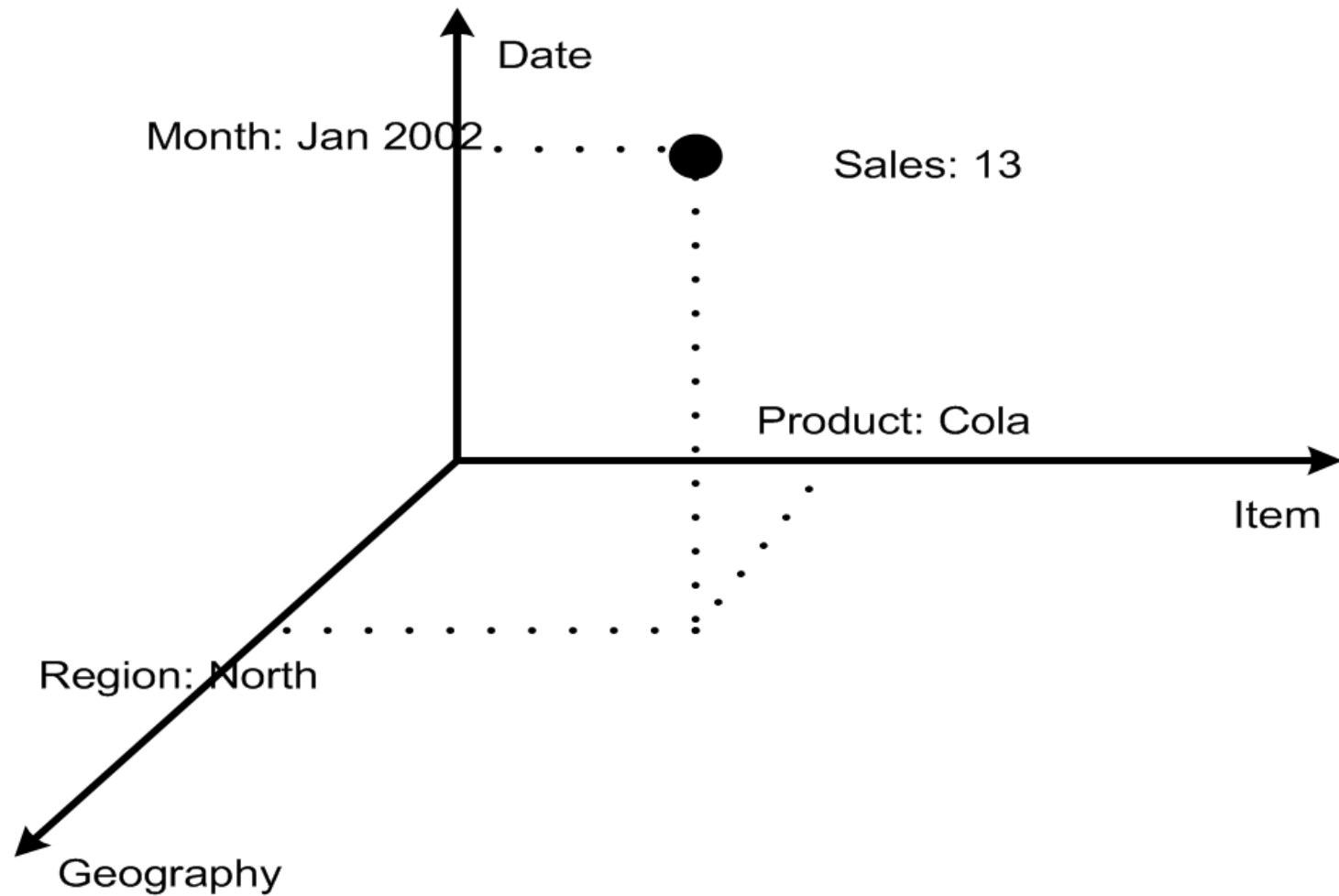
Βασικές Έννοιες OLAP

- Αφορά την ανάλυση κάποιων μετρήσιμων μεγεθών (μέτρων)
 - πωλήσεις, απόθεμα, κέρδος,...
- Διαστάσεις: παράμετροι που καθορίζουν το περιβάλλον (context) των μέτρων
 - ημερομηνία, προϊόν, τοποθεσία, πωλητής, ...
- Κύβοι: συνδυασμοί διαστάσεων που καθορίζουν κάποια μέτρα
 - Ο κύβος καθορίζει ένα πολυδιάστατο χώρο διαστάσεων, με τα μέτρα να είναι σημεία του χώρου αυτού

Κύβοι για OLAP



Κύβοι για OLAP



Βασικές Έννοιες OLAP

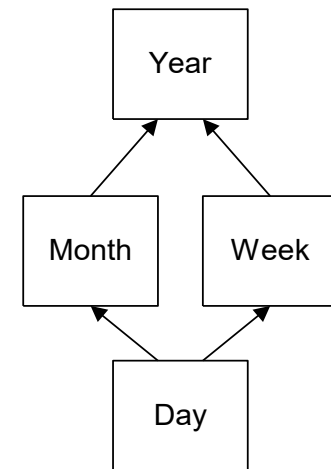
- Τα δεδομένα θεωρούνται αποθηκευμένα σε ένα **πολυδιάστατο πίνακα** (multi-dimensional array), ο οποίος αποκαλείται και **κύβος** ή **υπερκύβος** (**Cube** και **HyperCube** αντίστοιχα).
- Ο κύβος είναι μια ομάδα από **κελιά** δεδομένων (data cells). Κάθε κελί χαρακτηρίζεται μονοσήμαντα από τις αντίστοιχες τιμές των **διαστάσεων** (dimensions) του κύβου.
- Τα περιεχόμενα του κελιού ονομάζονται **μέτρα** (measures) και αναπαριστούν τις αποτιμώμενες αξίες του πραγματικού κόσμου.

Ιεραρχίες επιπέδων για OLAP

- Μια **διάσταση** μοντελοποιεί όλους τους τρόπους με τους οποίους τα δεδομένα μπορούν να συναθροιστούν σε σχέση με μια συγκεκριμένη παράμετρο του περιεχομένου τους.
 - ✦ **Ημερομηνία, Προϊόν, Τοποθεσία, Πωλητής, ...**
- Κάθε διάσταση έχει μια σχετική **ιεραρχία επιπέδων** συνάθροισης των δεδομένων (hierarchy of levels). Αυτό σημαίνει, ότι η διάσταση μπορεί να θεωρηθεί από πολλά επίπεδα αδρομέρειας.
 - ✦ **Ημερομηνία: μέρα, εβδομάδα, μήνας, χρόνος, ...**

Ιεραρχίες Επιπέδων

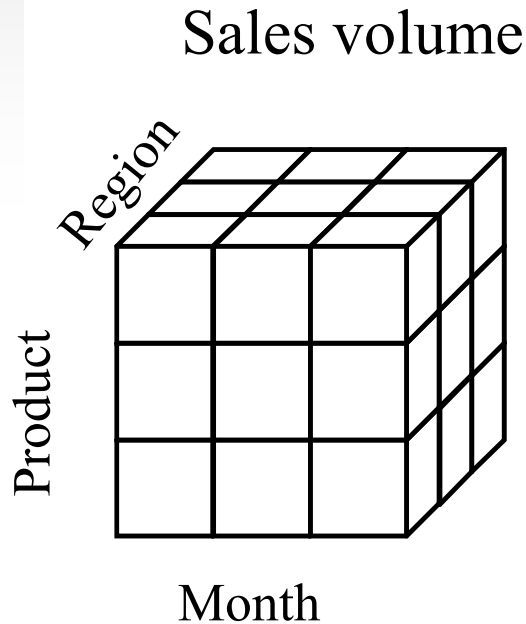
- **Ιεραρχίες Επιπέδων:** κάθε διάσταση οργανώνεται σε διαφορετικά επίπεδα αδρομέρειας
- Ο χρήστης μπορεί να πλοηγηθεί από το ένα επίπεδο στο άλλο, δημιουργώντας νέους κύβους κάθε φορά



Αδρομέρεια: το αντίθετο της λεπτομέρειας

-- ο σωστός όρος είναι αδρομέρεια...

Κύβοι & ιεραρχίες διαστάσεων για OLAP



Διαστάσεις: Product, Region, Date

Ιεραρχίες διαστάσεων:

Industry
|
Category
|
Product

Country
|
Region
|
City
|
Store

Year
|
Quarter
/ \
Month Week
\< /
Day

Εργασίες που κάνει ο χρήστης

- ✦ Συνήθειες πράξεις που κάνουμε σε κύβους
 - ✦ Συναθροίσεις (total sales, percent-to-total)
 - ✦ Συγκρίσεις (budget vs. expense)
 - ✦ Ταξινόμηση - κατάταξη (top 10)
 - ✦ Πρόσβαση σε πιο αναλυτική πληροφορία
 - ✦ Οπτικοποίηση με διαφορετικούς τρόπους

Βασικές (αλγεβρικές) πράξεις OLAP

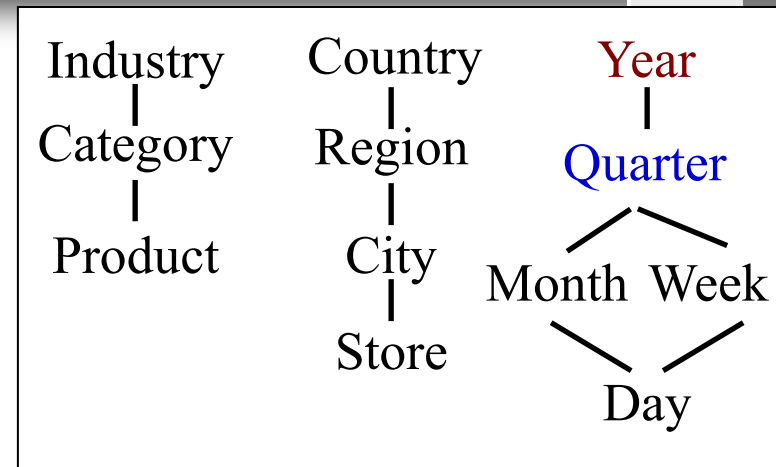
- **Συναθροιστική Άνοδος (Roll up)**: συνάθροιση της πληροφορίας = μετάβαση από χαμηλότερο σε υψηλότερο επίπεδο αδρομέρειας (π.χ. από day σε month)
- **Αναλυτική Κάθοδος (Drill down)**: το αντίστροφο του Roll up
- **Οριζόντιος Τεμαχισμός (Slice)**: (σχεσιακή) επιλογή
- **Κάθετος Τεμαχισμός (Dice)**: (σχεσιακή) προβολή
- **Περιστροφή (Pivot)**: αναδιάταξη της 2D προβολής του πολυδιάστατου κύβου στην οθόνη

Συναθροιστική άνοδος – Roll-up

- Η συναθροιστική άνοδος περιλαμβάνει τον υπολογισμό μίας συνολικής τιμής για μία θέση στην ιεραρχία μίας διάστασης δεδομένων.
- Για παράδειγμα, με ένα roll-up, οι πωλήσεις σε επίπεδο **τοπικών μαγαζιών** παράγουν τις **συνολικές** πωλήσεις σε επίπεδο **πόλης** και αυτές με τη σειρά τους με ένα ακόμα roll-up παράγουν τις πωλήσεις σε επίπεδο **περιοχής**.

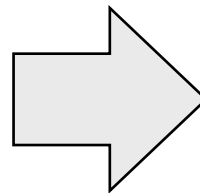
*Εναλλακτικά: consolidation, aggregation
(συνάθροιση)*

Roll up



Sales volume			
	Products	Store1	Store2
Q1	Electronics	\$5,2	\$5,6
	Toys	\$1,9	\$1,4
	Clothing	\$2,3	\$2,6
	Cosmetics	\$1,1	\$1,1
Q2	Electronics	\$8,9	\$7,2
	Toys	\$0,75	\$0,4
	Clothing	\$4,6	\$4,6
	Cosmetics	\$1,5	\$0,5

Χρόνος: Επίπεδο **Quarter**



Sales volume			
	Products	Store1	Store2
Year 1996	Electronics	\$14,1	\$12,8
	Toys	\$2,65	\$1,8
	Clothing	\$6,9	\$7,2
	Cosmetics	\$2,6	\$1,6

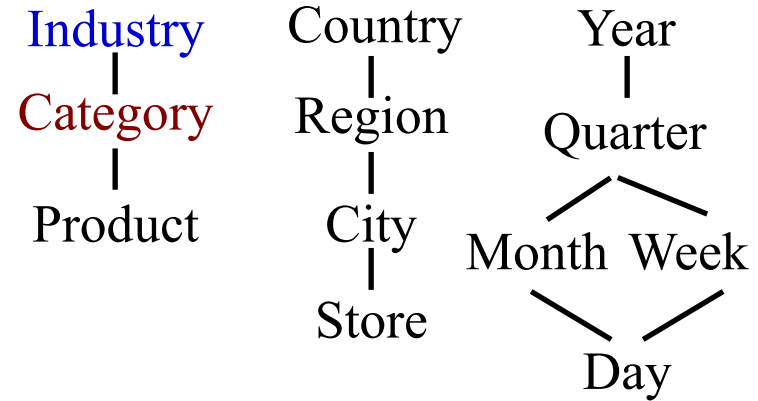
Χρόνος: Επίπεδο **Year**

SUM(Sales volumes)

Αναλυτική κάθοδος – Drill Down

- Ο χρήστης περνά από ένα ανώτερο επίπεδο μίας διάστασης που έχει συγκεντρωτικά δεδομένα σε ένα χαμηλότερο επίπεδο με πιο λεπτομερή δεδομένα. Πρόκειται για την αντίστροφη πράξη του roll-up.
- Για παράδειγμα, κατά το drill down, ξεκινάμε από τις πωλήσεις ανά **περιοχή** και παίρνουμε τις **αναλυτικές** πωλήσεις ανά **πόλη** και μετά τις πωλήσεις ανά **κατάστημα**.

Drill down



Sales volume

	Products	Store1	Store2
Q1	Electronics	\$5,2	\$5,6
	Toys	\$1,9	\$1,4
	Clothing	\$2,3	\$2,6
	Cosmetics	\$1,1	\$1,1
Q2	Electronics	\$8,9	\$7,2
	Toys	\$0,75	\$0,4
	Clothing	\$4,6	\$4,6
	Cosmetics	\$1,5	\$0,5

Item: Επίπεδο **Industry**

Sales volume

	Electronics	Store1	Store2
Q1	VCR	\$1,4	\$1,4
	Camcorder	\$0,6	\$0,6
	TV	\$2,0	\$2,4
	CD player	\$1,2	\$1,2
Q2	VCR	\$2,4	\$2,4
	Camcorder	\$3,3	\$1,3
	TV	\$2,2	\$2,5
	CD player	\$1,0	\$1,0

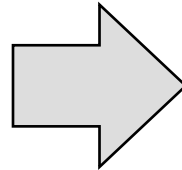
Item: Επίπεδο **Category**

Περιστροφή - Pivot

- Εναλλαγή των γραμμών και των στηλών του κύβου, όπως αυτός παρουσιάζεται στην οθόνη
- Δεν απαιτείται κανένας νέος υπολογισμός στη ΒΔ
- Ενδιαφέρον θέμα: πώς παρουσιάζω ένα κύβο με περισσότερες των 2 διαστάσεων? Μέχρι στιγμή, απλώς **φωλιάζουμε τις διαστάσεις** τη μία μέσα στην άλλη...

Pivot

Sales volume			
	Products	Store1	Store2
Q1	Electronics	\$5,2	\$5,6
	Toys	\$1,9	\$1,4
	Clothing	\$2,3	\$2,6
	Cosmetics	\$1,1	\$1,1
Q2	Electronics	\$8,9	\$7,2
	Toys	\$0,75	\$0,4
	Clothing	\$4,6	\$4,6
	Cosmetics	\$1,5	\$0,5



Sales volume			
	Products	Q1	Q2
Store 1	Electronics	\$5,2	\$8,9
	Toys	\$1,9	\$0,75
	Clothing	\$2,3	\$4,6
	Cosmetics	\$1,1	\$1,5
Store 2	Electronics	\$5,6	\$7,2
	Toys	\$1,4	\$0,4
	Clothing	\$2,6	\$4,6
	Cosmetics	\$1,1	\$0,5

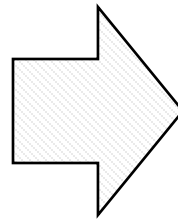
Εναλλαγή γραμμών και στηλών

Επιλογή τιμών -- Slice & Dice

- ✦ **Slice** : Συνθήκη της μορφής **Level = value**
 - ✦ Π.χ., κράτα μόνο το **Store 1** από τα καταστήματα
- ✦ **Dicing** : Οποιαδήποτε πιο πολύπλοκη συνθήκη
 - ✦ Π.χ., κράτα μόνο το **Store 1** από τα καταστήματα και τις βιομηχανίες **Clothing** και **Cosmetics**
 - ✦ Παλαιότερα, οι όροι είχαν άλλο νόημα: slicing ήταν η επιλογή τιμών και dicing η επιλογή στηλών.
 - ✦ Πλέον, οι δύο όροι είναι όπως παραπάνω...

Slice and Dice

Sales volume			
	Products	Store1	Store2
Q1	Electronics	\$5,2	\$5,6
	Toys	\$1,9	\$1,4
	Clothing	\$2,3	\$2,6
	Cosmetics	\$1,1	\$1,1
Q2	Electronics	\$8,9	\$7,2
	Toys	\$0,75	\$0,4
	Clothing	\$4,6	\$4,6
	Cosmetics	\$1,5	\$0,5



Sales volume		
	Products	Store1
Q1	Electronics	\$5,2
	Toys	\$1,9
Q2	Electronics	\$8,9
	Toys	\$0,75

Διώξε το **Store 2** και τις βιομηχανίες **Clothing & Cosmetics**

Τελεστές Rollup και Cube

➤ Τελεστής **Rollup**

✦ group by rollup product, store, city

➤ group by product, store, city

➤ group by store, city

➤ group by city

Όχι όμως ανά product, city!!

Όχι όμως ανά product!!

➤ Τελεστής **Cube** για όλους τους δυνατούς συνδυασμούς

✦ group by cube product, store, city

➤ group by κάθε υποσύνολο του {*product, store, city*},
ανεξάρτητα από τη σειρά που έδωσα στις στήλες αυτές στην
εντολή

Το αποτέλεσμα των τελεστών δεν παράγει πολλά μικρά πινακάκια, αλλά έναν πίνακα με εγγραφές με NULL όπου δεν μπαίνει τιμή...

Τελεστές Rollup και Cube

```
select color, make, year, sum(units)
from car_sales
where make in {"chevy", "ford"}
and year between 1990 and 1994
group by cube color, make, year
having sum(units) > 0;
```

Aggregate



Sum

Group By
(with total)

By Color

RED
WHITE
BLUE

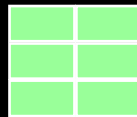


Sum

Cross Tab

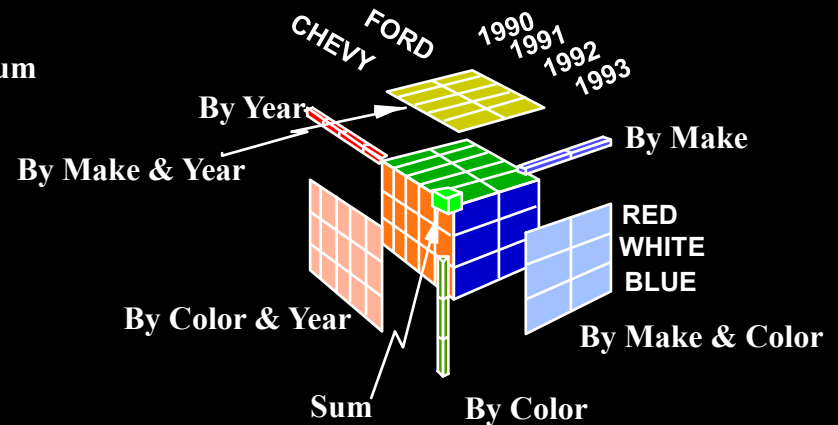
Chevy Ford By Color

RED
WHITE
BLUE



Sum

By Make



Jim Gray
Adam Bosworth
Andrew Layman
Microsoft

Hamid Pirahesh
IBM

Top N Queries

- Έστω ότι ψάχνουμε τα 10 πιο φθηνά αυτοκίνητα από ένα πίνακα.
- Κανονικά, η ερώτηση σημαίνει ότι θα ταξινομήσουμε τον πίνακα και θα βρούμε τα 10 πιο φθηνά
- Στην SQL δεν υποστηρίζεται προγραμματιστικά αυτό («δείξε μου ΜΟΝΟ τα 10 πρώτα»), αλλά κυρίως, δεν το λαμβάνει υπόψη του ο βελτιστοποιητής

Top N Queries

```
SELECT P.pid, P.pname, S.sales  
FROM Sales S, Products P  
WHERE S.pid=P.pid AND S.locid=1 AND S.timeid=3  
ORDER BY S.sales DESC  
OPTIMIZE FOR 10 ROWS
```

- Το **OPTIMIZE FOR** clause καλύπτει την απαίτηση προγραμματιστικά. Δεν είναι μέρος της SQL:1999 αλλά υπάρχει π.χ., στην DB2!
- Πώς θα εκτελούσαμε έξυπνα αυτό το query? Αν διαλέγαμε μια τιμή **c**, ώστε να είναι η τιμή του 10ου αυτοκίνητου, θα μπορούσαμε να βάζαμε την επιλογή **cost<c** και να εκτελούσαμε την ερώτηση.

Top N Queries



```
SELECT P.pid, P.pname, S.sales  
FROM Sales S, Products P  
WHERE S.pid=P.pid AND S.locid=1 AND S.timeid=3  
ORDER BY S.sales DESC  
OPTIMIZE FOR 10 ROWS
```

```
SELECT P.pid, P.pname, S.sales  
FROM Sales S, Products P  
WHERE S.pid=P.pid AND S.locid=1 AND S.timeid=3  
      AND S.sales > c  
ORDER BY S.sales DESC
```

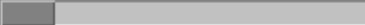
Top N Queries

- **Κέρδος:** Αποφύγαμε να ταξινομήσουμε!
- Η τιμή c επιλέγεται από τον optimizer με βάση κάποιο ιστόγραμμα.
 - ✦ Θυμηθείτε, το ιστόγραμμα **προβλέπει** κατανομή τιμών!
- Τι θα γίνει αν τελικά πάρουμε περισσότερες από 10 απαντήσεις?
- Τι θα γίνει αν τελικά πάρουμε λιγότερες από 10 απαντήσεις?

Online Aggregation

Postgres95 Online Aggregation Interface

	AVG	Confidence	Interval
	2.6336	95	0.0652539

14% done 

Postgres95 Online Aggregation Interface

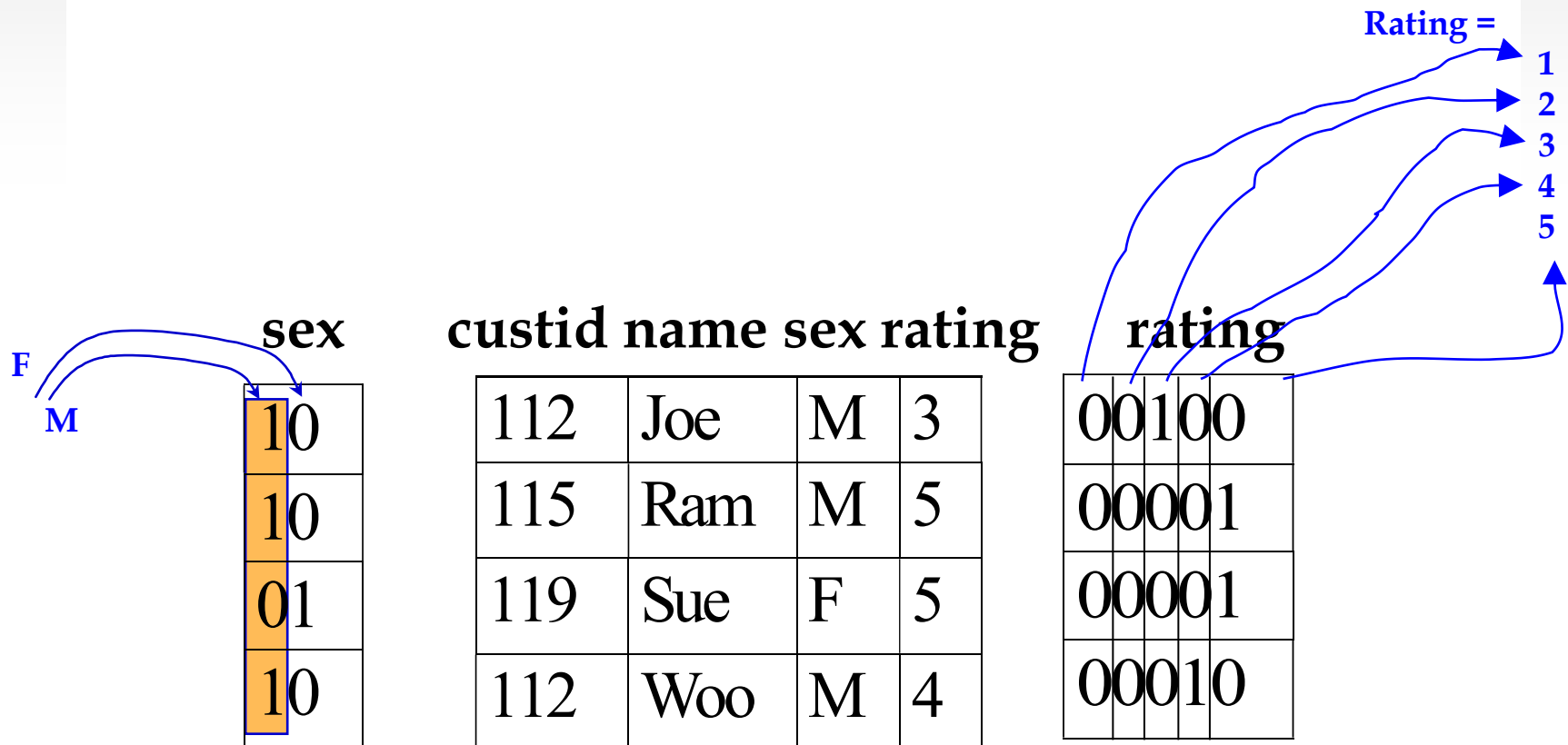
	major	AVG	Confidence	Interval
	1	2.27216	95	0.160417
	2	2.56146	95	0.160417
	3	2.66702	95	0.160417
	4	2.86235	95	0.160417
	5	3.12048	95	0.160417
	9	2.89645	95	0.160417

14% done 

Ευρετήρια για OLAP ερωτήσεις

- **Bitmap index:** για κάθε τιμή ενός γνωρίσματος, και μια στήλη με bits
- Ανάλογα με την τιμή, ακριβώς ένα bit γίνεται 1 και όλα τα άλλα είναι 0

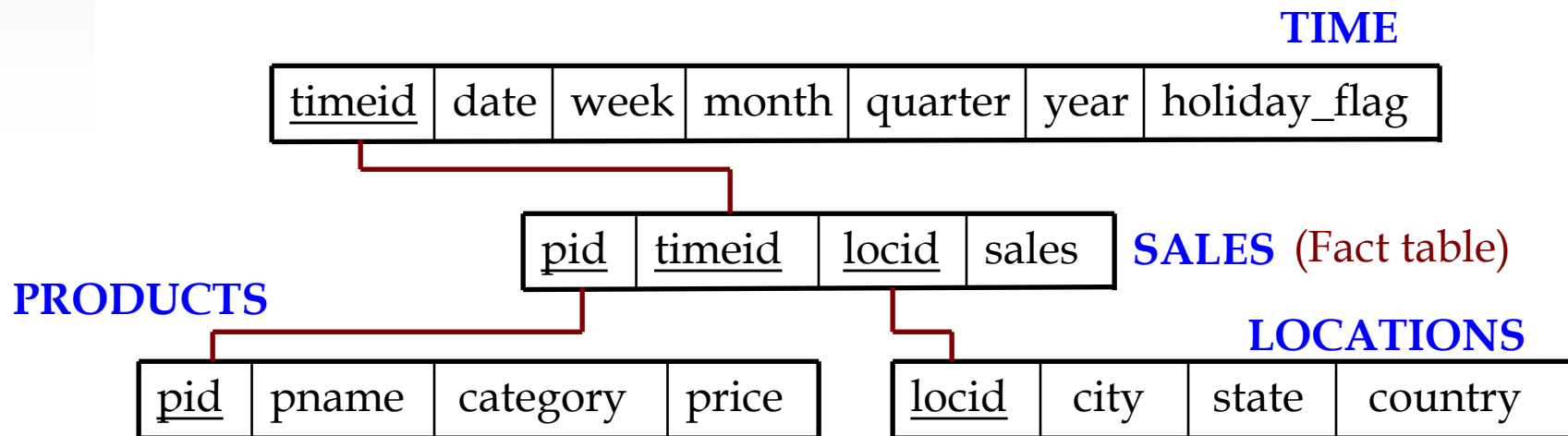
Ευρετήρια για OLAP ερωτήσεις



Ευρετήρια για OLAP ερωτήσεις

- Ποιοι ναύτες είναι άνδρες **ΚΑΙ/Η** έχουν rating = 5?
 - ✦ Bitwise **AND/OR** των αντίστοιχων bitmap indexes?
- **Πόσοι** ναύτες έχουν rating = 3 ?
 - ✦ Count “1” στη στήλη που αφορά το rating = 3

Ευρετήρια για OLAP ερωτήσεις



- Έστω το εν λόγω σχήμα με τρεις πίνακες διαστάσεων και ένα fact table

Join index

- Έστω ότι θέλω να ρωτήσω σύνολο πωλήσεων, δηλ. $\text{sum}(\text{sales})$ ανά year , pname με μια επιλογή $\text{country}=\text{"USA"}$.
- Ένας **join index** περιέχει τετράδες από rowId 's $[\text{s}, \text{p}, \text{t}, \text{l}]$ με $\text{s} \in \text{Sales}$, $\text{p} \in \text{Products}$, $\text{t} \in \text{Times}$ and $\text{l} \in \text{Locations}$.
- Τότε μπορώ να βρω γρήγορα ποιες εγγραφές αντιστοιχούν στο rowId που αντιστοιχεί σε $\text{country}=\text{"USA"}$ και να συναθροίσω μόνο αυτές στη συνέχεια.
- Οι επιμέρους συνδέσεις γίνονται πολύ πιο εύκολα, επίσης...

Join index

- **Βαριάντα:** για κάθε στήλη στην οποία κάνω συχνά επιλογές, (π.χ., country), φτιάχνουμε ένα join index with [**c**,**s**] αν η σχετική πλειάδα του πίνακα διάστασης **c** αντιστοιχίζεται με την εγγραφή **s** του fact πίνακα
- Αν οι εν λόγω indexes είναι bitmaps, έχουμε ένα **bitmapped join index**.

Join index

- Αν έχω δύο συνθήκες επιλογής, π.χ.,
`Locations.country = “USA”` και `Product.pname = “Coke Cola”`, τότε παίρνουμε την τομή των δύο indexes, πριν από οτιδήποτε άλλο, ώστε να γλιτώσουμε τη σύνδεση.

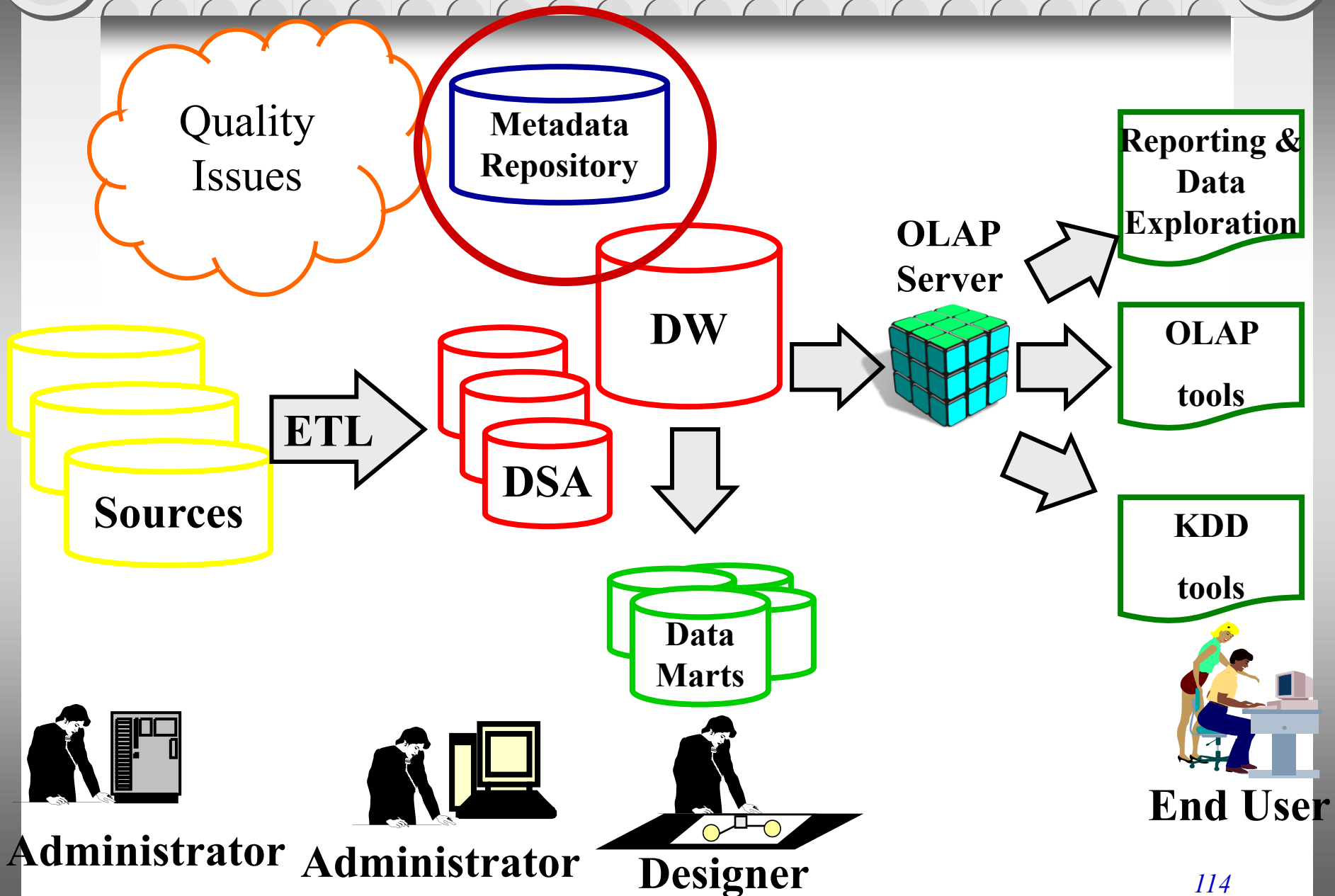
- Π.χ., μπορεί να έχω

➤ `[rid(USA), {rid1, rid2, rid4, rid12, rid18}]`

➤ `[rid(Coke), {rid2, rid12, rid23}]`

Αφορούν τον
πίνακα **Sales**

Λεξικό Μεταπληροφορίας – Metadata Repository



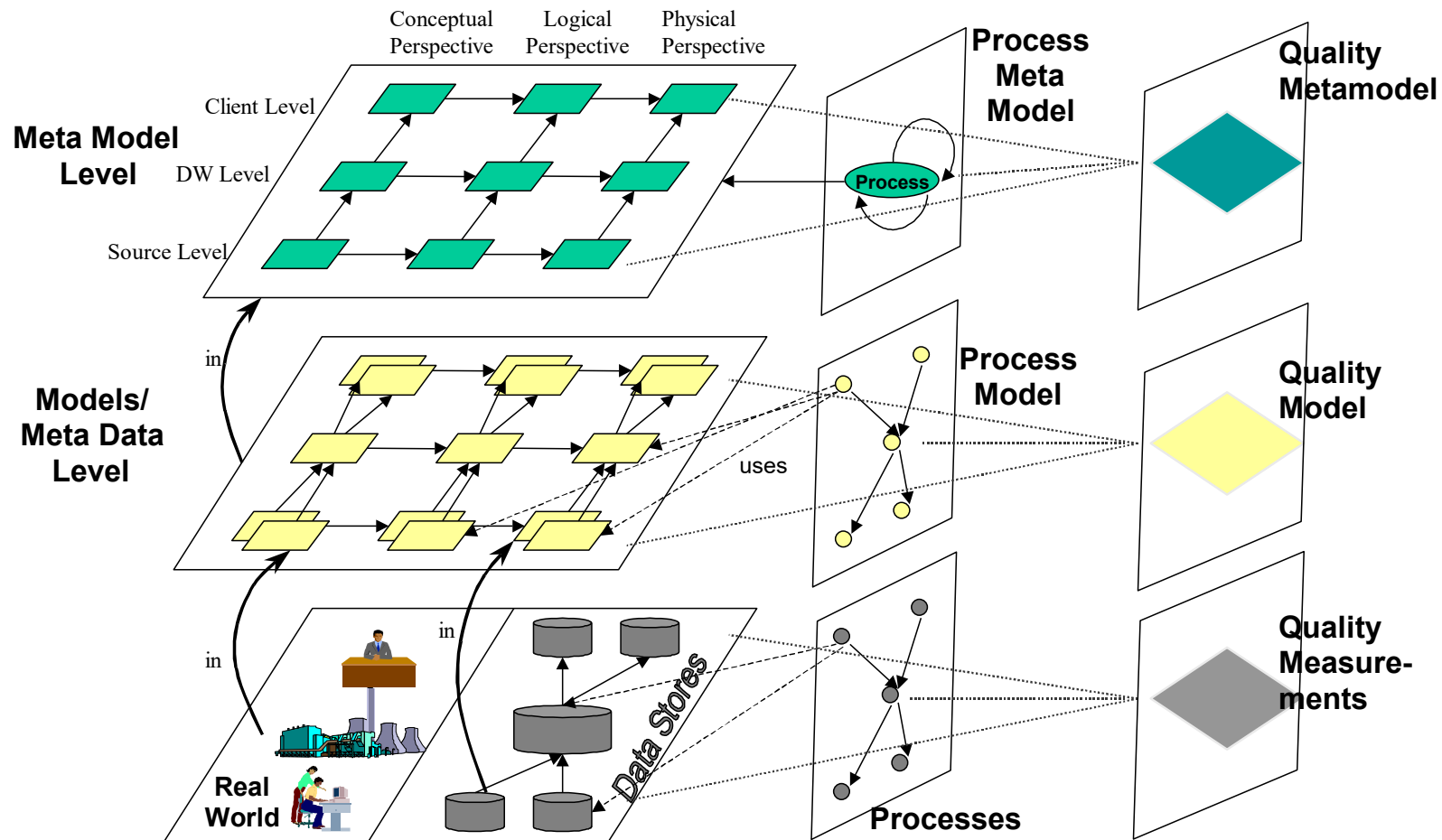
Λεξικό Μεταπληροφορίας

- Μεταπληροφορία Εννοιολογικού Σχήματος
 - ✦ Ορολογία και κανόνες του οργανισμού
- Μεταπληροφορία Λογικού Σχήματος
 - ✦ Περιγραφή του σχήματος των πηγών
 - ✦ Περιγραφή του σχήματος του DW, των Διαστάσεων και των ιεραρχιών τους, των όψεων και των συναθροίσεων
 - ✦ Περιγραφή των διαδικασιών του συστήματος και της ροής του συστήματος
 - ✦ Κανόνες ETL, ενημέρωσης και διαγραφής δεδομένων στο DW
 - ✦ Περιγραφή των υλοποιημένων reports

Λεξικό Μεταπληροφορίας

- Μεταπληροφορία Φυσικού Σχήματος
 - ✦ Φυσική περιγραφή υπολογιστών, δικτύων κλπ
 - ✦ Περιγραφή των (πηγαίων) αρχείων, tablespaces, ...
 - ✦ Περιγραφή των προγραμμάτων που υλοποιούν τις λειτουργικές διαδικασίες
 - ✦ Περιγραφή των gateways για τις πηγές
 - ✦ Προφίλ χρηστών, και δικαιώματα πρόσβασης
 - ✦ Στατιστικά που προκύπτουν από το monitoring του DW

The DWQ Approach



Στάνταρτ για τη μεταπληροφορία

- Metadata Coalition
 - ✦ MetaData Interchange Specification (MDIS)
 - ✦ Open Information Model (OIM)
- OMG
 - ✦ Common Warehouse Model (CWM)
- Microsoft Repository

Επίλογος

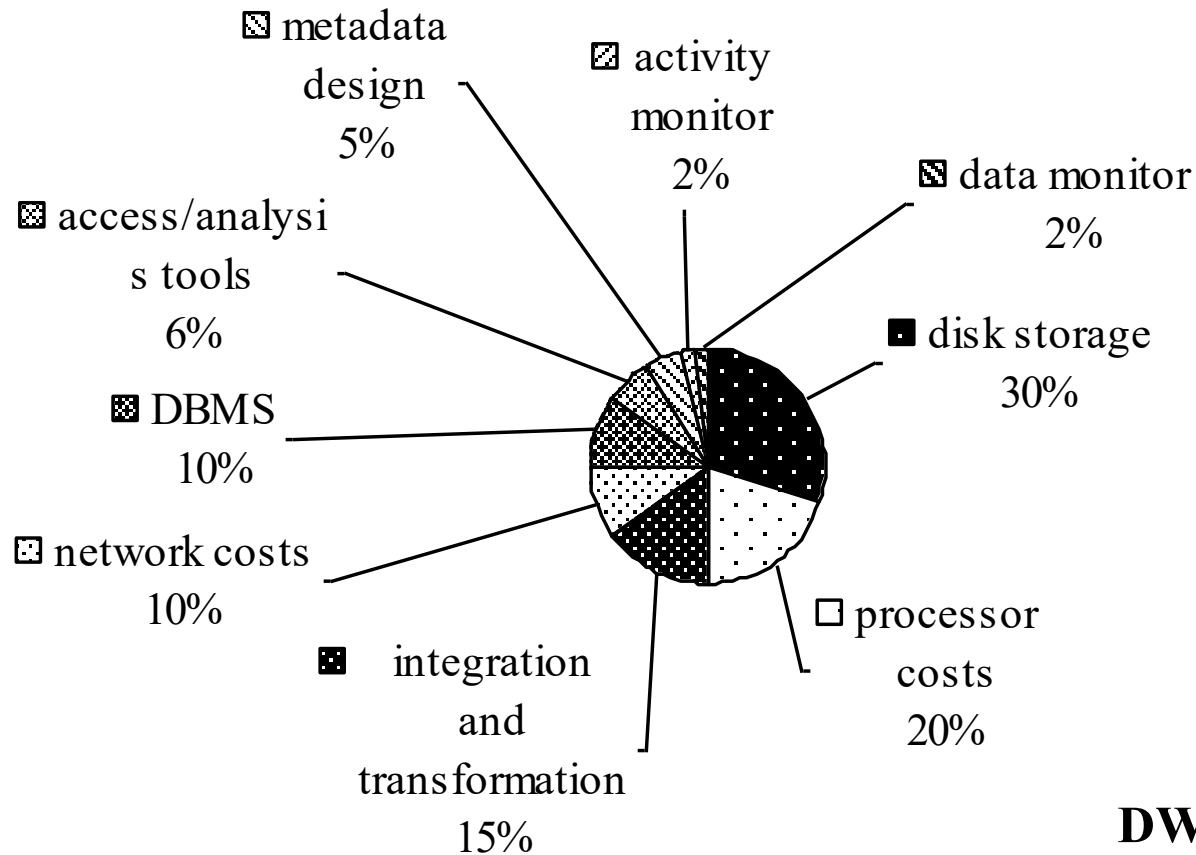
- Τι είναι οι Αποθήκες Δεδομένων
- Αρχιτεκτονική και σχήμα της Αποθήκης Δεδομένων
- Λειτουργικές διαδικασίες της Αποθήκης Δεδομένων
- Συστήματα επερώτησης της Αποθήκης Δεδομένων
- Λεξιικό Μεταπληροφορίας
- Το χρήμα ...

Το χρήμα...

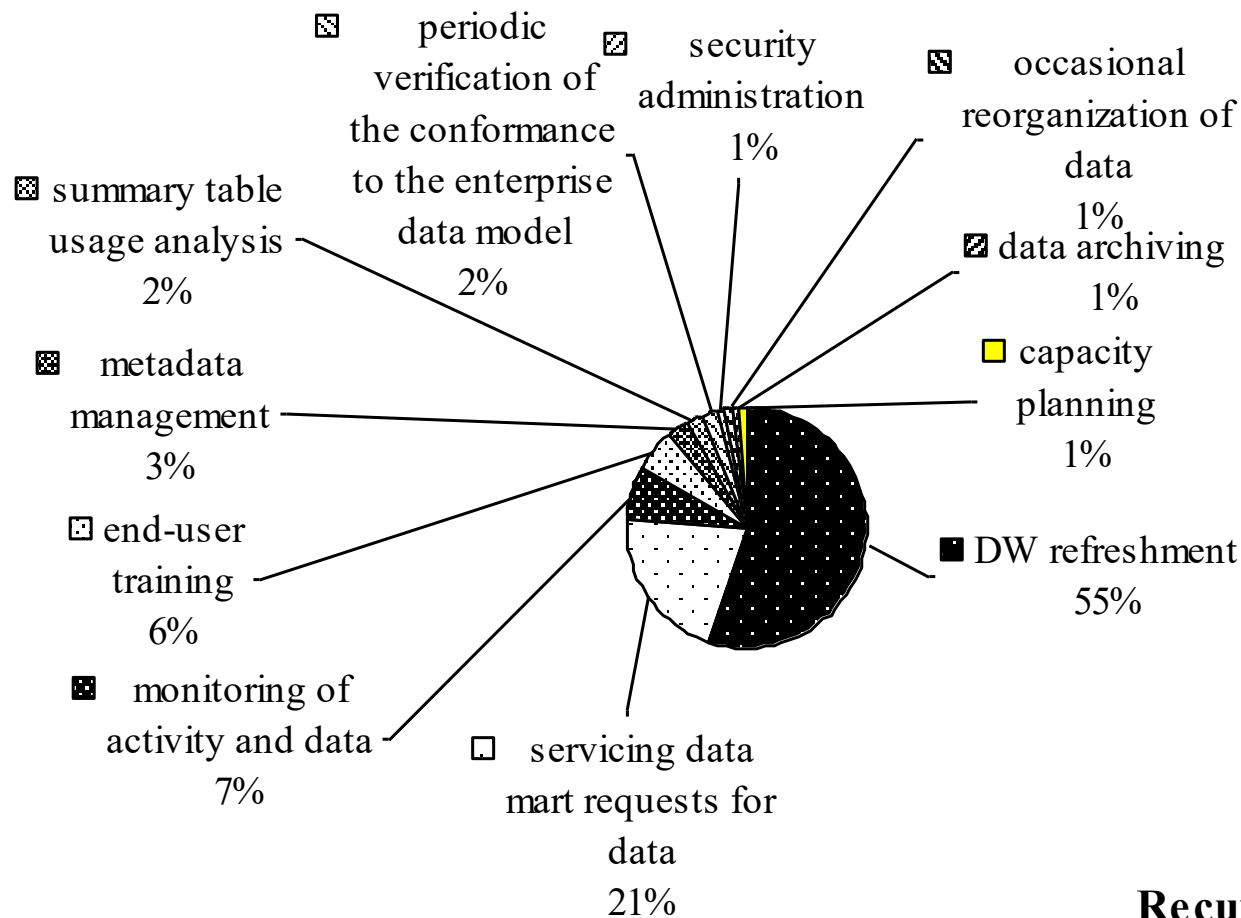
	1998	1999	2000	2001	2002	CAGR (%)
RDBMS sales for DW	900.0	1110.0	1390.0	1750.0	2200.0	25.0
Data Marts	92.4	125.0	172.0	243.0	355.0	40.0
ETL tools	101.0	125.0	150.0	180.0	210.0	20.1
Data Quality	48.0	55.0	64.5	76.0	90.0	17.0
Metadata Management	35.0	40.0	46.0	53.0	60.0	14.4
OLAP (including implementation services)*	2000	2500	3000	3600	4000	18.9

Estimated sales in **millions of dollars** [ShTy98]
(*estimates are from [Pend00]).

Κόστος σχεδίασης [B. Inmon]



Κόστος «καθημερινής» χρήσης [B. Inmon]



Recurring DW costs