Εισαγωγή στις Αποθήκες Δεδομένων

Διαφάνειες βασισμένες σε σχετικές διαφάνειες του Πάνου Βασιλειάδη

Εισαγωγή: OLTP

Παραδοσιακή Διαχείριση Δεδομένων με ΣΔΒΔ

Σύστημα Επεξεργασίας Δοσοληψιών - On-Line Transaction Processing (OLTP)

Ένα πλήρες σύστημα που περιέχει εργαλεία για προγραμματισμό εφαρμογών, εκτέλεση και διαχείριση των δοσοληψιών

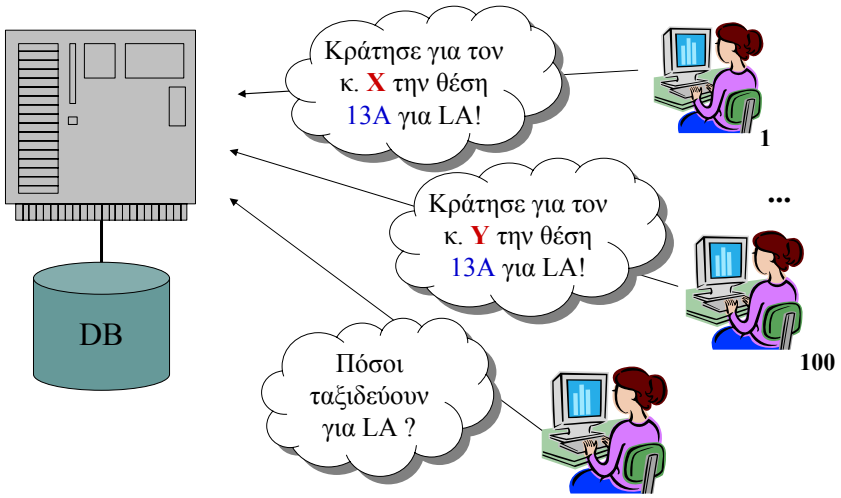
Μια τέτοια εφαρμογή πρέπει να δουλεύει συνεχώς, να αντεπεξέρχεται αποτυχιών, εξελίσσεται συνεχώς, είναι συνήθως κατανεμημένη και περιλαμβάνει:

- Βάση Δεδομένων
- Δίκτυο
- Προγράμματα εφαρμογής

Εξαιρετικά κρίσιμη για τη λειτουργία κάθε οργανισμού

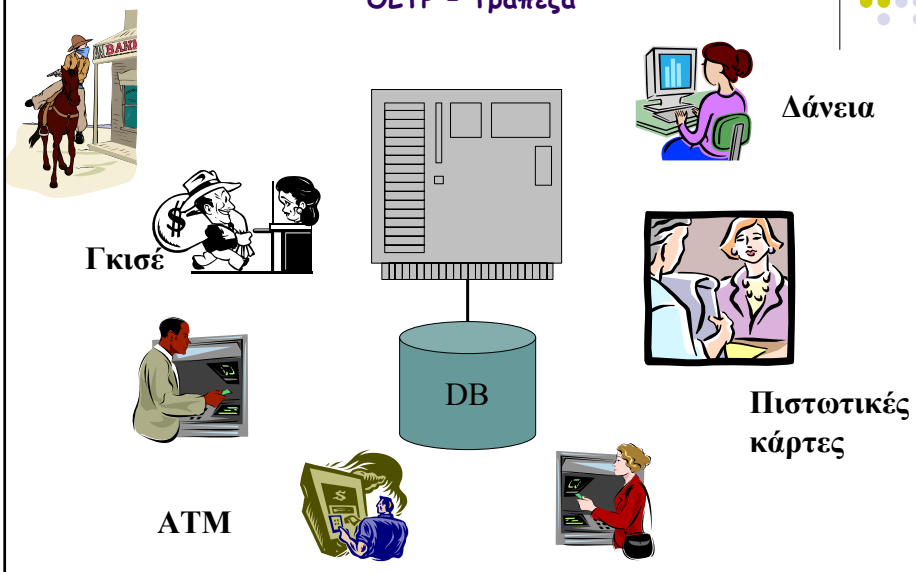
Εισαγωγή: OLTP

OLTP – Αεροπορική Εταιρεία



Εισαγωγή: OLTP

OLTP – Τράπεζα



Εισαγωγή: OLTP

OLTP – Βασικά Χαρακτηριστικά

- Ελάχιστος χρόνος διαθέσιμος για την εκτέλεση μιας δοσοληψίας.
- 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.

Εισαγωγή: OLAP

OLAP

- 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 προτύπων σε τεράστιες βάσεις δεδομένων
  - OLAP + Data Mining => On-line Analytical Mining

Εισαγωγή: OLAP

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 κέρδη των πωλήσεων τα 10 τελευταία χρόνια;
- Ποια ήταν τα δέκα πρώτα καταστήματα σε πωλήσεις CD;
- Πόσους δίσκους πουλήσαμε στην Πελοπόννησο 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;

Εισαγωγή: OLAP

Λειτουργικά Χαρακτηριστικά Απαιτήσεων OLAP

- Πρόσβαση σε *μεγάλο όγκο* δεδομένων
- Συμμετοχή *αθροιστικών* και *ιστορικών* δεδομένων σε πολύπλοκες ερωτήσεις
- Μεταβολή της *οπτικής γωνίας* ή *βαθμού αφαίρεσης* 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 ("On-Line").

Πως θα το πετύχουμε;

Εισαγωγή

Δύο κεντρικά θέματα

- Απόδοση
  - Αν μια πολύπλοκη OLAP ερώτηση χρειαστεί να κλειδώσει ένα ολόκληρο πίνακα, τότε όλες οι OLTP δοσοληψίες την περιμένουν μέχρι να τελειώσει
- 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…

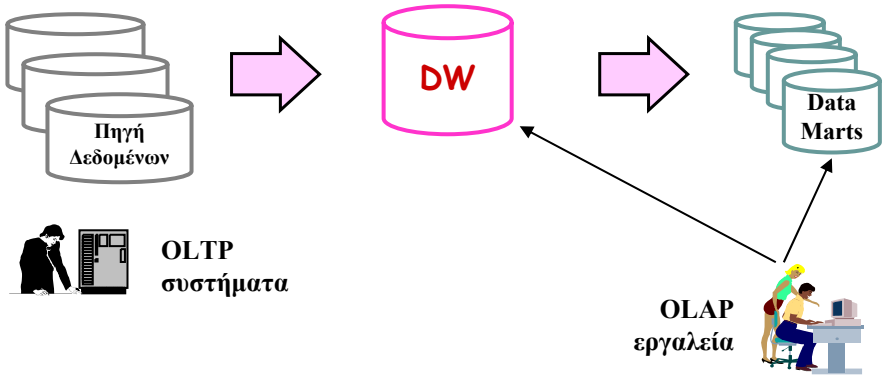
Εισαγωγή: Αποθήκη Δεδομένων

Αποθήκες Δεδομένων

- Μια **κεντρικοποιημένη 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
- Μπορεί να συμπληρώνεται και από εξειδικευμένα *θεματικά υποσύνολα* (Data Marts) για περαιτέρω απόδοση των OLAP εφαρμογών

Εισαγωγή: Αποθήκη Δεδομένων

Γενική Αρχιτεκτονική



Εισαγωγή: Αποθήκη Δεδομένων

Αποθήκες Δεδομένων: Δύο ορισμοί

- Μια ΒΔ υποστήριξης αποφάσεων, που διατηρείται 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, έχει ολοκληρωμένα (ενοποιημένα) δεδομένα, τα οποία διατηρούνται σε βάθος χρόνου χωρίς να διαγράφονται.

  W.H. Inmon, Building the Data Warehouse, 1992 (ο εφευρέτης του όρου)

Εισαγωγή: Αποθήκη Δεδομένων

Προτερήματα/Ιδιότητες

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

Ποιότητα Δεδομένων

Η ποιότητα των δεδομένων στις πηγές είναι συχνά προβληματική (τα δεδομένα μπορεί να μην είναι πλήρη, να έχουν ασυνέπειες, να είναι παλιά, να παραβιάζουν τους λογικούς και δομικούς κανόνες αξιοπιστίας, κλπ)

- Έχει βρεθεί ότι τουλάχιστο 10% των δεδομένων είναι προβληματικά στις πηγές, με αποτέλεσμα οικονομικές απώλειες του 25-40%
- Πριν την εισαγωγή στις αποθήκες δεδομένων καθαρισμός, επίσης λειτουργεί και ως ένα ενδιάμεσο σύστημα στον οποίο καθαρίζουμε τα δεδομένα

Source 1: Personnel (Cobol)

EMP ID	Name	DoB	Salary	Total Income	DeptID
110	Kostas	1/1/72	1500	1200	132
…	…				…

Source 2: Accounting (DB2)

EMP ID	IL_ID	Amount
110	10	1500
110	30	300

EMP INCOME

EMP ID	Name	Age
110	Kostas	30
120	Mitsos	48
130	Roula	29

EMP

IL_ID	Descr
10	Μισθός
20	Επίδομα Τέκνων
30	Φόρος
…	…

Income Lookup

Εισαγωγή: Αποθήκη Δεδομένων

Προτερήματα/Ιδιότητες

Απόδοση

Οι εφαρμογές OLAP επιταχύνονται αν τα δεδομένα οργανωθούν με μη παραδοσιακούς τρόπους (π.χ., απο-κανονικοποιημένα)

ΣΔΒΔ για OLTP (ευρετήρια, επεξεργασία δοσοληψιών)

Οι σύνθετες OLAP ερωτήσεις θα συγκρούονταν με τις παραδοσιακές OLTP δοσοληψίες, με αποτέλεσμα την υπερφόρτωση του συστήματος

Θεματικά προσανατολισμένη: Διατήρηση μόνο των σχετικών 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,...)

Εισαγωγή: Αποθήκη Δεδομένων

Προτερήματα/Ιδιότητες

Ιστορικά Δεδομένα

- Ο χρονικός ορίζοντας μια αποθήκης δεδομένων είναι πολύ μεγαλύτερος από ότι ενός συστήματος σε λειτουργία
- Η ΒΔ έχει τα τωρινά δεδομένα ενώ οι αποθήκες διατηρούν και παλιά δεδομένα (πχ τα προηγούμενα 5-10 χρόνια)

Τροποποιήσεις

- Οι τροποποιήσεις στις πηγές δεδομένων δεν φαίνονται άμεσα στις αποθήκες δεδομένων, συνήθως περιοδικά
- Μόνο δύο βασικές λειτουργίες: αρχικό φόρτωμα των δεδομένων (loading) και προσπέλαση δεδομένων (access)

Εισαγωγή: Αποθήκη Δεδομένων

OLTP vs OLAP

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

Εισαγωγή: Αποθήκη Δεδομένων

OLTP vs OLAP

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

Εισαγωγή: Αποθήκη Δεδομένων

OLTP vs OLAP

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

Εισαγωγή: Αποθήκη Δεδομένων

Σύγκριση με ενοποίηση ετερογενών ΣΔΒΔ

Wrapper/mediators

Με βάση την ερώτηση, μεταφράζεται ανάλογα, εκτελείται σε κάθε ΣΔΒΔ και τα αποτελέσματα ενοποιούνται σε μια ολική απάντηση

Μοντέλο Δεδομένων και Λειτουργίες

Εισαγωγή

Με λίγα λόγια …

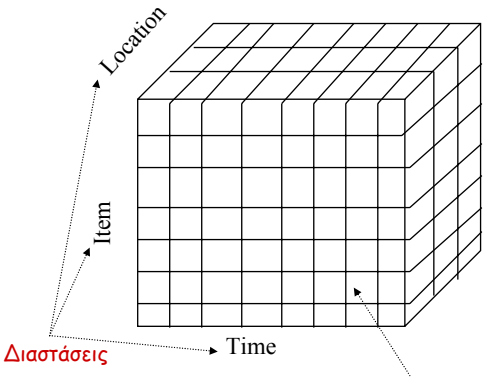
Μια αποθήκη δεδομένων βασίζεται σε ένα πολυδιάστατο μοντέλο δεδομένων (multidimensional data model) που αναπαριστά τα δεδομένα με τη μορφή ενός κύβου δεδομένων (data cube)

Ένας κύβος δεδομένων (data cube) επιτρέπει την μοντελοποίηση και την θεώρηση των δεδομένων από πολλές οπτικές γωνίες – Διαστάσεις (dimensions)-

Για συγκεκριμένες τιμές στις διαστάσεις μια Μέτρηση (Measure) – αυτό που μας ενδιαφέρει να μετρήσουμε

Παράδειγμα

Κύβος ΠΩΛΗΣΕΙΣ



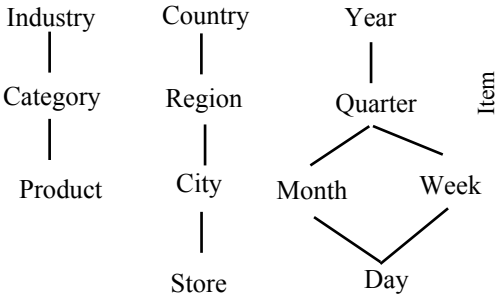
Μέτρηση: Αριθμός Πωλήσεων για τις συγκεκριμένες διαστάσεις (Location, Item, Time)

Εννοιολογική Ιεραρχία

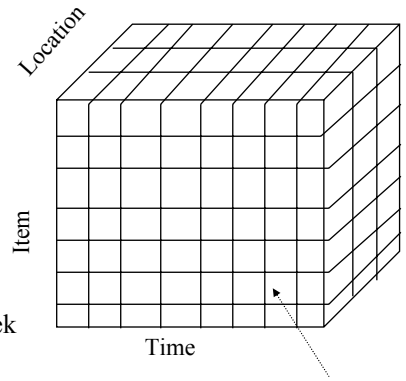
Ιεραρχίες Διαστάσεων

Κάθε διάσταση παίρνει τιμές από διαφορετικά επίπεδα, μπορεί να εκφραστεί σε διαφορετικά επίπεδα λεπτομέρειας

Διαστάσεις: Product, Region, Date
Ιεραρχίες διαστάσεων:



Κύβος ΠΩΛΗΣΕΙΣ

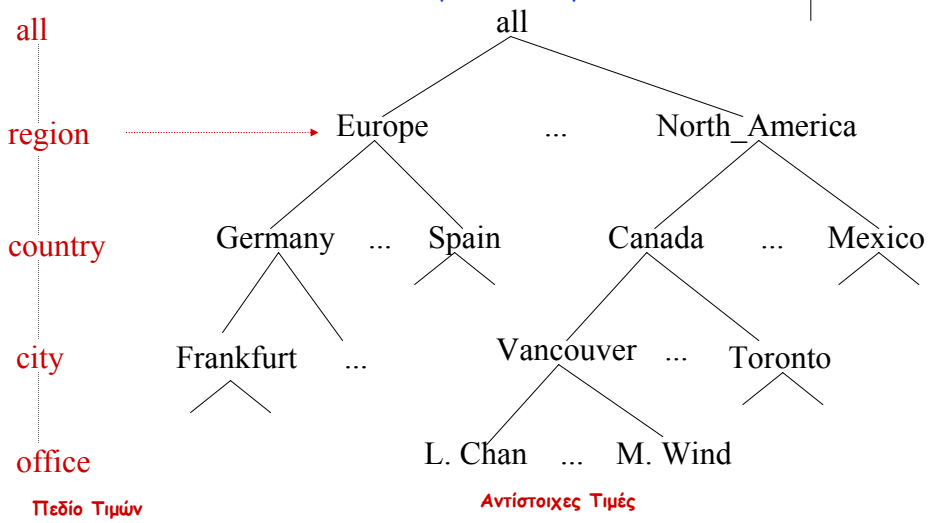


Μέτρηση: Αριθμός Πωλήσεων για τις συγκεκριμένες διαστάσεις (Location, Item, Time)

Εννοιολογική Ιεραρχία

Παράδειγμα:

Εννοιολογική ιεραρχία (Concept Hierarchy) για Location



Πεδίο Τιμών

Αντίστοιχες Τιμές

Εννοιολογικό Μοντέλο

Μοντέλο Δεδομένων (Σχήμα)

Σε σχεσιακό μοντέλο

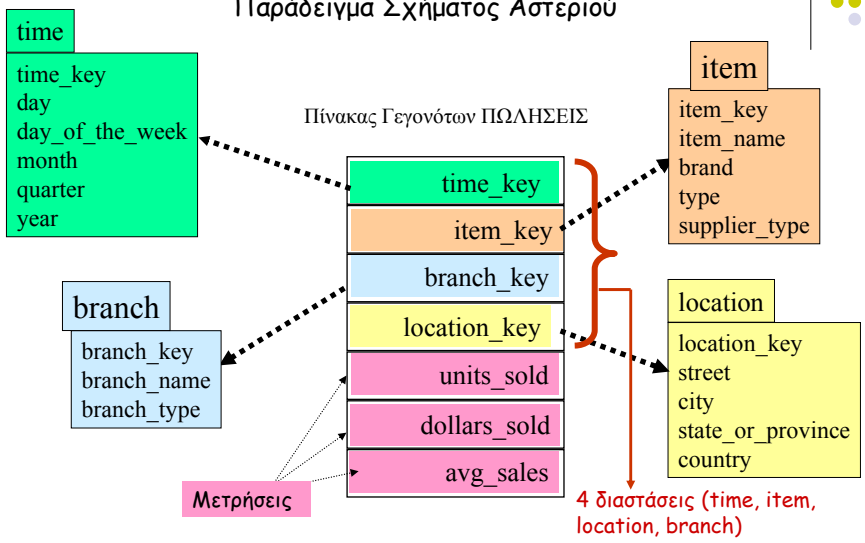
Πίνακες Διαστάσεων
Πίνακας με πληροφορία σχετικά με κάθε διάσταση
Item (item_name, brand, type),
Time(day, week, month, quarter, year)

Πίνακας γεγονότων (Fact Table) έχει ως γνωρίσματα:
- τις μετρήσεις (πχ αριθμός πωλήσεων, τιμή σε δολάρια, κλπ) +
- το πρωτεύον κλειδί κάθε σχετικού πίνακα διαστάσεων

Σχήμα Αστέρι (Star schema)
Πίνακας γεγονότων στο κέντρο που συνδέεται με ένα σύνολο από πίνακες διαστάσεων

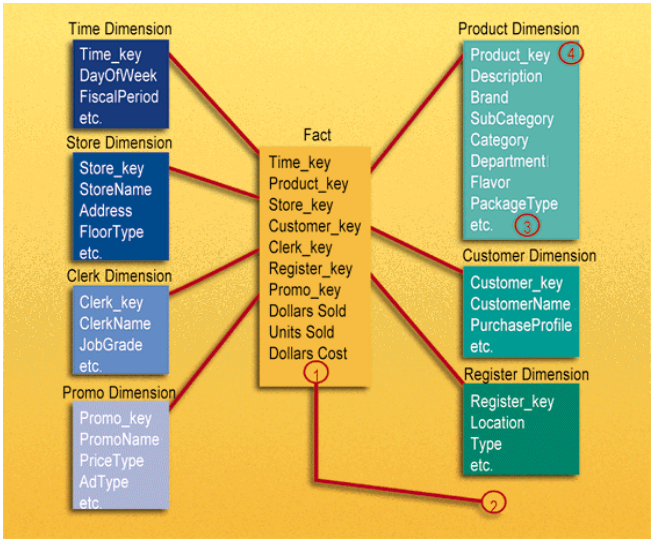
Εννοιολογικό Μοντέλο

Παράδειγμα Σχήματος Αστεριού



Εννοιολογικό Μοντέλο

Παράδειγμα Σχήματος Αστεριού



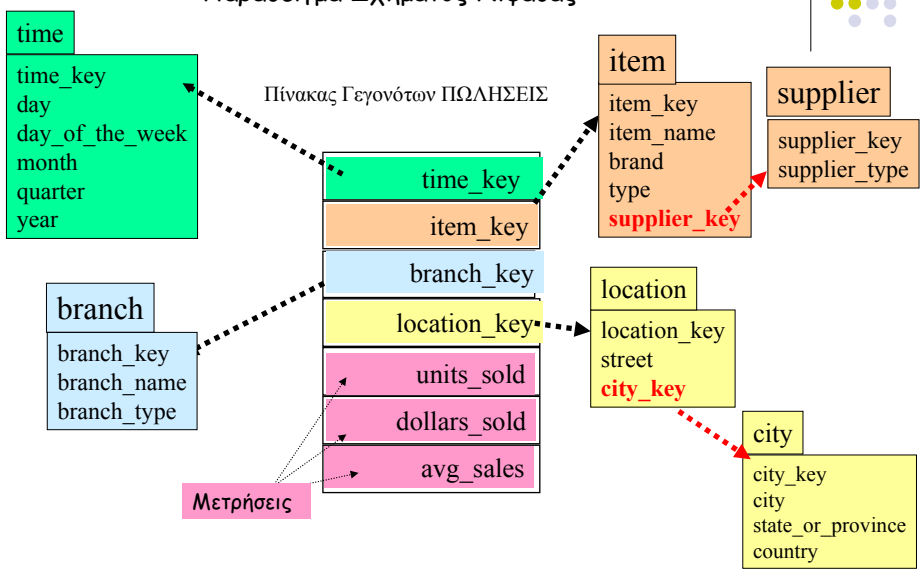
Εννοιολογικό Μοντέλο

Σχήμα Νιφάδας (Snowflake schema)

Μια βελτίωση του σχήματος αστέρι όπου η ιεραρχία διαστάσεων κανονικοποιείται σε ένα σύνολο από μικρότερους πίνακες διαστάσεων

Εννοιολογικό Μοντέλο

Παράδειγμα Σχήματος Νιφάδας



Εννοιολογικό Μοντέλο

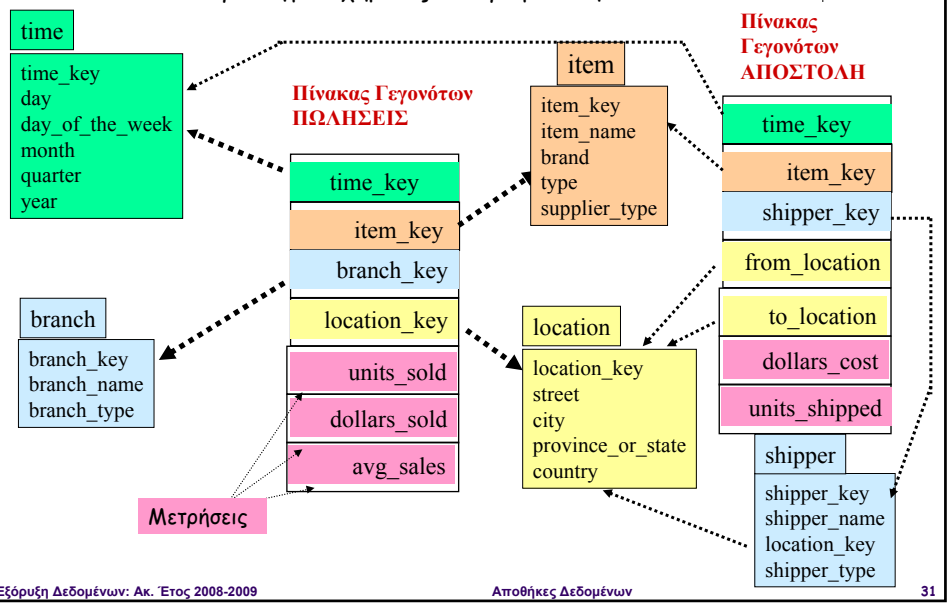
Αστερισμοί Γεγονότων (Fact constellations)

Πολλαπλοί Πίνακες Γεγονότων που μοιράζονται τους Πίνακες Διαστάσεων, μπορούμε να τους δούμε ως συλλογή από αστέρια και άρα ως

Αστερισμό Γεγονότων ή Σχήμα Γαλαξία (galaxy schema)

Εννοιολογικό Μοντέλο

Παράδειγμα Σχήματος Αστερισμού Γεγονότων



Κύβος Δεδομένων

Ορολογία

Συχνά ο n-D κύβος λέγεται βασικός κυβοειδής (base cuboid).

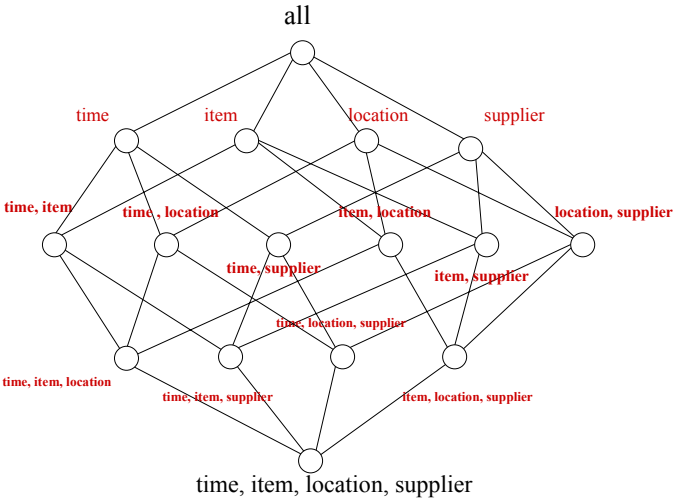
Στο παράδειγμα ο κύβος με τις τέσσερεις διαστάσεις (Item, Time, Branch, Location)

Ο 0-D cuboid που περιέχει τη μεγαλύτερο επίπεδο περίληψης, apex cuboid.

Το πλέγμα των κυβοειδών κύβος δεδομένων.

Κύβος Δεδομένων

Πλέγμα Κυβοειδών – Κύβος δεδομένων



0-D (apex) cuboid

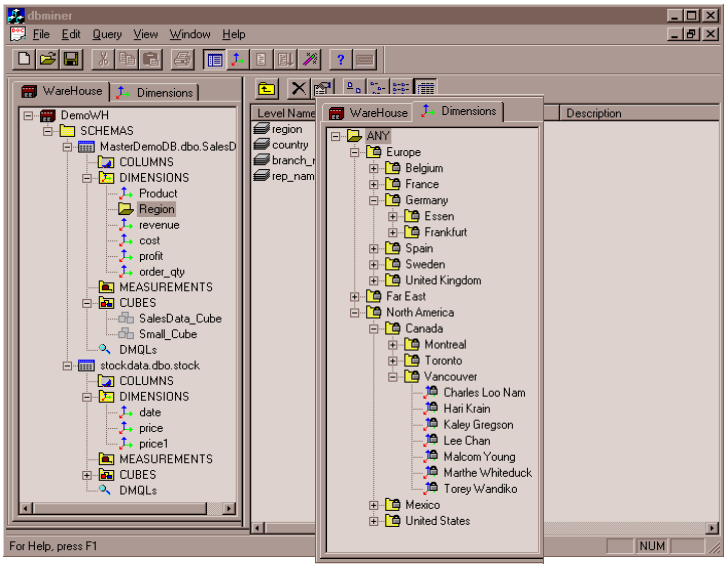
1-D cuboids

2-D cuboids

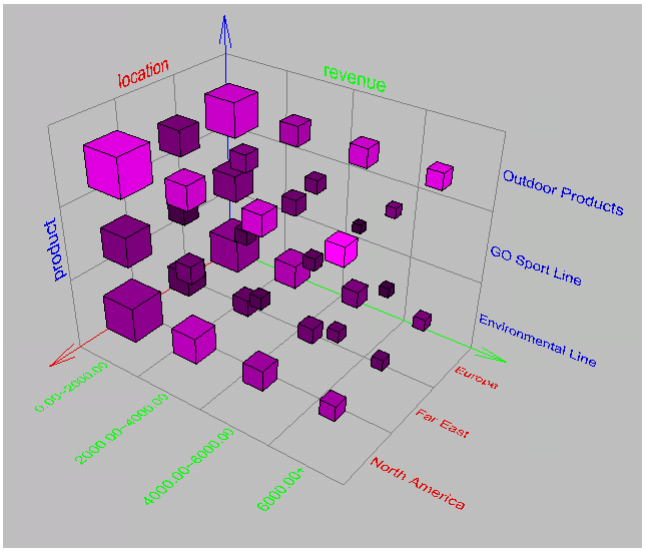
3-D cuboids

4-D (base) cuboid

Παράδειγμα Ιεραρχιών



Οπτικοποίηση Κύβου



Servers & Τεχνολογικές λύσεις

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

Σχεσιακά ΣΔΒΔ & 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)

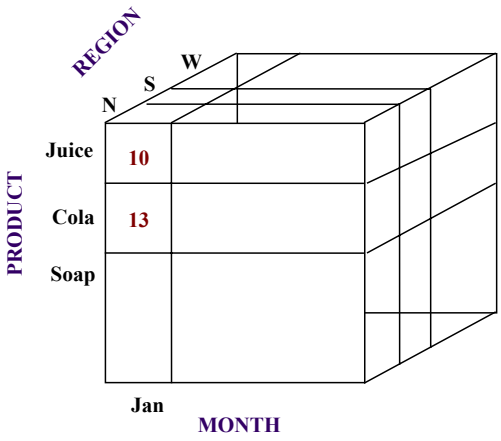
+ Δυνατότητα υποβολής οποιασδήποτε ερώτησης
+ Εύκολη χρήση από τους administrators που γνώριζαν τη σχεσιακή τεχνολογία

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
```


Πολυδιάστατοι πίνακες

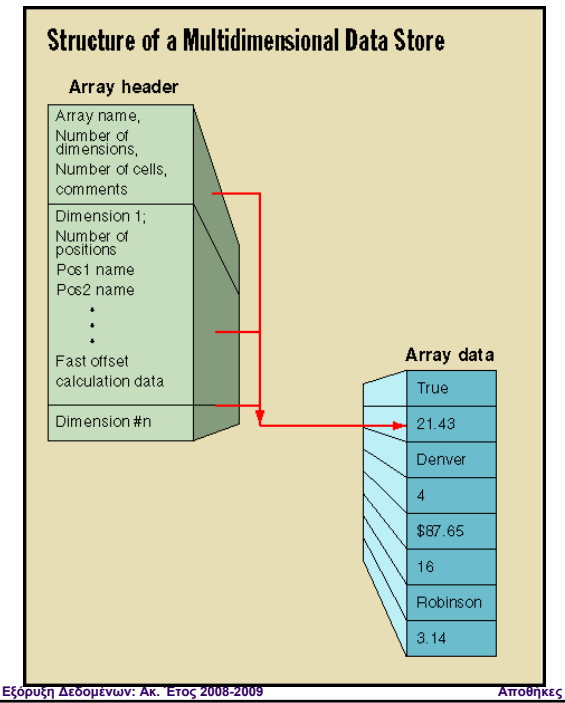


MOLAP Servers

- Η αποθήκευση γίνεται σε πολυδιάστατους πίνακες (multi-dimensional arrays)
  - «πίνακες» με την έννοια της άλγεβρας / γλωσσών προγραμματισμού /..., και όχι του σχεσιακού μοντέλου
- Χρήση τεχνικών συμπιέσεως (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



Υλοποίηση πολυδιάστατων πινάκων

Μετρήσεις - Συναθροίσεις

- 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

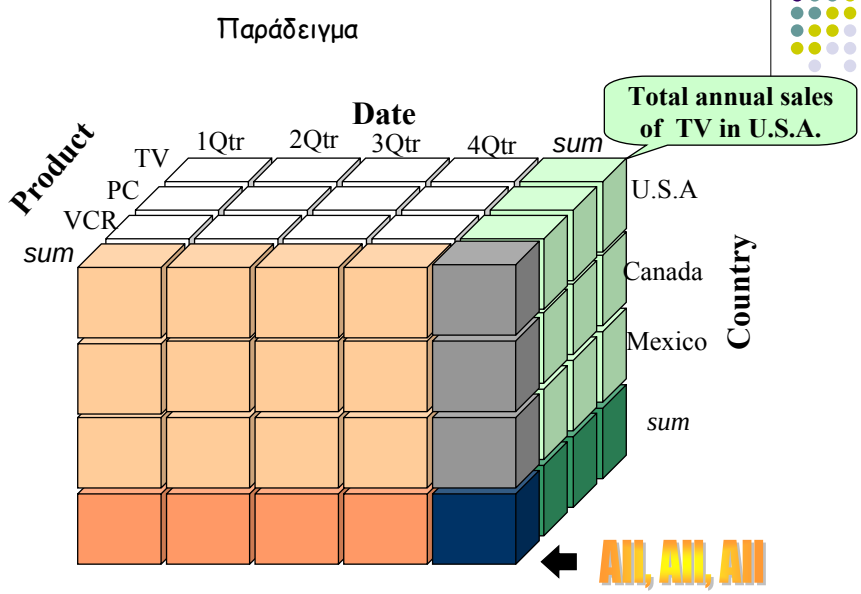
Μετρήσεις - Συναθροίσεις

Υπάρχουν τρεις κατηγορίες μετρήσεων:

- Κατανεμημένες (Distributive): αν μπορούμε να διαμερίσουμε τα δεδομένα και να υπολογίσουμε τη συναθροιστική συνάρτηση σε κάθε διαμέριση ξεχωριστά και σχεδόν άμεσα από αυτές τις τιμές να υπολογίσουμε την ολική τιμή Πχ count(), sum(), min(), max()
- Αλγεβρικές (Algebraic): πάλι μπορούμε να υπολογίσουμε την ολική τιμή της συνάρτησης από τις τιμές της συνάρτησης στις διαμερίσεις χρησιμοποιώντας *M* γνωρίσματα (όπου *M* σταθερά), Πχ. avg(), min_N(), standard_deviation()
- Ολιστικές (Holistic): δεν υπάρχει όριο (πολυπλοκότητα) σταθερής τάξης για το χώρο αποθήκευσης που χρειαζόμαστε για τον υπολογισμό της ολικής τιμής από τις τιμές στις διαμερίσεις, Πχ. median(), mode(), rank()

Βασικές Πράξεις

Παράδειγμα



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: Ενσωμάτωση των αθροιστικών εγγραφών στον βασικό (base/basic) fact table + μια επιπλέον στήλη που να εξηγεί το επίπεδο συνάθροισης

sum

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**

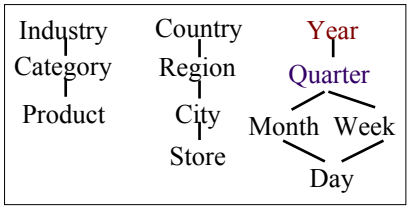
Βασικές Αλγεβρικές Πράξεις

- 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 (π.χ month σε day)
- Οριζόντιος Τεμαχισμός (Slice): (σχεσιακή) επιλογή
- Κάθετος Τεμαχισμός (Dice): (σχεσιακή) προβολή
- Περιστροφή (Pivot): αναδιάταξη της 2D προβολής του πολυδιάστατου κύβου στην οθόνη

Βασικές Αλγεβρικές Πράξεις

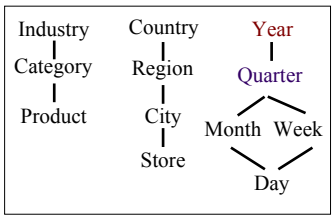
Roll-up

- 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 (Store) παράγουν τις συνολικές πωλήσεις σε επίπεδο πόλης (City) και αυτές με τη σειρά τους με ένα ακόμα roll-up παράγουν τις πωλήσεις σε επίπεδο περιοχής (Region).



Βασικές Αλγεβρικές Πράξεις

Roll-up (Παράδειγμα)



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

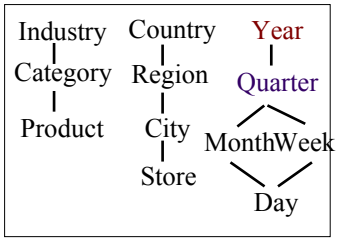
Χρόνος: Επίπεδο Year

SUM(Sales volumes)

Βασικές Αλγεβρικές Πράξεις

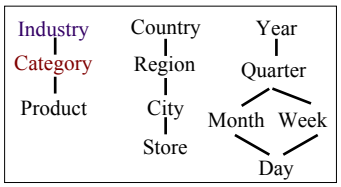
Drill-Down

- 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 (Region) και παίρνουμε τις αναλυτικές πωλήσεις ανά πόλη (City) και μετά τις πωλήσεις ανά κατάστημα (Store).



Βασικές Αλγεβρικές Πράξεις

Drill-down (Παράδειγμα)



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

Βασικές Αλγεβρικές Πράξεις

Περιστροφή (Pivot)

- Εναλλαγή των γραμμών και των στηλών του κύβου, όπως αυτός παρουσιάζεται στην οθόνη
- Δεν απαιτείται κανένας νέος υπολογισμός στη ΒΔ

Βασικές Αλγεβρικές Πράξεις

Pivot (Παράδειγμα)

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

Βασικές Αλγεβρικές Πράξεις

Οριζόντιος (slice) και Κάθετος (dice) Τεμαχισμός

- Slice : Επιλογή συγκεκριμένων τιμών σε κάποια διάσταση (select)
  - Π.χ., διώξε το Store 2 από τα καταστήματα και τις βιομηχανίες Clothing και Cosmetics
- Dicing : Σβήσιμο μιας ολόκληρης διάστασης (project)
  - Π.χ., από ένα κύβο πωλήσεων ανά προϊόν, ημερομηνία και περιοχή, να δειχθεί ο μέσος όρος πωλήσεων ανά προϊόν και ημερομηνία.

Βασικές Αλγεβρικές Πράξεις

Slice&Dice (Παράδειγμα)

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

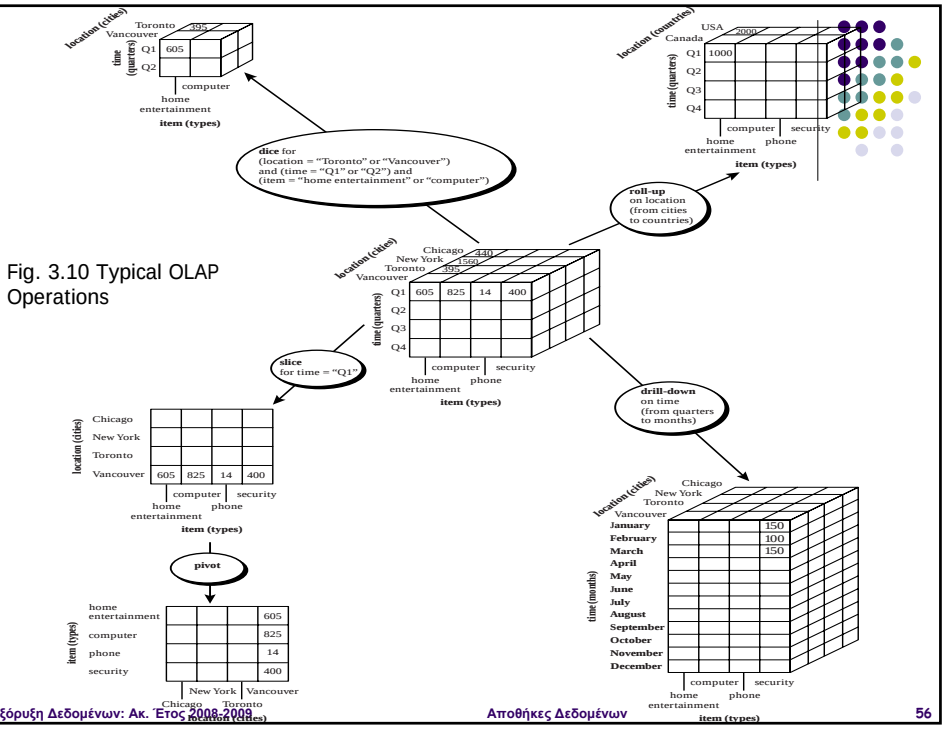


Fig. 3.10 Typical OLAP Operations

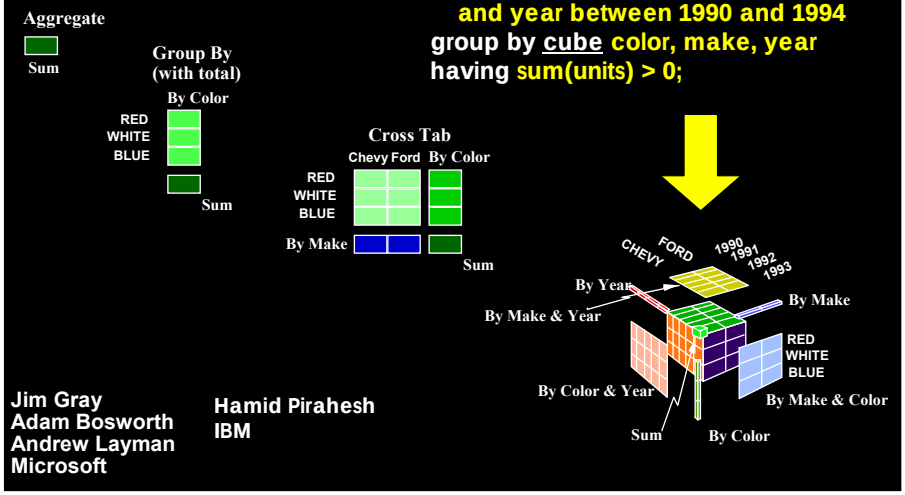
Rollup & Cube

- Τελεστής **Rollup**
  - **group by rollup product, store, city**
    - group by product, store, city
    - group by store, city
    - group by city
- 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 πολλούς μικρούς πίνακες, αλλά έναν πίνακα με εγγραφές με NULL όπου δεν αντιστοιχεί τιμή

Τελεστές Rollup και Cube

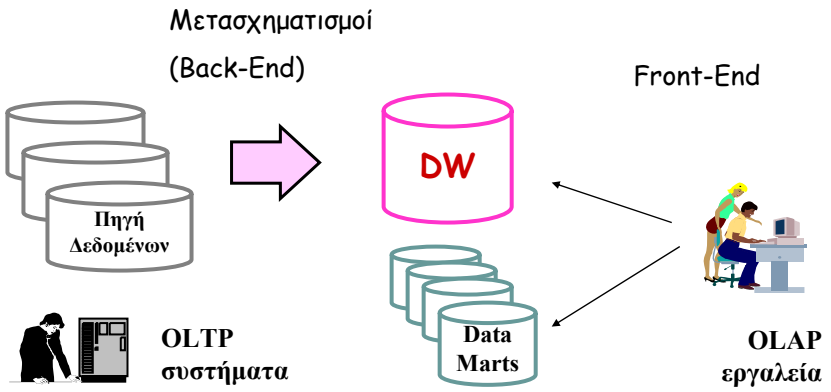
```
select color, make, year, sum(units)
from car_sales
where make in {"chevy", "ford"}
  and year between 1990 and 1994
group by cube color, make, year
having sum(units) > 0;
```

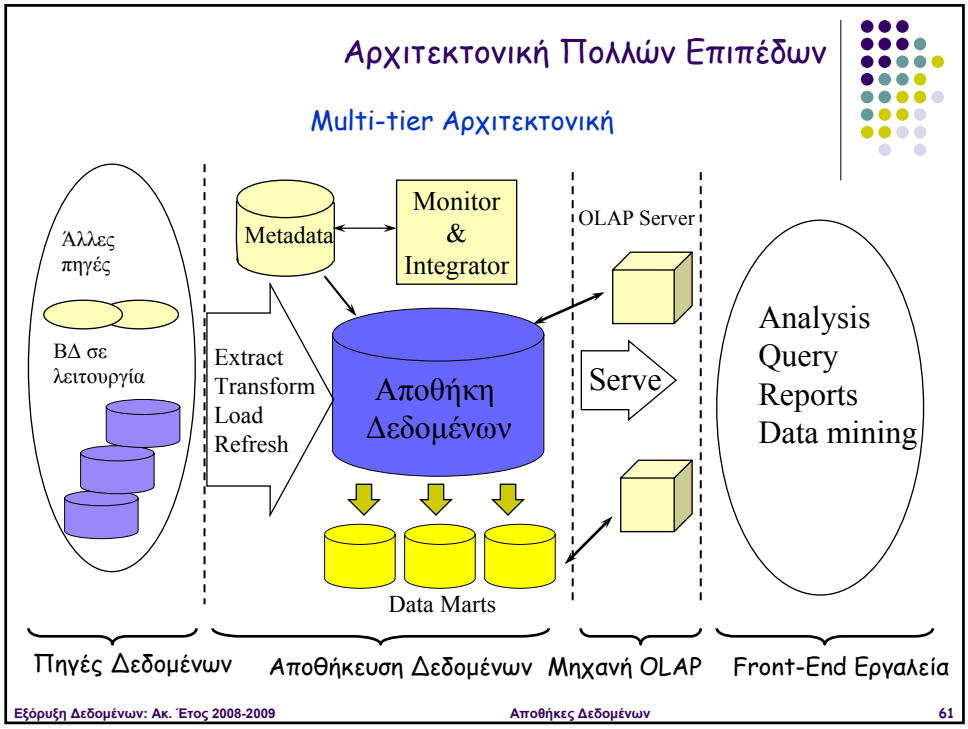


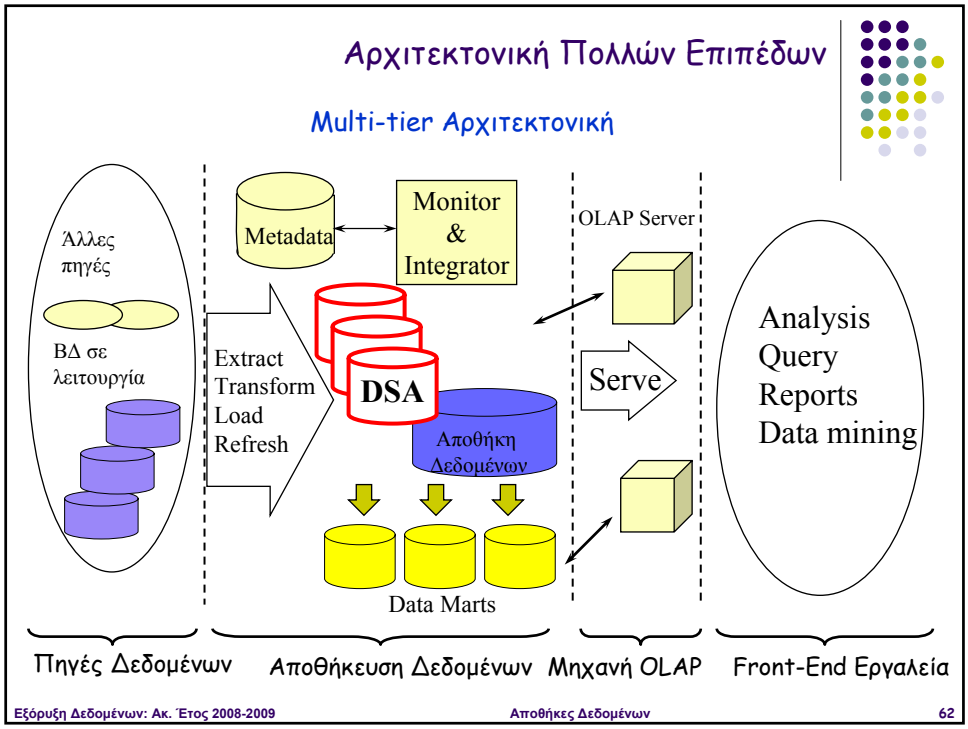
Jim Gray
Adam Bosworth
Andrew Layman
Microsoft

Hamid Pirahesh
IBM

Αρχιτεκτονική




Αρχιτεκτονική Πολλών Επιπέδων
Multi-tier Αρχιτεκτονική
Άλλες πηγές
ΒΔ σε λειτουργία
Metadata
Monitor & Integrator
Extract
Transform
Load
Refresh
Αποθήκη Δεδομένων
Data Marts
OLAP Server
Serve
Analysis
Query
Reports
Data mining
Πηγές Δεδομένων
Αποθήκευση Δεδομένων
Μηχανή OLAP
Front-End Εργαλεία
Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009
Αποθήκες Δεδομένων
61


Αρχιτεκτονική Πολλών Επιπέδων
Multi-tier Αρχιτεκτονική
Άλλες πηγές
ΒΔ σε λειτουργία
Metadata
Monitor & Integrator
Extract
Transform
Load
Refresh
DSA
Αποθήκη Δεδομένων
Data Marts
OLAP Server
Serve
Analysis
Query
Reports
Data mining
Πηγές Δεδομένων
Αποθήκευση Δεδομένων
Μηχανή OLAP
Front-End Εργαλεία
Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009
Αποθήκες Δεδομένων
62

Αρχιτεκτονικές Μονάδες

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

Λεξικό Μεταπληροφορίας

Τα μετα-δεδομένα είναι τα δεδομένα που ορίζουν τα αντικείμενα της αποθήκης δεδομένων. Περιέχουν

- Περιγραφή της δομής της αποθήκης δεδομένων
  Σχήμα, όψεις, διαστάσεις, ιεραρχίες, την τοποθεσία των data mart και το περιεχόμενο τους, κλπ
- Λειτουργικά μεταδεδομένα
  data lineage (την ιστορία των δεδομένων που μεταφέρθηκαν και ποιοι μετασχηματισμοί χρησιμοποιήθηκαν), στοιχεία για το πόσο ενημερωμένα/πρόσφατα είναι, πληροφορία επίβλεψης (monitoring) για τη λειτουργία της αποθήκης (στατιστικά στοιχεία λειτουργίας, error reports, audit trails)
- Τους αλγορίθμους που χρησιμοποιηθήκαν για τις περιλήψεις
- Την απεικόνιση του λειτουργικού περιβάλλοντος στην αποθήκη δεδομένων
- Δεδομένα σχετικά με την απόδοση του συστήματος
- Business data
  Πολιτικές χρέωσης, ιδιοκτησίας δεδομένων, κλπ

Αρχιτεκτονική: Μετασχηματισμοί

Back-End Εργαλεία

- **ETL (Extract-Transform-Load) εφαρμογές**: Εφαρμογές που εκτελούν τις διαδικασίες
  - Εξαγωγής,
  - μεταφοράς,
  - μετασχηματισμού,
  - καθαρισμού και
  - φόρτωσης των δεδομένων
  - από τις πηγές στην Αποθήκη Δεδομένων.

Front-End Εργαλεία

- **Εφαρμογές Ανάλυσης**: Εφαρμογές παραγωγής αναφορών, OLAP , DSS, Data Mining

Μετασχηματισμοί

Back-End Εργαλεία

- Data extraction
  - Φέρε δεδομένα από πολλαπλές, ετερογενείς και εξωτερικές πηγές
- Data cleaning
  - Εντοπισμός λαθών στα δεδομένα και διόρθωση τους όταν είναι δυνατόν

  Παραδείγματα: Δεδομένα που παραβιάζουν τους κανόνες της βάσης: διπλοεγγραφές, παραβιάσεις πρωτεύοντος ή ξένου κλειδιού, τιμές εκτός ορίων, παραβιάσεις λογικών κανόνων, κλπ Συνώνυμα και συγκρούσεις Ελλιπή δεδομένα
  - Ομογενοποίηση κλειδιού
- Data transformation
  - Μετατροπή των δεδομένων από το τοπικό format στο format της αποθήκης

Μετασχηματισμοί

Back-End Εργαλεία

- Load
  - Ταξινόμηση, δημιουργία περίληψης, ενοποίηση (consolidate), υπολογισμός όψεων, έλεγχος integrity, δημιουργία ευρετηρίων και διαμερίσεων

    Η ενημέρωση / εισαγωγή δεδομένων στην πράξη δε γίνεται μέσω SQL, συνήθως μέσω εργαλείων batch loading που διαθέτουν όλα τα ΣΔΒΔ
- Refresh
  - Μετέφερε τις τροποποιήσεις από τις πηγές δεδομένων στην αποθήκη δεδομένων

Εργαλεία για την Υποστήριξη Αποφάσεων

Front-End Εργαλεία

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