ΒΔ για Λήψη Αποφάσεων

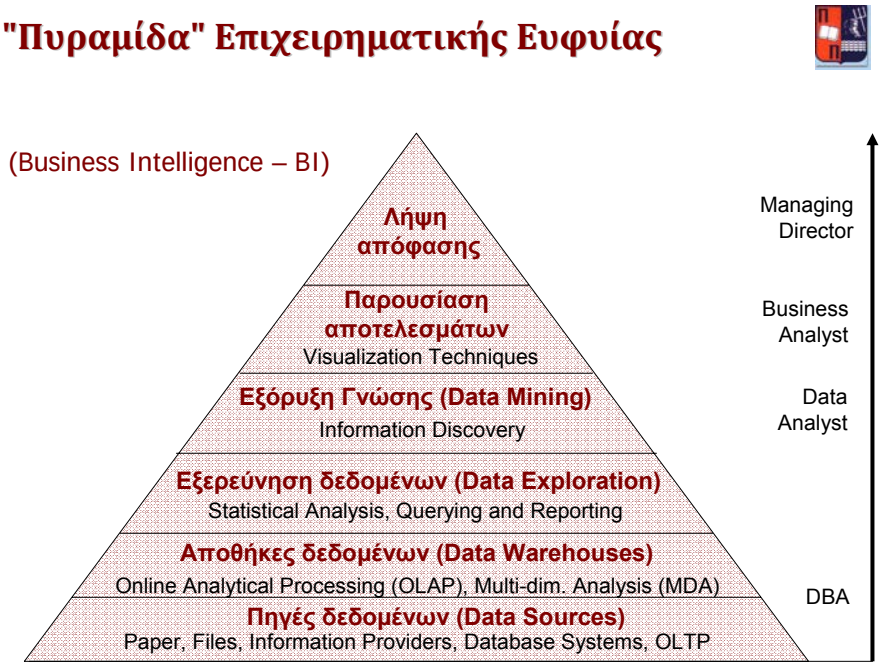
- Ανάλυση Δεδομένων (Data Analysis)
  - Άμεση Αναλυτική Επεξεργασία (OLAP)
  - Λειτουργίες Συνάθροισης στην SQL (windowing, ranking)
- Οργάνωση ιστορικής πληροφορίας σε Αποθήκες Δεδομένων (Data Warehousing)
- Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

Εργαστήριο Πληροφοριακών Συστημάτων, Παν/μιο Πειραιώς (http://infolab.cs.unipi.gr/)
έκδοση: Φεβ. 2010

Πηγές

- Το κύριο μέρος των διαφανειών προέρχεται από:
  - Silberschatz et al., “Database System Concepts”, 4/e.
- Οι μεταφράσεις των Αγγλικών όρων στα Ελληνικά βασίστηκαν κυρίως στο:
  - Π. Βασιλειάδης, Σ. Σκιαδόπουλος, Τ. Σελλής «Θέματα Μοντελοποίησης και Επερώτησης Πολυδιάστατων Βάσεων Δεδομένων», 1ο Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων, Αθήνα, 22-23 Ιουλίου 2002, σελ. 3-22.

"Πυραμίδα" Επιχειρηματικής Ευφυίας



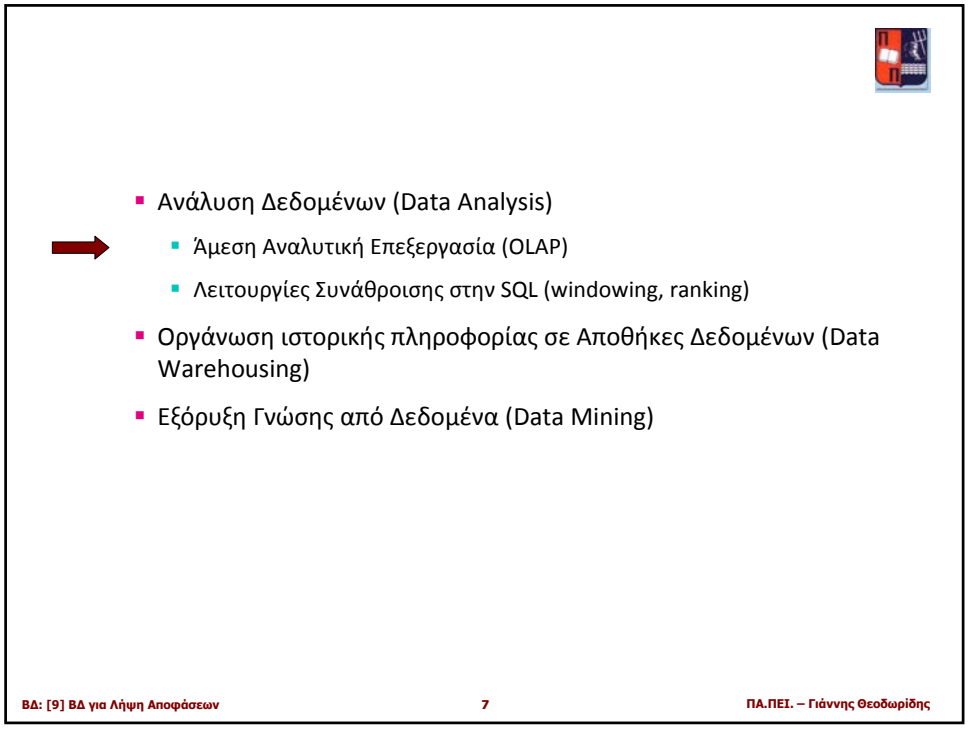
- Ανάλυση Δεδομένων (Data Analysis)
  - Άμεση Αναλυτική Επεξεργασία (OLAP)
  - Λειτουργίες Συνάθροισης στην SQL (windowing, ranking)
- Οργάνωση ιστορικής πληροφορίας σε Αποθήκες Δεδομένων (Data Warehousing)
- Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

Συστήματα Υποστήριξης Αποφάσεων

- **Τα Συστήματα Υποστήριξης Αποφάσεων – ΣΥΑ (Decision-support systems – DSS)** χρησιμοποιούνται για την υποστήριξη λήψης αποφάσεων στηριγμένων πάνω στα επιχειρησιακά δεδομένα που συγκεντρώνονται από **On-Line Transaction Processing (OLTP)** συστήματα.
- Παραδείγματα αποφάσεων:
  - Τι ποσότητες από κάθε προϊόν να έχουμε στην αποθήκη;
  - Σε ποιους από τους πελάτες μας να στείλουμε διαφημιστικό υλικό;
- Παραδείγματα δεδομένων που χρησιμοποιούνται για τη λήψη αποφάσεων
  - συναλλαγές πωλήσεων προϊόντων
  - στοιχεία πελατών (ηλικία, φύλο, εισόδημα κλπ.)

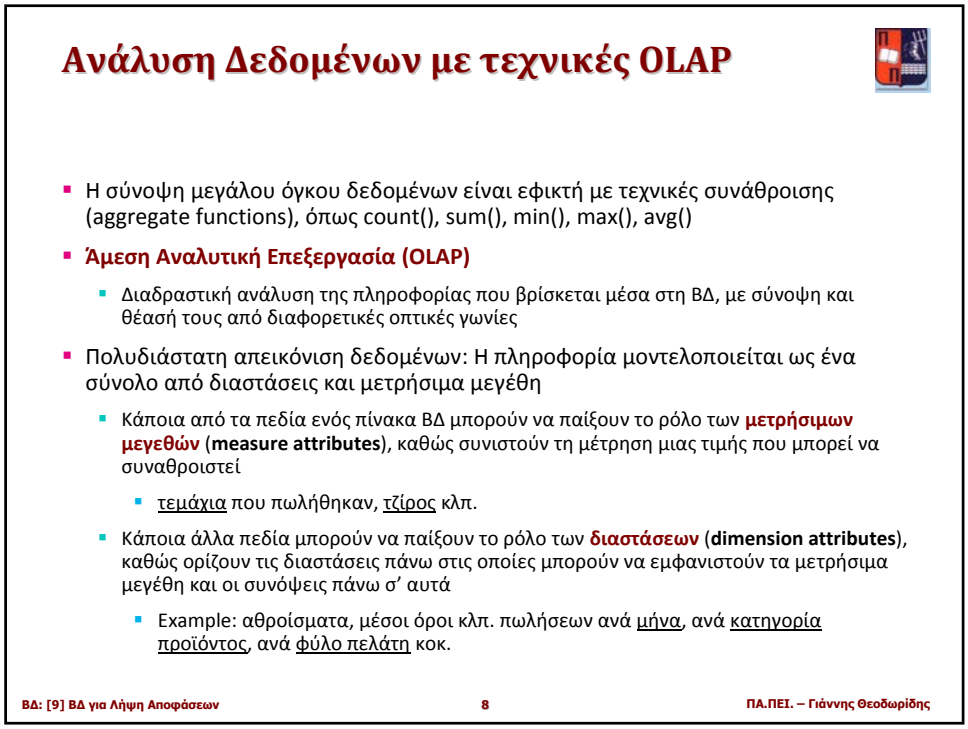
Εργαλεία Ανάλυσης Δεδομένων

- **Άμεση αναλυτική επεξεργασία δεδομένων (Online Analytical Processing – OLAP)** – πρακτικά, αποτελούν σύνθετα SQL αιτήματα
  - Παράδειγμα: να υπολογιστεί το σύνολο των πωλήσεων στο τελευταίο τρίμηνο του έτους ανά κατηγορία προϊόντων και περιοχή πώλησης, και να συγκριθεί με τις αντίστοιχες τιμές του προηγούμενου έτους.
  - Για να γίνει η σχετική ανάλυση, μπορεί να συνδεθεί ένα πακέτο στατιστικής ανάλυσης (π.χ. SPSS) με το ΣΔΒΔ στο οποίου διατηρούνται τα δεδομένα
- **Αποθήκες Δεδομένων – ΑΔ (Data Warehouse – DW)** – οργανώνουν και αποθηκεύουν κεντρικά ιστορική πληροφορία για τους σκοπούς της ανάλυσης
  - Πολύ χρήσιμες για μεγάλους οργανισμούς που συγκεντρώνουν δεδομένα από πολλαπλές τοποθεσίες
- **Τεχνικές εξόρυξης γνώσης από δεδομένα (Data mining)** – επιδιώκουν την ανακάλυψη 'κρυμμένης' γνώσης, σε μορφή κανόνων και προτύπων, από μεγάλες ΒΔ/ΑΔ.



- Ανάλυση Δεδομένων (Data Analysis)
  - Άμεση Αναλυτική Επεξεργασία (OLAP)
  - Λειτουργίες Συνάθροισης στην SQL (windowing, ranking)
- Οργάνωση ιστορικής πληροφορίας σε Αποθήκες Δεδομένων (Data Warehousing)
- Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

ΒΔ: [9] ΒΔ για Λήψη Αποφάσεων 7 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης



Ανάλυση Δεδομένων με τεχνικές OLAP

- Η σύνοψη μεγάλου όγκου δεδομένων είναι εφικτή με τεχνικές συνάθροισης (aggregate functions), όπως count(), sum(), min(), max(), avg()
- **Άμεση Αναλυτική Επεξεργασία (OLAP)**
  - Διαδραστική ανάλυση της πληροφορίας που βρίσκεται μέσα στη ΒΔ, με σύνοψη και θέασή τους από διαφορετικές οπτικές γωνίες
- Πολυδιάστατη απεικόνιση δεδομένων: Η πληροφορία μοντελοποιείται ως ένα σύνολο από διαστάσεις και μετρήσιμα μεγέθη
  - Κάποια από τα πεδία ενός πίνακα ΒΔ μπορούν να παίξουν το ρόλο των **μετρήσιμων μεγεθών** (**measure attributes**), καθώς συνιστούν τη μέτρηση μιας τιμής που μπορεί να συναθροιστεί
    - τεμάχια που πωλήθηκαν, τζίρος κλπ.
  - Κάποια άλλα πεδία μπορούν να παίξουν το ρόλο των **διαστάσεων** (**dimension attributes**), καθώς ορίζουν τις διαστάσεις πάνω στις οποίες μπορούν να εμφανιστούν τα μετρήσιμα μεγέθη και οι συνόψεις πάνω σ' αυτά
    - Example: αθροίσματα, μέσοι όροι κλπ. πωλήσεων ανά μήνα, ανά κατηγορία προϊόντος, ανά φύλο πελάτη κοκ.

ΒΔ: [9] ΒΔ για Λήψη Αποφάσεων 8 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Cross-tabs

- Ένας **πίνακας διασταύρωσης δεδομένων** (**cross-tab**) σχηματίζεται ως εξής:
  - Οι τιμές ενός από τα πεδία-διαστάσεις αποτελούν τις επικεφαλίδες των γραμμών του πίνακα, ενώ οι τιμές ενός άλλου από τα πεδία-διαστάσεις αποτελούν τις επικεφαλίδες των στηλών του πίνακα
    - Τα υπόλοιπα πεδία-διαστάσεις εμφανίζονται εκτός πίνακα
  - Παράδειγμα:
    - Πίνακας ΒΔ: sales
    - Διαστάσεις: item-name, color, size
    - Μετρήσιμο μέγεθος: number

size: **all**

color

item-name	dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

Cross-tabs

- Οι πίνακες cross-tab μπορούν να υλοποιηθούν ως σχεσιακοί πίνακες
  - 3 στήλες (για τις 2 διαστάσεις και το 1 μετρήσιμο μέγεθος)
  - χρήση της ειδικής τιμής **all** όπου υπάρχει συνάθροιση
    - η SQL:1999 χρησιμοποιεί τιμές null αντί της all

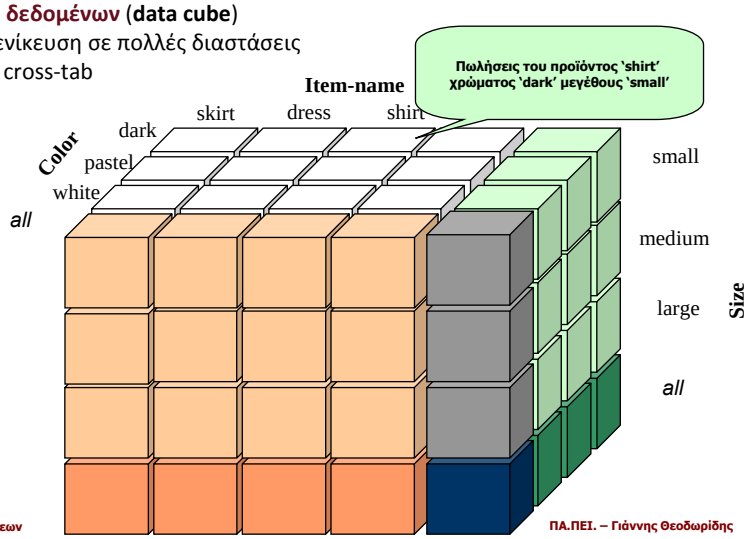
color

item-name	dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

item-name	*color*	*number*
skirt	dark	8
skirt	pastel	35
skirt	white	10
skirt	**all**	53
dress	dark	20
dress	pastel	10
dress	white	5
dress	**all**	35
shirt	dark	14
shirt	pastel	7
shirt	white	28
shirt	**all**	49
pant	dark	20
pant	pastel	2
pant	white	5
pant	**all**	27
all	dark	62
all	pastel	54
all	white	48
all	**all**	164

Πολυδιάστατος κύβος δεδομένων

- Ένας **κύβος δεδομένων** (**data cube**) αποτελεί γενίκευση σε πολλές διαστάσεις του πίνακα cross-tab



OLAP πάνω σε κύβους δεδομένων

size: **all**

color

item-name	dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

- Η εναλλαγή των διαστάσεων στην εμφάνιση ενός cross-tab ονομάζεται **pivoting** («περιστροφή»)
  - Παράδειγμα: ο αναλυτής βλέπει cross-tab πάνω στις διαστάσεις item-name, color και θέλει να δει cross-tab πάνω στις διαστάσεις item-name, size.
- Η επιλογή ορισμένων μόνο στοιχείων ενός cross-tab ονομάζεται **slicing** («τεμαχισμός»)
  - Παράδειγμα: ο αναλυτής θέλει να δει cross-tab πάνω στις διαστάσεις item-name, color για συγκεκριμένο όμως μέγεθος (π.χ. size='large') αντί της συνάθροισης όλων των μεγεθών.

OLAP πάνω σε κύβους δεδομένων

size: **all**

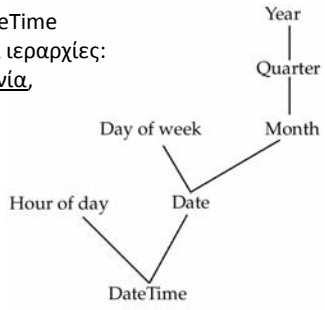
color

item-name	dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

- Η μετάβαση από μεγαλύτερη σε μικρότερη λεπτομέρεια / κλίμακα εμφάνισης πληροφορίας ονομάζεται **roll-up** («συναθροιστική άνοδος»)
  - Παράδειγμα: ο αναλυτής βλέπει (πλήρες) cross-tab πάνω στις διαστάσεις item-name, color και θέλει να δει (συνοπτικό, color = 'all') cross-tab πάνω στη διάσταση item-name.
- Η αντίθετη λειτουργία (μετάβαση από μικρότερη σε μεγαλύτερη λεπτομέρεια) ονομάζεται **drill-down** («αναλυτική κάθοδος»)
  - Παράδειγμα: ο αναλυτής βλέπει (συνοπτικό, color = 'all') cross-tab πάνω στη διάσταση item-name και θέλει να δει αναλυτικά τι πωλήθηκε για κάθε χρώμα, δηλαδή cross-tab πάνω στις διαστάσεις item-name, color .

Ιεραρχίες πάνω στις διαστάσεις

- Μπορούμε να ορίσουμε **ιεραρχίες** (δενδρικές δομές) πάνω στις διαστάσεις ώστε να εμφανίζουμε την πληροφορία σε διαφορετικά επίπεδα λεπτομέρειας, π.χ.
  - πάνω στη διάσταση DateTime μπορούν να οριστούν οι ιεραρχίες: ώρα (00 – 23), ημερομηνία, ημέρα (Δε – Κυ), μήνας, τρίμηνο, έτος
  - πάνω στη διάσταση Location μπορούν να οριστούν οι ιεραρχίες (υποθέστε μια πολυ-εθνική επιχείρηση): πόλη, νομός, χώρα, ήπειρος



a) Time Hierarchy

b) Location Hierarchy

Cross-tabs με ιεραρχίες

- Οι πίνακες Cross-tabs μπορούν εύκολα να ενσωματώσουν ιεραρχίες
  - Οπότε μπορούν να γίνουν λειτουργίες drill-down ή roll-up πάνω σε ιεραρχίες

category	*item-name*					
		dark	pastel	white	total	
womenswear	skirt	8	8	10	53	
	dress	20	20	5	35	
	subtotal	28	28	15		88
menswear	pants	14	14	28	49	
	shirt	20	20	5	27	
	subtotal	34	34	33		76
total		62	62	48		164

Υλοποίηση κύβων δεδομένων

- Τα πρώτα συστήματα OLAP αποθήκευαν τους κύβους δεδομένων ως πολυδιάστατους πίνακες κύριας μνήμης (arrays). Ονομάζονται **multidimensional OLAP** (MOLAP) συστήματα
- Αργότερα προέκυψαν υλοποιήσεις σε σχεσιακούς πίνακες ΒΔ (relations). Ονομάζονται **relational OLAP** (ROLAP) συστήματα
- Σήμερα, υπάρχουν υβριδικές υλοποιήσεις που αποθηκεύουν
  - τις συναθροίσεις των υψηλών επιπέδων σε arrays
  - τα λεπτομερή δεδομένα και τις συναθροίσεις χαμηλών επιπέδων σε relations
- ... και ονομάζονται **hybrid OLAP** (HOLAP) συστήματα.

- Ανάλυση Δεδομένων (Data Analysis)
  - Άμεση Αναλυτική Επεξεργασία (OLAP)
  - Λειτουργίες Συνάθροισης στην SQL (windowing, ranking)
- Οργάνωση ιστορικής πληροφορίας σε Αποθήκες Δεδομένων (Data Warehousing)
- Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

Επεκτάσεις SQL για ανάλυση δεδομένων

- Η ανάλυση δεδομένων έχει οδηγήσει σε νέες απαιτήσεις ερωτημάτων πάνω στη ΒΔ:
  - κύβοι δεδομένων και πολύπλοκες λειτουργίες συνάθροισης
    - ολιστικές λειτουργίες, π.χ. διάμεσος (median), διακύμανση (variance), συσχέτιση (correlation), παλινδρόμηση (regression), ...
    - λειτουργίες 'παραθύρου', π.χ. κινούμενοι μέσοι όροι
  - αιτήματα κατάταξης (π.χ. «να ανατεθεί ένας α/α σε κάθε φοιτητή με βάση τη συνολική του βαθμολογία») και 'παραθύρου' (π.χ. «να συγκριθεί ο τζίρος του φετινής χρονιάς
- Η SQL:1999 παρέχει επεκτάσεις προς αυτή την κατεύθυνση:
  - λειτουργίες **cube** και **rollup** για πολυδιάστατη ανάλυση δεδομένων
  - λειτουργία **rank** για αιτήματα κατάταξης
  - λειτουργίες 'παραθύρου' (**windowing**)

Λειτουργίες cube και rollup

- Η λειτουργία **cube()** ουσιαστικά είναι ένα σύνολο από λειτουργίες group by (μια για κάθε διακριτή τιμή των πεδίων – διαστάσεων)
- Π.χ. η εντολή SQL

```
select item-name, color, size, sum(number)
from sales
group by cube(item-name, color, size)
```

  υπολογίζει $2^3 = 8$ διαφορετικές ομαδοποιήσεις πάνω στον πίνακα sales:

  { (item-name, color, size), (item-name, color), (item-name, size), (color, size), (item-name), (color), (size), ( ) }

  όπου το κενό σύνολο ( ) υποδηλώνει κενή λίστα group by.
- Σε κάθε ομαδοποίηση, τα χαρακτηριστικά που δεν συμμετέχουν δηλώνονται στο αποτέλεσμα με τιμές null.

Λειτουργίες cube και rollup

- Η λειτουργία **rollup()** δημιουργεί κι αυτή ένα σύνολο από ομαδοποιήσεις, με τη διαφορά ότι σχηματίζονται ομάδες μόνο με βάση τα προθέματα της λίστας χαρακτηριστικών
- Π.χ. η εντολή SQL

```
select item-name, color, size, sum(number)
from sales
group by rollup(item-name, color, size)
```

  υπολογίζει 3+1 = 4 διαφορετικές ομαδοποιήσεις πάνω στον πίνακα sales:

  { (item-name, color, size), (item-name, color), (item-name), ( ) }
- Προφανώς, λιγότερες από αυτές που παράγει η λειτουργία cube: για n χαρακτηριστικά, n+1 έναντι 2^n !!

Λειτουργίες cube και rollup

- Οι λειτουργίες rollup και cube μπορούν να εμφανίζονται πολλαπλά σε μια γραμμή group by
  - Καθεμία λειτουργία σχηματίζει ένα σύνολο από ομαδοποιήσεις, και στη συνέχεια υπολογίζεται το καρτεσιανό γινόμενο αυτών των συνόλων
- Π.χ. η εντολή SQL

```
select item-name, color, size, sum(number)
from sales
group by rollup(item-name), rollup(color, size)
```

υπολογίζει 2 X 3 = 6 διαφορετικές ομαδοποιήσεις πάνω στον πίνακα sales

{item-name, ()} X {(color, size), (color), ()}

= { (item-name, color, size), (item-name, color), (item-name), (color, size), (color), ( ) }

Λειτουργίες κατάταξης

- Η λειτουργία κατάταξης **rank()** συνδυάζεται με μια διάταξη που ορίζεται στη γραμμη order by.
- Π.χ. αν υποθέσουμε τον πίνακα student-marks(student-id, mark) η εντολή SQL

```
select student-id, rank( ) over (order by marks desc) as s-rank
from student-marks
order by s-rank
```

εμφανίζει τους φοιτητές με βάση την κατάταξη που τους δίνει η τιμή του πεδίου mark

- Η λειτουργία rank μπορεί να αφήσει κενά στην αρίθμηση, π.χ. αν δύο φοιτητές ισοβαθμούν στην 1η θέση, θα έχουν και οι δύο τιμή κατάταξης 1 ενώ η επόμενη τιμή θα είναι η 3
  - Αυτό το αποφεύγουμε με τη λειτουργία dense_rank()

Λειτουργίες κατάταξης

- Η κατάταξη μπορεί να συνδυαστεί με διαμέριση των εγγραφών ενός πίνακα.
- Παράδειγμα: "Βρες την κατάταξη των φοιτητών μέσα σε κάθε τμήμα"

```
select student-id, section,
     rank ( ) over (partition by section order by marks desc) as sec-rank
from student-marks, student-section
where student-marks.student-id = student-section.student-id
order by section, sec-rank
```

- Η γραμμή select μπορεί να περιέχει πολλαπλές λειτουργίες rank
- Εάν υπάρχει γραμμή group by, η λειτουργία rank ενεργοποιείται μετά την ομαδοποίηση

Λειτουργίες 'παραθύρου'

- Παράδειγμα: δοθέντος του τζίρου κάθε ημέρας, για κάθε ημέρα να υπολγιστεί ο μ.ο. του τζίρου της ίδιας, της προηγούμενης και της επόμενης ημέρας
- Αυτές είναι ερωτήσεις τύπου **κινούμενου μέσου** (moving average) και χρησιμοποιούνται στη στατιστική για εξομάλυνση των τιμών, απάλειψη 'θορύβου' κλπ.
- Μια βασική διαφορά από το group by: η ίδια εγγραφή συμμετέχει σε περισσότερα από ένα 'παράθυρα'
- Πώς γίνεται αυτό στην SQL:
  - Καταρχήν διατάσσονται οι εγγραφές, κατόπιν ορίζεται το παράθυρο, και στη συνέχεια υλοποιείται η όποια λειτουργία συνάθροισης.
  - Π.χ. αν υποθέσουμε τον πίνακα sales(date, value) η εντολή SQL υλοποιεί το ανωτέρο παράδειγμα

```
select date, avg(value) over
    (order by date between rows 1 preceding and 1 following)
from sales
```

Λειτουργίες 'παραθύρου'

- Παραδείγματα εναλλακτικών 'παραθύρων':
  - between rows unbounded preceding and current
  - rows unbounded preceding
  - range between 10 preceding and current row
- Η λειτουργία 'παραθύρου' μπορεί να συνδυαστεί με διαμέριση των εγγραφών ενός πίνακα.
- Παράδειγμα: έστω ο πίνακας transaction(account-number, date-time, value) με τις δοσοληψίες των τραπεζικών λογαριασμών. Τι υλοποιεί η ακόλουθη εντολή SQL;

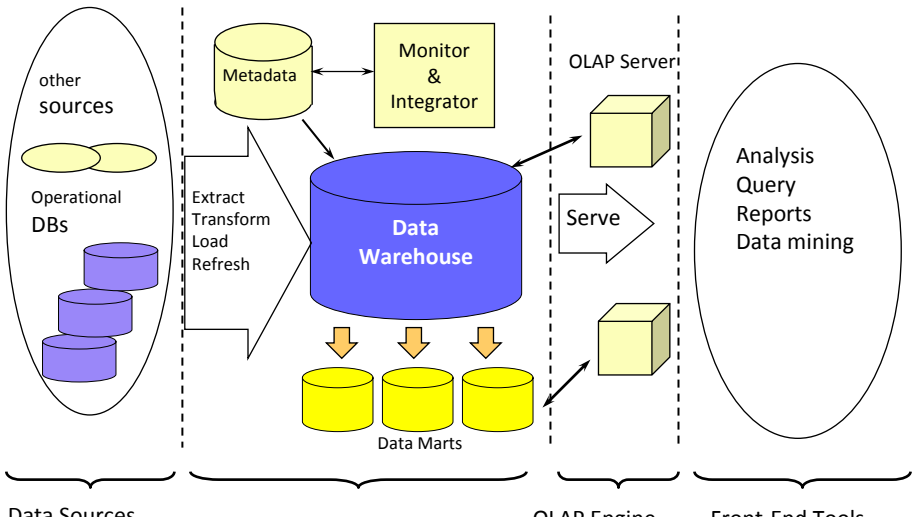
```
select account-number, date-time, sum(value) over
    (partition by account-number
     order by date-time
     between rows unbounded preceding and current) as balance
from transaction
order by account-number, date-time
```

- Ανάλυση Δεδομένων (Data Analysis)
  - Άμεση Αναλυτική Επεξεργασία (OLAP)
  - Λειτουργίες Συνάθροισης στην SQL (windowing, ranking)
- Οργάνωση ιστορικής πληροφορίας σε Αποθήκες Δεδομένων (Data Warehousing)
- Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

Αποθήκες Δεδομένων

- Οι μεγάλοι οργανισμοί συχνά έχουν μεγάλο όγκο δεδομένων να τηρείται σε διαφορετικές τοποθεσίες (π.χ. υποκαταστήματα), με διαφορετικά OLTP συστήματα και διαφορετικά σχήματα ΒΔ
- Οι διαδικασίες λήψης αποφάσεων απαιτούν ανάλυση των δεδομένων του οργανισμού, κοιτώντας αρκετά πίσω στο χρόνο (δεν αρκεί η τρέχουσα πληροφορία)
- Μια **αποθήκη δεδομένων – ΑΔ** (**data warehouse – DW**) είναι μια ΒΔ αρχειακής μορφής με πληροφορία που συλλέγεται από πολλαπλές πηγές και συγκεντρώνεται σε μία τοποθεσία κάτω από ένα ενιαίο σχήμα ΒΔ
  - Απλοποιεί τις λειτουργίες ανάλυσης δεδομένων, επιτρέπει την ιστορική ανάλυση της πληροφορίας, δεν απασχολεί τα OLTP συστήματα

Αρχιτεκτονική ΑΔ

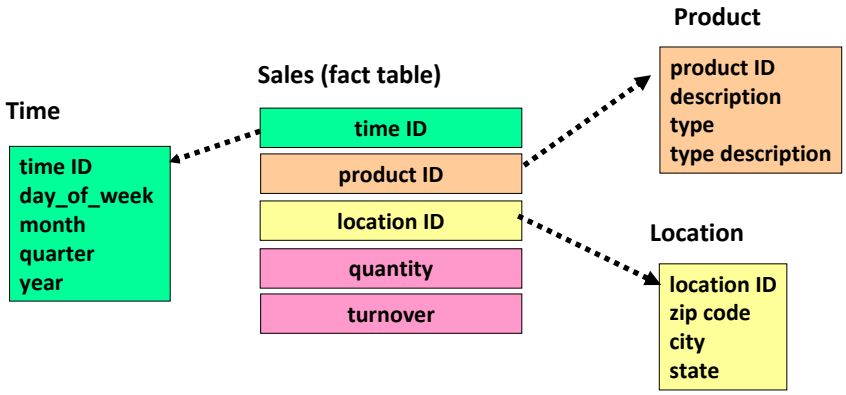


Συστατικά μιας ΑΔ

- Δεδομένα που συλλέγονται περιοδικά από τις πολλαπλές πηγές
  - Δεν υπάρχει λόγος να είναι απόλυτα συγχρονισμένη η πληροφορία μεταξύ των OLTP συστημάτων και της ΑΔ
- ... κάτω από ένα ενιαίο σχήμα ΒΔ
  - Ζητήματα ενοποίησης σχημάτων ΒΔ
- ... αφού έχουν υποστεί κατάλληλο καθαρισμό / μετασχηματισμό
  - Καθαρισμός δεδομένων (συμπλήρωση ελλιπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών)
  - Μετασχηματισμός δεδομένων (εξομάλυνση, κανονικοποίηση τιμών)
  - Μείωση δεδομένων (μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων)
  - Διακριτοποίηση δεδομένων (από συνεχείς σε διακριτές τιμές ή διαστήματα τιμών)

Σχεδίαση ΑΔ

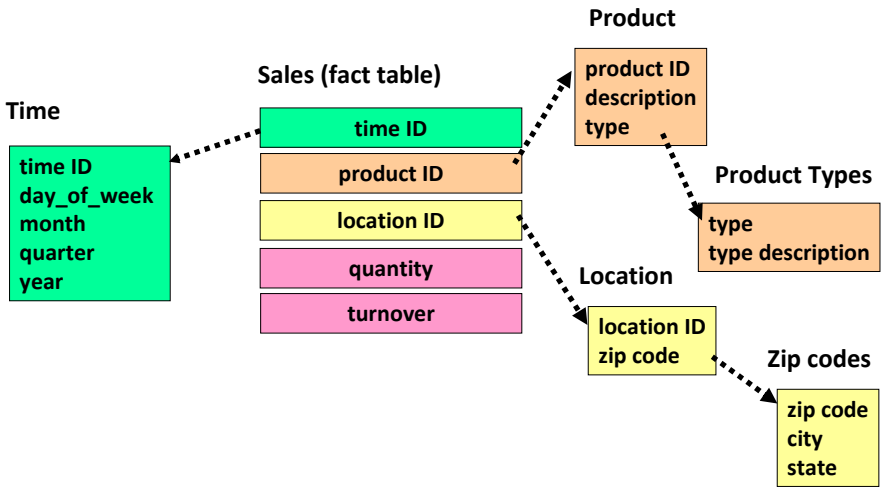
- Ένα τυπικό σχήμα ΑΔ: ένας **πίνακας γεγονότων** (**fact table**) που περιβάλλεται από **πίνακες διαστάσεων** (**dimension tables**)



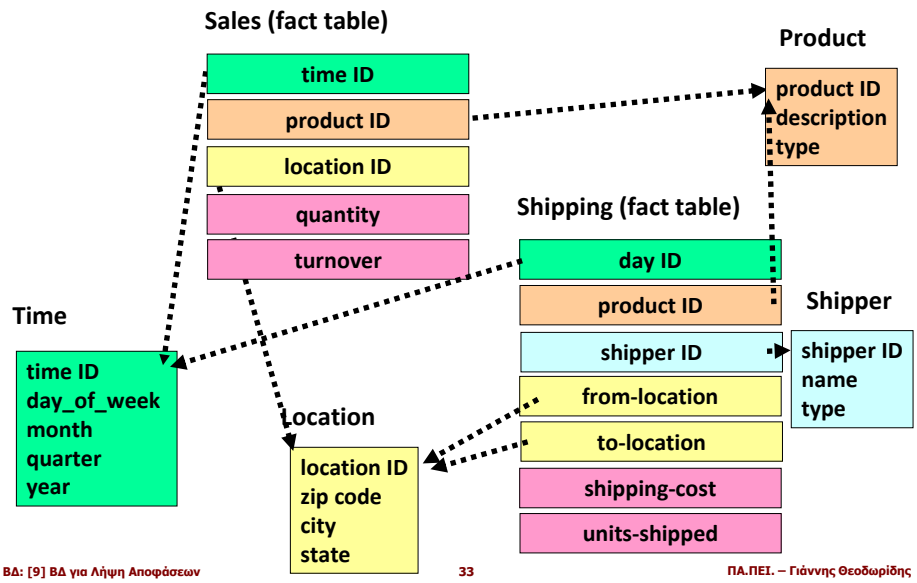
Σχεδίαση ΑΔ

- Συνήθως τα δεδομένα μιας ΑΔ είναι πολυδιάστατα (άρα, πολλοί πίνακες διαστάσεων) με πίνακες γεγονότων πάρα πολλών εγγραφών
  - Παραδείγματα διαστάσεων: προϊόν, χρόνος πώλησης, υποκατάστημα πώλησης, πελάτης, ...
  - Παραδείγματα μετρήσιμων μεγεθών: τεμάχια που πωλήθηκαν, αξία, ...
- Οι τιμές των διαστάσεων συνήθως κωδικοποιούνται με μικρές ακέραιες τιμές (id's) που παραπέμπουν στους πίνακες διαστάσεων για την πλήρη πληροφορία
  - Το ανωτέρω ονομάζεται **σχήμα αστέρα** (**star schema**)
- Πιο σύνθετα σχήματα ΑΔ
  - **Σχήμα χιονονιφάδας** (**snowflake schema**): πίνακες διαστάσεων πολλαπλών επιπέδων
  - **Αστερισμός γεγονότων** (**fact constellation**): πολλαπλοί πίνακες γεγονότων

Σχήμα χιονονιφάδας



Αστερισμός γεγονότων

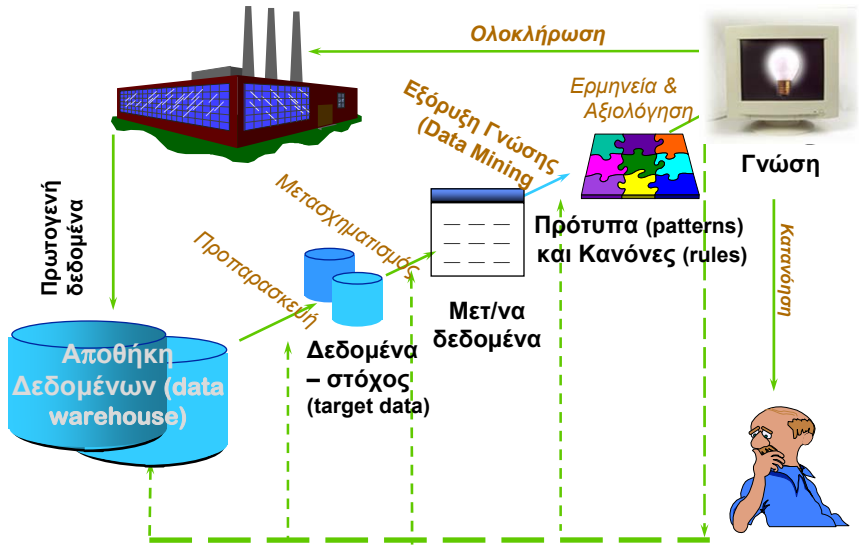


- Ανάλυση Δεδομένων (Data Analysis)
  - Άμεση Αναλυτική Επεξεργασία (OLAP)
  - Λειτουργίες Συνάθροισης στην SQL (windowing, ranking)
- Οργάνωση ιστορικής πληροφορίας σε Αποθήκες Δεδομένων (Data Warehousing)
- Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

Data Mining

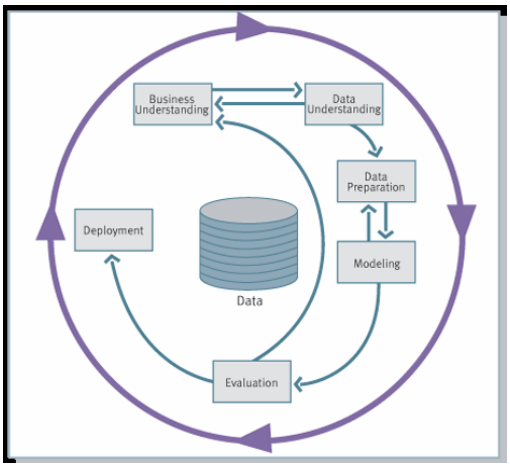
- Η **ανακάλυψη γνώσης από τα δεδομένα** (**knowledge discovery in data – KDD**) είναι η διαδικασία ημι-αυτόματης ανάλυσης μεγάλων ΒΔ/ΑΔ με στόχο την ανακάλυψη 'γνώσης' σε μορφή κανόνων ή προτύπων
  - Διαφέρει από τη στατιστική ανάλυση δεδομένων κυρίως στο ότι χειρίζεται μεγάλους όγκους δεδομένων
- Η 'γνώση' που εξάγεται μπορεί να είναι διαφόρων τύπων
  - Κατηγοριοποίηση / ταξινόμηση δεδομένων, εύρεση ομάδων στα δεδομένα, ανακάλυψη συσχετίσεων κλπ.
- ... σε μορφή κανόνων ή εξισώσεων ή συναρτήσεων κλπ.
  - π.χ. "γυναίκες, νέες σε ηλικία και με εισόδημα άνω των 50.000 είναι πιο πιθανό να αγοράσουν σπορ αυτοκίνητο"
- ... αφού έχει προηγηθεί κάποια προπαρασκευή των δεδομένων
  - π.χ. εντοπισμός και απομόνωση ανωμαλιών ή 'θορύβου' στα δεδομένα

Διαδικασία εξόρυξης γνώσης



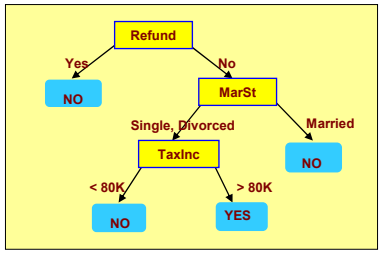
Βήματα εξόρυξης γνώσης

- **μεθοδολογία CRISP-DM**
  - Κατανόηση του οργανισμού
  - Κατανόηση των δεδομένων
  - Προετοιμασία δεδομένων
  - Εύρεση μοντέλων ('γνώσης')
    - με χρήση ποικίλων τεχνικών data mining
  - Αποτίμηση αποτελεσμάτων
  - Εφαρμογή αποτελεσμάτων στις διαδικασίες του οργανισμού



Εφαρμογές εξόρυξης γνώσης

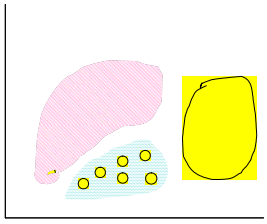
- Πρόβλεψη μελλοντικών τιμών βασισμένη στα ιστορικά δεδομένα
  - Πρόβλεψη αξιοπιστίας ενός πελάτη τράπεζας, με βάση τα χαρακτηριστικά του (εισόδημα, επάγγελμα, ηλικία κλπ.) και την καταγεγραμμένη αξιοπιστία άλλων πελατών
  - Πρόβλεψη διαρροής πελατών στον ανταγωνισμό
  - Πιθανή αναγνώριση απάτης σε μια συναλλαγή (πιστωτική κάρτα, τηλεφωνική κλήση, ...)
- Μηχανισμοί πρόβλεψης:
  - **Κατηγοριοποίηση** / **Ταξινόμηση** (**classification**)
    - Δοθέντος ενός συνόλου εγγραφών ΒΔ που ανήκουν σε γνωστές κλάσεις/κατηγορίες, να γίνει πρόβλεψη για την κλάση μιας νέας εγγραφής
  - **Παλινδρόμηση** (**regression**)
    - Δοθέντος ενός συνόλου τιμών που ταιριάζουν στα αποτελέσματα μιας (άγνωστης) συνάρτησης, να βρεθεί η συνάρτηση και να εφαρμοστεί σε μια νέα εγγραφή



Εφαρμογές εξόρυξης γνώσης

- Περιγραφικά πρότυπα
  - **Κανόνες συσχέτισης** (**association rules**)
    - Εύρεση των προϊόντων που συνήθως αγοράζονται μαζί από τους ίδιους πελάτες. Εφόσον ένας νέος πελάτης αγοράσει ένα τέτοιο προϊόν είναι πολύ πιθανό να αγοράσει και τα υπόλοιπα.
    - Εφαρμογές: λιανεμπόριο, e-shopping, κλπ.
  - **Συστάδες** (**clusters**)
    - εντοπισμός προφίλ πελατών,
    - συγκέντρωση συμβάντων επιδημίας γύρω από ένα μολυσμένο σημείο κοκ.

$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%



Σύνοψη ενότητας

- Η ανάλυση των δεδομένων αποτελεί ισχυρό εργαλείο στη στρατηγική των οργανισμών και τη λήψη αποφάσεων
- Οι βασικοί μηχανισμοί ανάλυσης είναι:
  - η πολυδιάστατη ανάλυση δεδομένων (OLAP) πάνω σε κύβους / αποθήκες δεδομένων και
  - η εξόρυξη γνώσης από δεδομένα (data mining)
    - Ανάλυση συστάδων, κατηγοριοποίηση, ανακάλυψη συσχετίσεων κλπ.

Προτεινόμενη βιβλιογραφία

Διαθέσιμη στο ./courses/dwdm/readings/

- Γενικά για Αποθήκες Δεδομένων και Εξόρυξη Γνώσης
  - S. Chaudhuri et al.: Database technology for decision support systems. IEEE Computer, Dec. 2001, pp. 48-55.
  - R. J. Brachman et al.: Mining business databases. Communications of the ACM, Nov. 1996, pp. 42-48.