

Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Αποθήκες Δεδομένων

Γιάννης Θεοδορίδης, Νίκος Πελέκης

Εργαστήριο Πληροφοριακών Συστημάτων

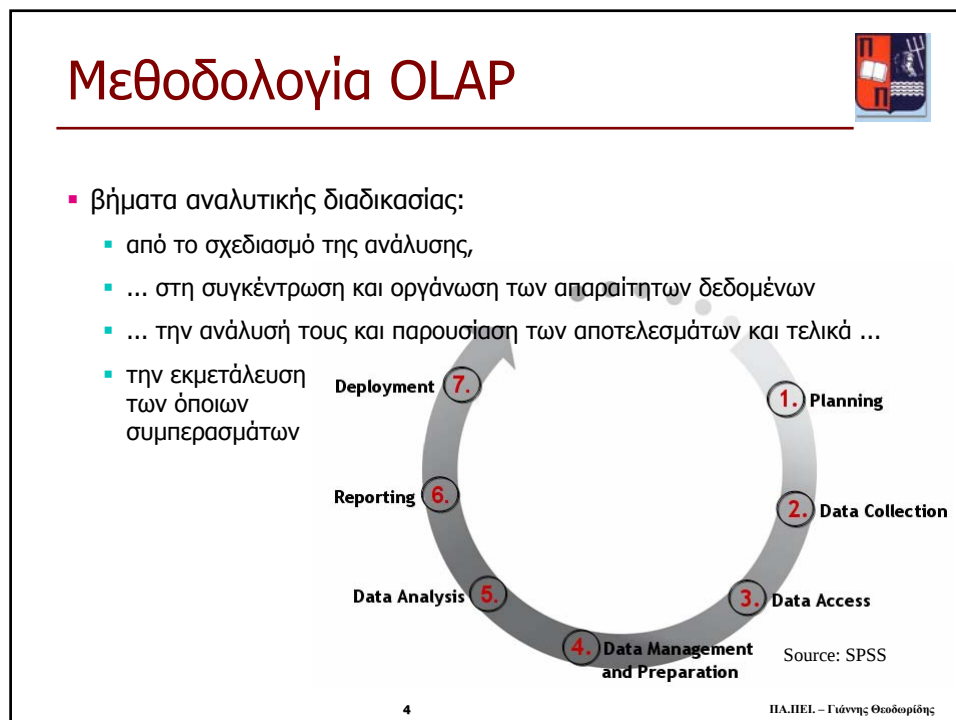
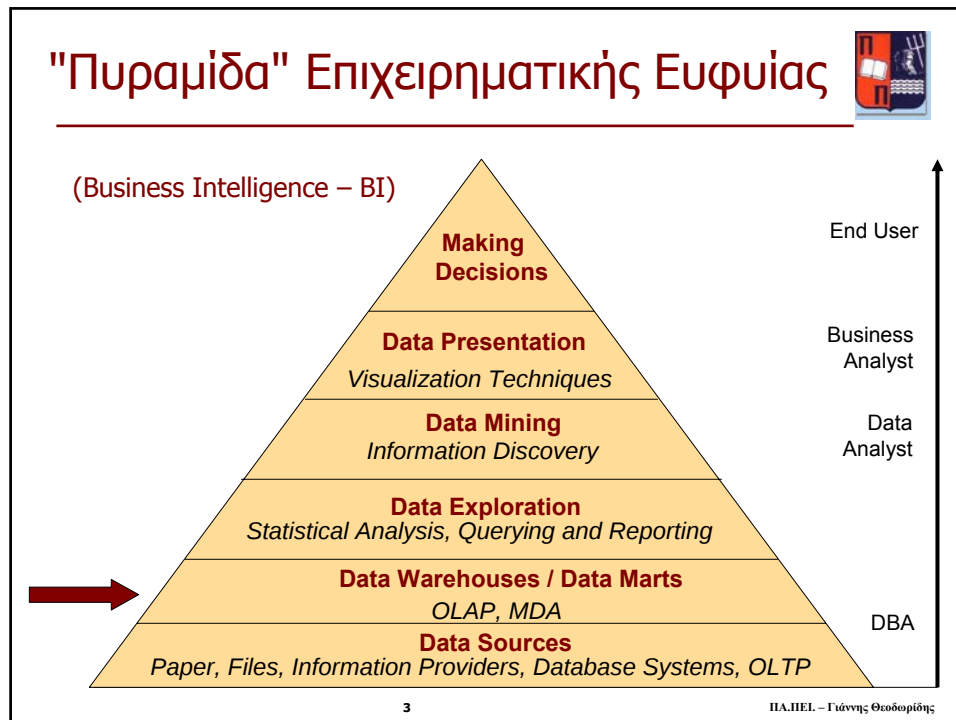
<http://infolab.cs.unipi.gr>

05 Νοε. 08

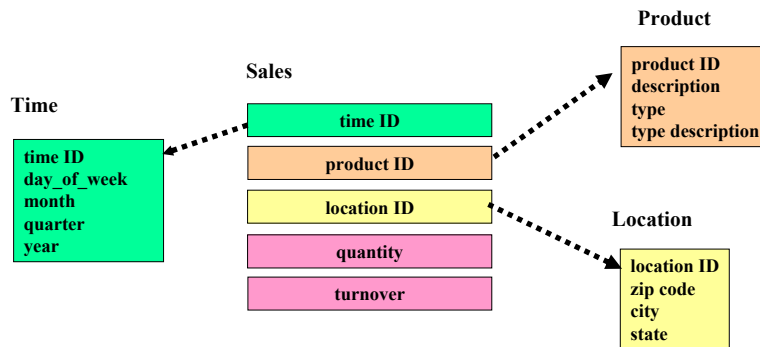
Πηγές



- Το κύριο μέρος των διαφανειών προέρχεται από:
 - Dunham: Data Mining – Introductory and Advanced Topics. Prentice Hall, 2003.
- Επιπρόσθετα, χρησιμοποιήθηκε υλικό από:
 - Han, Kamber: Data Mining – Concepts and Techniques. Morgan Kaufmann, 2001.



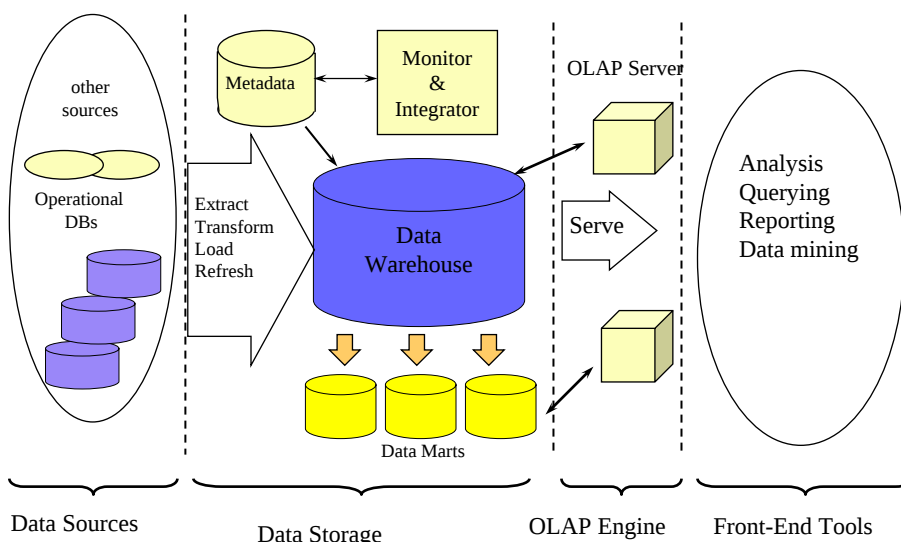
Αποθήκες Δεδομένων – ένα παράδειγμα



5

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Αρχιτεκτονική ΑΔ



6

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Αποθήκες Δεδομένων - ορισμός



- Ορισμός (κατά W. Inmon):

Οι **Αποθήκες Δεδομένων** (data warehouses) αποτελούν

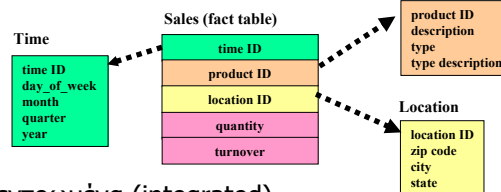
- θεματο-κεντρικά (subject-oriented), συγκεντρωμένα (integrated),
- με χρονική διάσταση (time-variable), μη ευμετάβλητα (non-volatile)

συστήματα διαχείρισης πληροφοριακών δεδομένων για την υποστήριξη των διαδικασιών λήψης αποφάσεων

- **Πληροφοριακά δεδομένα:** υποστηρίζουν άλλες (πέρα των καθημερινών) λειτουργίες της επιχείρησης, όπως σχεδιασμό και πρόβλεψη.

7

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης



Περισσότερα για τον ορισμό των ΑΔ



- **Θεματο-κεντρικά:**

- Οργανώνονται γύρω από συγκεκριμένα θέματα, όπως πελάτες, προϊόντα, πωλήσεις.
- Δεν συμπεριλαμβάνουν πλευρές (δεδομένα) των θεμάτων που δεν συνεισφέρουν στη διαδικασία λήψης αποφάσεων

- **Συγκεντρωμένα:**

- Κατασκευάζονται με συγκέντρωση πολλαπλών, πιθανώς ετερογενών πηγών δεδομένων (σχεσιακές ΒΔ, αρχεία κ.α.)
- Εφαρμόζονται τεχνικές καθαρισμού και ολοκλήρωσης δεδομένων (για την εξασφάλιση συνέπειας)

▪ ...

8

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Περισσότερα για τον ορισμό των ΑΔ



■ ...

■ **Με χρονική διάσταση:**

- Ο χρονικός ορίζοντας στις ΑΔ είναι ευρύτερος αυτού των ΒΔ (ιστορική πληροφορία π.χ. για τα τελευταία 5-10 έτη)
- Η έννοια του χρόνου είναι αναπόσπαστο τμήμα μιας ΑΔ

■ **Μη ευμετάβλητα:**

- Οι ΑΔ αποθηκεύονται ξεχωριστά από τις (επιχειρησιακές) ΒΔ.
- Δεν υπάρχει η έννοια της τροποποίησης δεδομένων (άρα δεν υπάρχουν θέματα επεξεργασίας συναλλαγών, ανάνηψης, ελέγχου συνδρομικότητας).
- Υπάρχει μόνο η λειτουργία **φόρτωσης δεδομένων** είτε πλήρως (full loading) είτε αυξητικά (incremental loading)

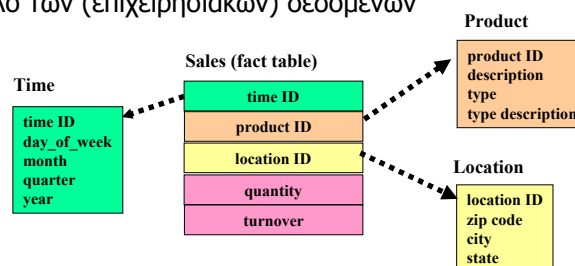
9

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Από τις ΒΔ στις ΑΔ



- Μια **αποθήκη δεδομένων** κρατά συνοπτική πληροφορία για κάποιο υποσύνολο των (επιχειρησιακών) δεδομένων



- Μια ΑΔ ορίζεται πάνω στα δεδομένα μιας (επιχειρησιακής) ΒΔ και η πλήρωσή της γίνεται με κατάλληλη διαδικασία 'φόρτωσης'
- «διαστάσεις» (dimensions) της ΑΔ: οι πίνακες **Time**, **Product**, **Location**
 - «μετρήσιμα μεγέθη» (measures) της ΑΔ: οι συναθροιστικές τιμές **quantity**, **turnover**

10

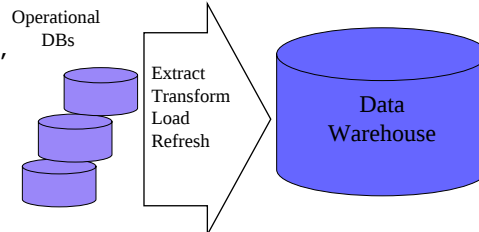
ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Παράδειγμα φόρτωσης ΑΔ



- Παράδειγμα ΒΔ:**
`product (product_ID, description, type, type_description)`
`location (location_ID, zip code, city, state)`
`sales-transaction (timestamp, product_ID, location_ID, units_sold, unit_price)`
- Παράδειγμα ETL:**

```
INSERT INTO sales
( SELECT date(timestamp) AS time_id, product_ID, location_ID,
  sum(units_sold) AS quantity, sum(units_sold*unit_price) AS turnover
FROM sales-transaction
GROUP BY time_id, product_ID, location_ID HAVING quantity > 0 )
```
- Παράδειγμα ΑΔ:**
`product (product_ID, description, type, type_description)`
`location (location_ID, zip code, state, city)`
`time (time_ID, day_of_week, month, quarter, year)`
`sales (time_ID, product_ID, location_ID, quantity, turnover)`



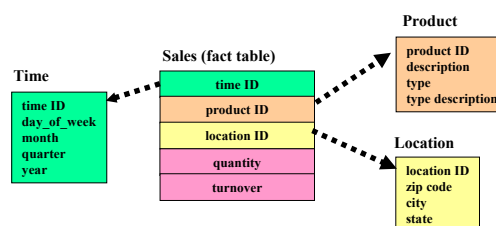
11

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Πολυδιάστατο μοντέλο δεδομένων



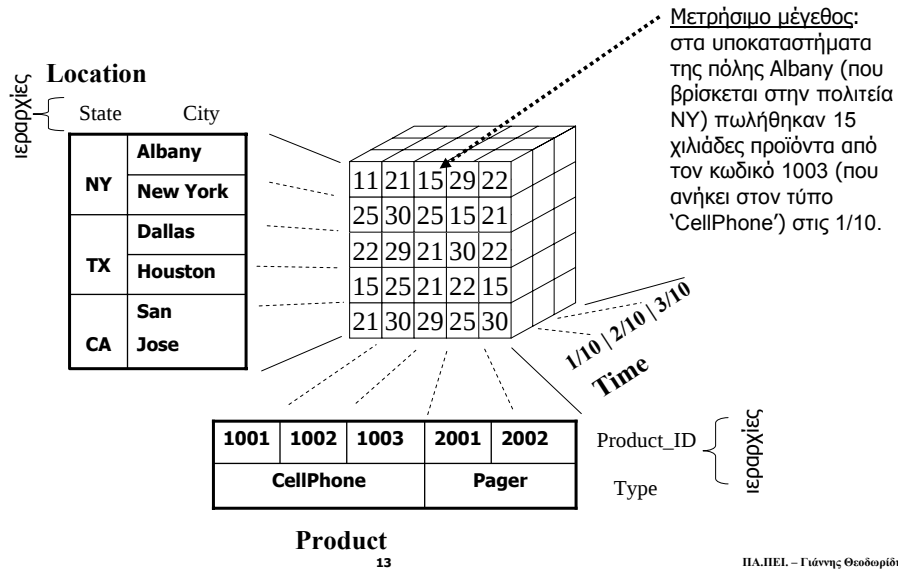
- Το **πολυδιάστατο μοντέλο δεδομένων** απεικονίζει τα δεδομένα σε μορφή κύβου
- Ένας **κύβος δεδομένων** επιτρέπει τη μοντελοποίηση και θεώρηση των δεδομένων σε πολλαπλές διαστάσεις. Το σχήμα αποτελείται από:
 - Πίνακες διαστάσεων** (dimension tables) με πληροφορία για τις διαστάσεις του κύβου
 - Πίνακες γεγονότων** (fact tables) με **μετρήσιμα μεγέθη** και κλειδιά προς τους σχετιζόμενους πίνακες διαστάσεων



12

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

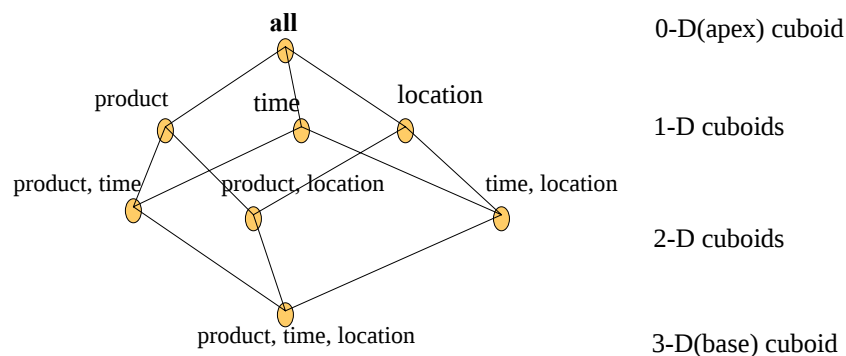
Όψη κύβου δεδομένων



Κύβος και Κυβοειδή



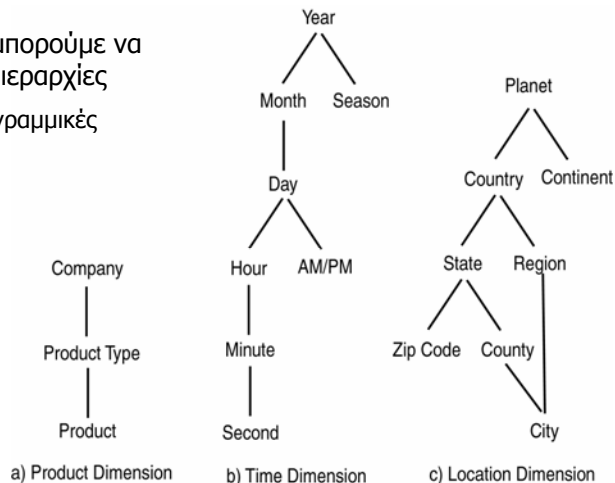
- Από τον αρχικό n-διάστατο κύβο, προκύπτουν κυβοειδή μικρότερης διάστασης



Ιεραρχίες Διαστάσεων



- Σε κάθε διάσταση μπορούμε να ορίσουμε πλούσιες ιεραρχίες
 - όχι κατ' ανάγκην γραμμικές



15

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Λογικό Σχήμα ΑΔ

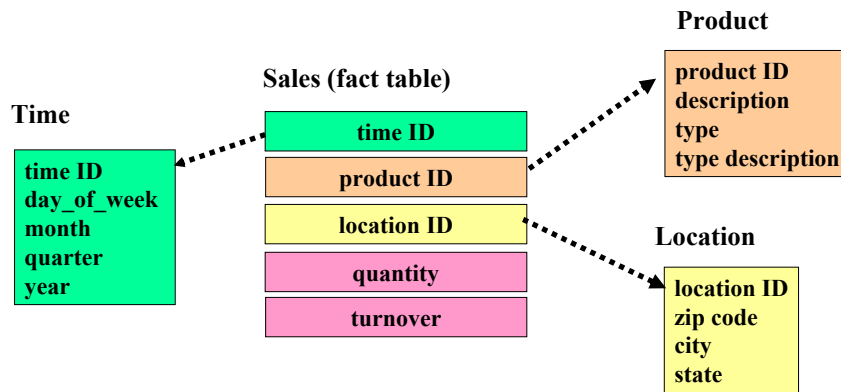


- Μοντελοποίηση ΑΔ: μετρήσιμα μεγέθη και διαστάσεις
 - **Σχήμα αστέρα (star schema):** ένας πίνακας γεγονότων στο κέντρο που περιβάλλεται από πίνακες διαστάσεων
 - **Σχήμα χιονονιφάδας (snowflake schema):** βελτίωση του σχήματος αστέρα, με κάποια διάσταση να είναι κανονικοποιημένη σε σύνολο μικρότερων πινάκων διαστάσεων
 - **Αστερισμοί γεγονότων (fact constellations):** περισσότεροι του ενός πίνακες γεγονότων που μοιράζονται από κοινού πίνακες διαστάσεων
 - μπορεί να θεωρηθεί και ως συλλογή σχημάτων αστέρων – γι' αυτό και εναλλακτικά ονομάζεται **σχήμα γαλαξία**.

16

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

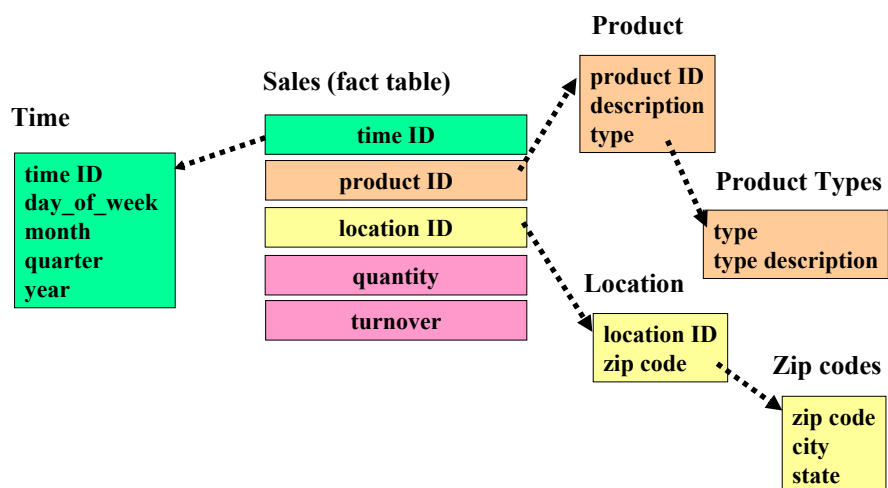
Σχήμα αστέρα



17

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

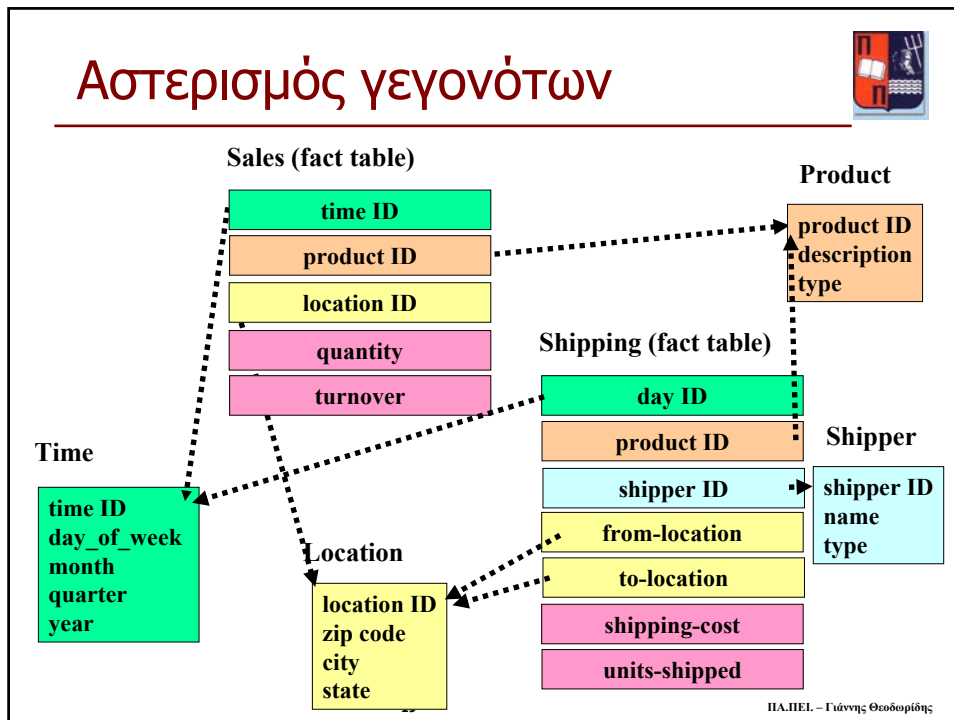
Σχήμα χιονονιφάδας



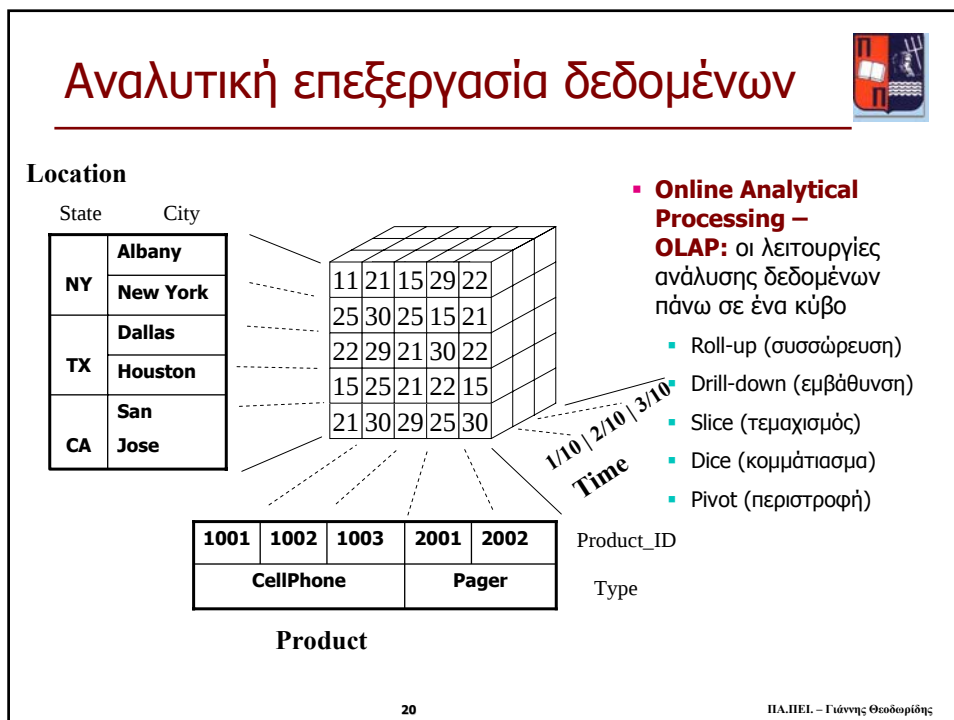
18

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Αστερισμός γεγονότων



Αναλυτική επεξεργασία δεδομένων



Βασικές λειτουργίες OLAP



▪ Συσώρευση (roll-up):

- άνοδος στο lattice των κυβοειδών (με αφαίρεση διάστασης) ή
- άνοδος στην ιεραρχία μιας διάστασης

▪ Εμβάθυνση (drill-down):

το αντίστροφο της λειτουργίας roll-up

Quantity (in 1000's)		2008			
		Q1	Q2	Q3	Q4
TX	Dallas	78	98	34	56
	Houston	90	94	87	91

Roll Up
Dimension: time
(month → quarter)

Quantity (in 1000's)		Q1			Q2			...
		Jan	Feb	Mar	Apr	May	Jun	...
TX	Dallas	30	26	22	20	36	42	...
	Houston	28	30	32	38	30	26	...

21

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Βασικές λειτουργίες OLAP (συν.)



▪ Τεμαχισμός (slice) και

Κομμάτιασμα (dice):

εξέταση υποσυνόλου του κύβου με επιβολή φίλτρων σε κάποιες διαστάσεις

Quantity (in 1000's)		2008 (any product)			
		Q1	Q2	Q3	Q4
TX	Dallas	78	45	34	56
	Houston	90	67	87	91

Slice
product.type = 'CellPhone'

Quantity (in 1000's)		2008 (CellPhones only)			
		Q1	Q2	Q3	Q4
TX	Dallas	53	35	20	48
	Houston	76	57	40	80

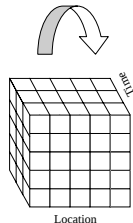
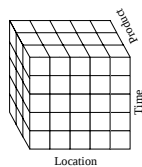
22

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Βασικές λειτουργίες OLAP (συν.)



- **Περιστροφή (pivot):** εμφάνιση κύβου από διαφορετική οπτική γωνία



Quantity (in 1000's)		CellPhone			Pager	
		1001	1002	1003	2001	2002
TX	Dallas	20	33	12	8	12
	Houston	32	45	34	20	23

Pivot

Quantity (in 1000's)		2008 (any product)			
		Q1	Q2	Q3	Q4
TX	Dallas	78	45	34	56
	Houston	90	67	87	91

23

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Αρχιτεκτονικές OLAP Servers



- **Relational OLAP (ROLAP)**
 - ένα σχεσιακό DBMS αποθηκεύει και διαχειρίζεται τα δεδομένα του warehouse
 - ένα OLAP middle ware υποστηρίζει τις λειτουργίες OLAP
 - εκμεταλλεύεται τις δυνατότητες του DBMS (βελτιστοποίηση κλπ.)
 - κλιμακώνεται καλά
- **Multidimensional OLAP (MOLAP)**
 - μηχανή αποθήκευσης πολυδιάστατων δεδομένων βασισμένη σε arrays / matrices
 - γρήγορη δεικτοδότηση πάνω στις (προ-υπολογισμένες) συνόψεις
- **Hybrid OLAP (HOLAP)**
 - ευέλικτο σχήμα, σχεσιακοί πίνακες στο χαμηλό επίπεδο και μήτρες στο υψηλό επίπεδο

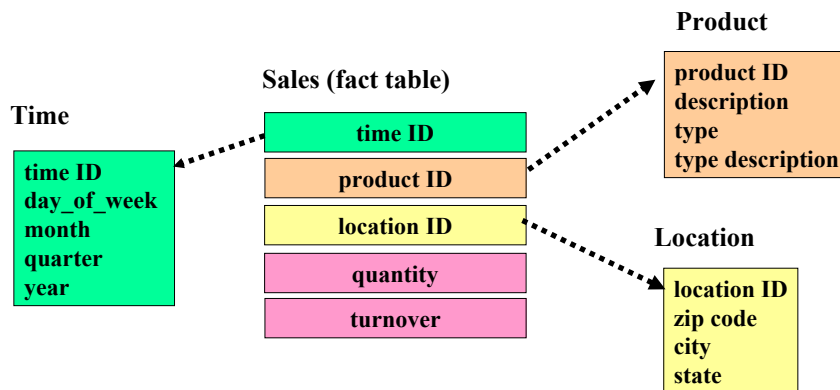
24

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Υλοποίηση τύπου ROLAP



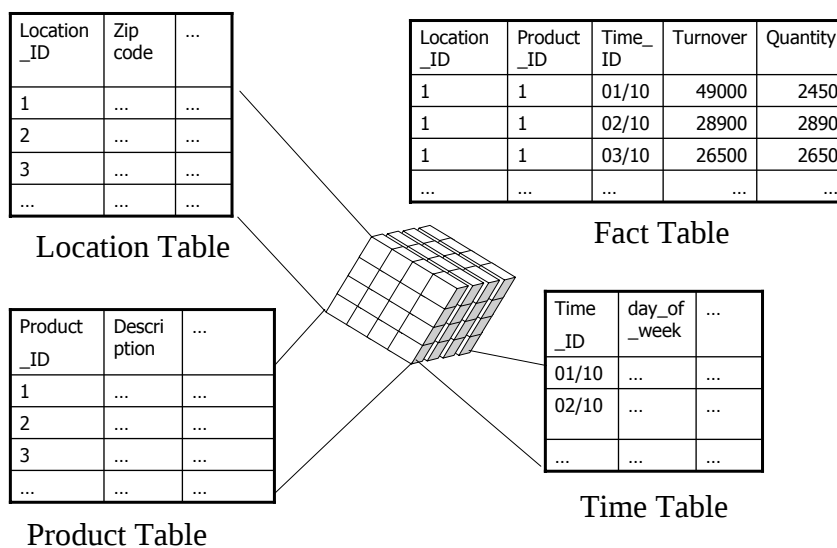
- Τόσο τα dimension tables όσο και το fact table είναι σχεσιακοί πίνακες



25

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα ROLAP



26

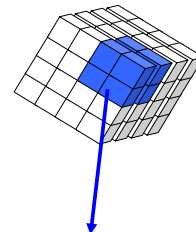
ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα ROLAP (συν.)



- Παράδειγμα Slice: "Ανάλυση πωλήσεων μόνο για τα κινητά τηλέφωνα που πωλήθηκαν το 2007 στην πολιτεία Texas"

```
SELECT Quantity
FROM Location L, Product P,
     Time T, Fact F
WHERE L.Location_ID = F.Location_ID
      AND P.Product_ID = F.Product_ID
      AND T.Time_ID = F.Time_ID
      AND L.State = 'TX'
      AND P.Type = 'CellPhone'
      AND T.Year = 2007
```



Slice: Loc=1..2, Prod=3..4
Time=3..4

27

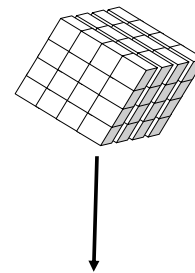
ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα ROLAP (συν.)



- Οι λειτουργίες Roll-up, Drill-down υλοποιούνται με κατάλληλα 'GROUP BY'
- Παράδειγμα Roll-up: "Ανάλυση πωλήσεων ανά πολιτεία"

```
SELECT SUM(F.Quantity)
FROM Location L, Product P,
     Time T, Fact F
WHERE L.Location_ID = F.Location_ID
      AND P.Product_ID = F.Product_ID
      AND T.Time_ID = F.Time_ID
GROUP BY L.State
```



Roll-up: Location.state

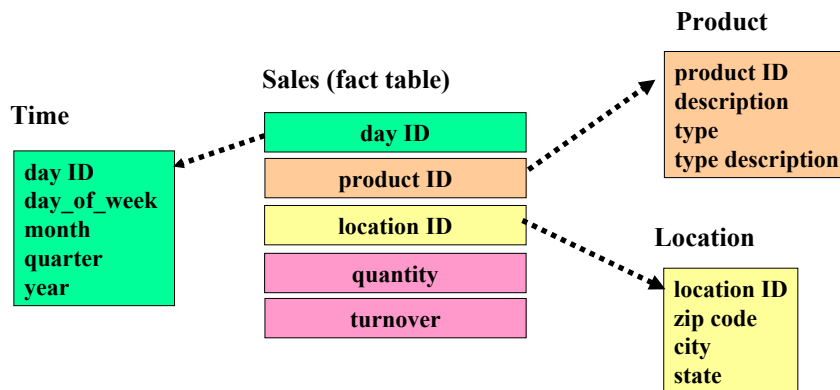
28

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Υλοποίηση τύπου MOLAP



- Το fact table είναι υλοποιημένο σε δομή πολυδιάστατης μήτρας (matrix) ή πίνακα (array)



29

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα MOLAP



- Ο κύβος είναι υλοποιημένος ως ένας 3-διάστατος πίνακας:

$\text{Fact}(\text{Loc}, \text{Prod}, \text{Time})$

- Επίσης, είναι προκατασκευασμένα όλα τα κυβοειδή !

- Παράδειγμα Slice: “Ανάλυση πωλήσεων μόνο για τα κινητά τηλέφωνα που πωλήθηκαν το 2007 στην πολιτεία Texas”

$$\sum_{\text{Loc}=1}^2 \sum_{\text{Prod}=3}^4 \sum_{\text{Time}=3}^4 \text{Fact}(\text{Loc}, \text{Prod}, \text{Time})$$

- Παράδειγμα Roll-up: “Ανάλυση πωλήσεων ανά πολιτεία”

- Δεν απαιτείται κάποια ειδική λειτουργία αφού είναι διαθέσιμα όλα τα κυβοειδή, άρα και το $\text{Fact}_{\text{Prod}, \text{Time}}(\text{Loc})$

30

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Κατακλείδα



- Οι ΑΔ τηρούνται ξεχωριστά από τις (επιχειρησιακές) ΒΔ, για δύο κυρίως λόγους:
 - Θέματα επίδοσης των δύο συστημάτων
 - ΣΔΒΔ — ρυθμισμένο να έχει καλές επιδόσεις σε OLTP: μέθοδοι προσπέλασης, ευρετήρια, έλεγχος συνδρομικότητας, ανάνηψη
 - ΑΔ — ρυθμισμένη να έχει καλές επιδόσεις σε OLAP: σύνθετες λειτουργίες OLAP, πολυδιάστατη θεώρηση δεδομένων.
 - Διαφορετικές λειτουργίες πάνω σε διαφορετικά δεδομένα:
 - **Ιστορικά δεδομένα:** Οι αποφάσεις των στελεχών εμπλέκουν δεδομένα του παρελθόντος, οι ΒΔ συντηρούν δεδομένα του παρόντος
 - **Συγκέντρωση δεδομένων:** Οι ΑΔ απαιτούν συγκεντρωτικά πληροφορία (συναθροίσεις, συνόψεις) από δεδομένα που προέρχονται από ετερογενείς πηγές ΒΔ