Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Σχεδιασμός και διαχείριση Αποθηκών και Βάσεων δεδομένων

Οι πιο συνηθισμένες πηγές δεδομένων είναι οι παρακάτω

- Σχεσιακές ΒΔ
- Αποθήκες δεδομένων
- Από Αρχεία

Σχεσιακές Βάσεις Δεδομένων

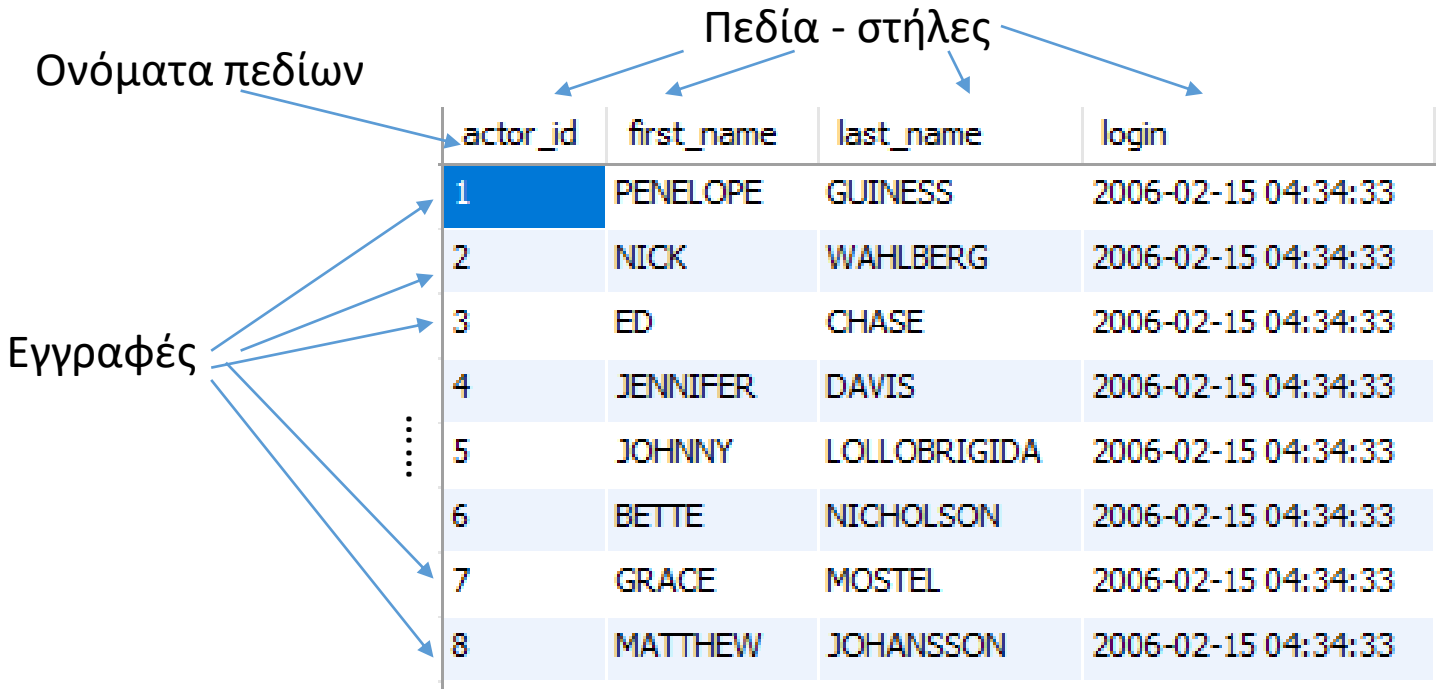
- Ο πιο ευρέως χρησιμοποιούμενος τρόπος συλλογής και αποθήκευσης δεδομένων σε ένα πληροφοριακό σύστημα είναι οι σχεσιακές βάσεις δεδομένων.
- Οι πιο δημοφιλής Βάσεις δεδομένων είναι
  - Mysql
  - Oracle και
  - SQl Server

- Τα δεδομένα αποθηκεύονται σε πίνακες. Για τη δημιουργία των πινάκων και την άντληση στοιχείων από αυτούς χρησιμοποιείται η γλώσσα SQL (Structured Query Language).

Σχεσιακές Βάσεις Δεδομένων

Κάθε πίνακας περιέχει εγγραφές με τα δεδομένα της οντότητας/αντικειμένου για το οποίο σχεδιάστηκε.

Για παράδειγμα στην παρακάτω εικόνα βλέπουμε τον πίνακα actor, ο οποίος περιέχει 8 εγγραφές με δεδομένα. Επίσης βλέπουμε 4 στήλες οι οποίες αποτελούν τα πεδία του πίνακα actor. Δηλαδή για κάθε actor αποθηκεύουμε έναν κωδικό, το όνομα , το επίθετο και την ημερομηνία που συνδέθηκε στο σύστημα .



actor_id	first_name	last_name	login
1	PENELOPE	GUINESS	2006-02-15 04:34:33
2	NICK	WAHLBERG	2006-02-15 04:34:33
3	ED	CHASE	2006-02-15 04:34:33
4	JENNIFER	DAVIS	2006-02-15 04:34:33
5	JOHNNY	LOLLOBRIGIDA	2006-02-15 04:34:33
6	BETTE	NICHOLSON	2006-02-15 04:34:33
7	GRACE	MOSTEL	2006-02-15 04:34:33
8	MATTHEW	JOHANSSON	2006-02-15 04:34:33

Στήλες=4
Εγγραφές=8

Σχεσιακές Βάσεις Δεδομένων

Κατά τον σχεδιασμό ενός πίνακα θα πρέπει να ορίσουμε ότι μια στήλη θα πρέπει να είναι μοναδική ή μοναδικό κλειδί (primary key), δηλαδή όταν εισάγουμε μια νέα εγγραφή δεν θα πρέπει να υπάρχει προηγούμενη εγγραφή με την ίδια τιμή.

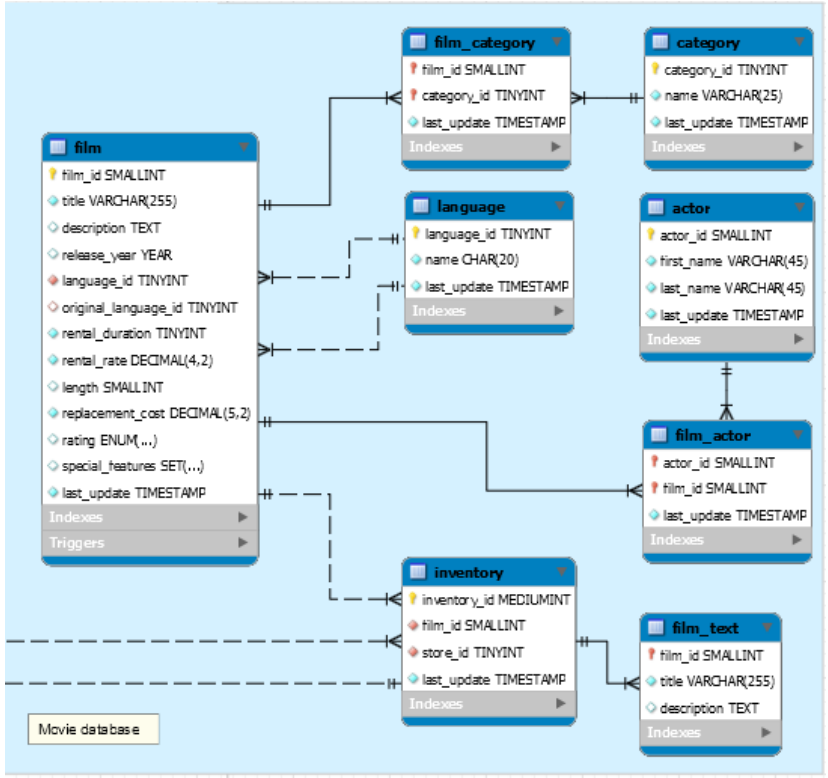
Πιο συγκεκριμένα έστω ότι έχουμε καταχωρήσει την ηθοποιό Penelope Guiness με κωδικό 1 και στον συγκεκριμένο πίνακα έχουμε ορίσει την στήλη actor_id σαν μοναδική στήλη.

Αν πάμε να καταχωρήσουμε τον ηθοποιό Nick Papas με κωδικό 1 , το σύστημα δεν θα μας το επιτρέψει. Προκειμένου να καταχωρηθεί σωστά θα πρέπει να δώσουμε σαν κωδικό, π.χ., το 2.

Οπότε όταν μια στήλη ορίζεται ως το μοναδικό κλειδί του πίνακα τότε δεν μπορούμε να καταχωρήσουμε μια νέα εγγραφή με την ίδια τιμή.

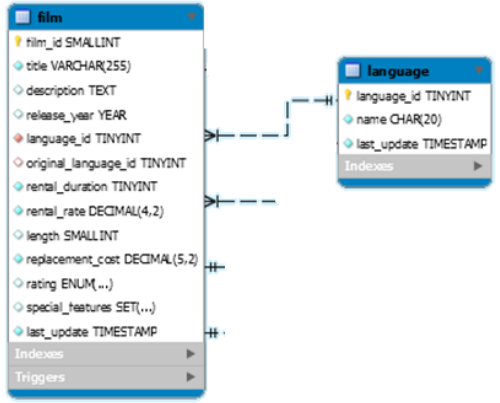
Σχεσιακές Βάσεις Δεδομένων

Στην παρακάτω εικόνα βλέπετε ένα σχεσιακό μοντέλο το οποίο αποτελείται από πολλούς πίνακες και οι οποίοι συνδέονται μεταξύ τους μέσω των συσχετίσεων των πινάκων.



Σχεσιακές Βάσεις Δεδομένων

Για παράδειγμα ο πίνακας με τις ταινίες (film) συνδέεται ή συσχετίζεται με τον πίνακα με τις γλώσσες.



Ο πίνακας με τις ταινίες (film) περιλαμβάνει τα πεδία: κωδικό, περιγραφή, χρονιά δημιουργίας, **κωδικός γλώσσας,** ...

Ενώ ο πίνακας γλώσσες (language) περιλαμβάνει τα πεδία:

κωδικός γλώσσας και περιγραφή.

Σχεσιακές Βάσεις Δεδομένων

Όπως παρατηρούμε στον πίνακα film αντί για το όνομα της γλώσσας έχουμε τον κωδικό της γλώσσας.

film_id	title	description	release_year	language_id	c
1	ACADEMY DINOSAUR	A Epic Drama of a Feminist And a Mad Scientist ...	2006	1	NU
2	ACE GOLDFINGER	A Astounding Epistle of a Database Administrat...	2006	1	NU
3	ADAPTATION HOLES	A Astounding Reflection of a Lumberjack And a ...	2006	1	NU

language_id	name
1	English
2	Italian
3	Japanese
4	Mandarin
5	French
6	German

Οπότε αν θέλουμε να μάθουμε τη γλώσσα της ταινίας με κωδικό (film_id)=1, θα πρέπει να δούμε στον πίνακα Language ποια γλώσσα έχει κωδικό=1.

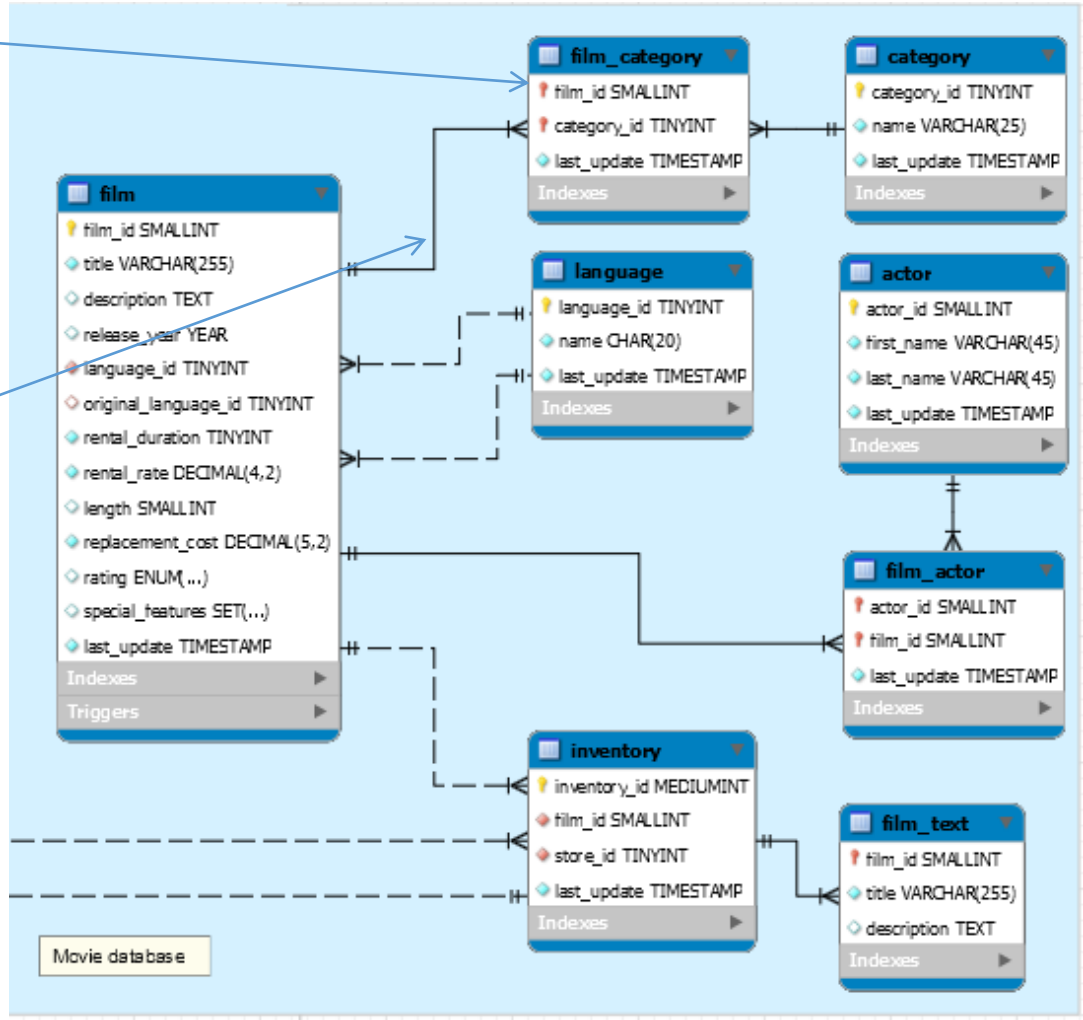
Όπως παρατηρούμε η γλώσσα με κωδικό=1 είναι τα Αγγλικά.

Οπότε υπάρχει σύνδεση του πίνακα film με τον πίνακα language μέσω του πεδίου language_id. Σε αυτή την περίπτωση λέμε ότι υπάρχει συσχέτιση μεταξύ των 2 πινάκων.

Σχεσιακές Βάσεις Δεδομένων

Κύριο Κλειδί
(Primary Key)

Συσχέτιση -Ξένο κλειδί
(Foreign Key)



Αποθήκες δεδομένων (Data Warehouse)

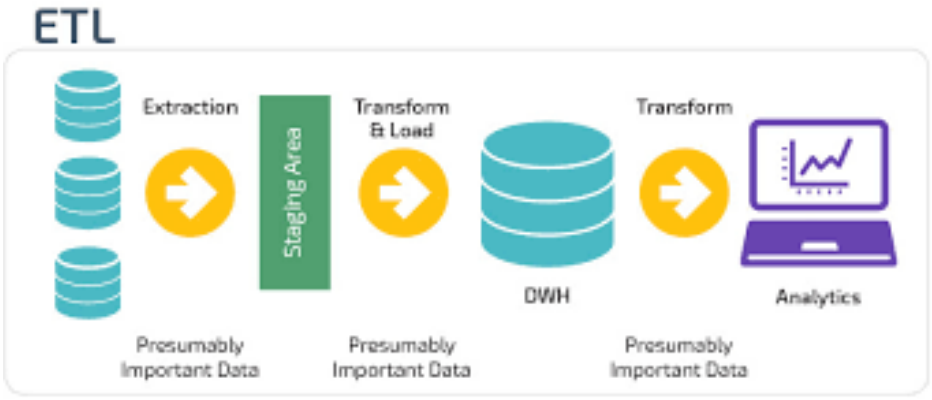
- Οι αποθήκες δεδομένων σχεδιάζονται με σκοπό να συλλέγουν συγκεντρωτικά στοιχεία προκειμένου να υποστηρίξουν συστήματα λήψης αποφάσεων.
- Οι αποθήκες δεν υποστηρίζουν την λειτουργία ενός πληροφοριακού συστήματος και δεν υποκαθιστούν την βάση δεδομένων, αλλά λειτουργούν συμπληρωματικά με αυτή.
- Οι χρήστες στους οποίους απευθύνονται είναι οι αναλυτές δεδομένων και τα στελέχη των εταιριών.

Αποθήκες δεδομένων

- Μία αποθήκη δεδομένων αποτελεί μία συλλογή δεδομένων που προέρχονται από ετερογενείς πηγές (σχεσιακή βάση, αρχείο csv, log files,….) για την επεξεργασία των οποίων χρησιμοποιείται η τεχνική ETL (Extract Transform Load)
- Μια αποθήκη δεδομένων μπορεί να αποτελείται από έναν ή περισσότερους κύβους.
- Στη συνέχεια και με σκοπό την λήψη αποφάσεων σε έναν οργανισμό, τα δεδομένα αναλύονται με χρήση αλγορίθμων μηχανικής μάθησης.

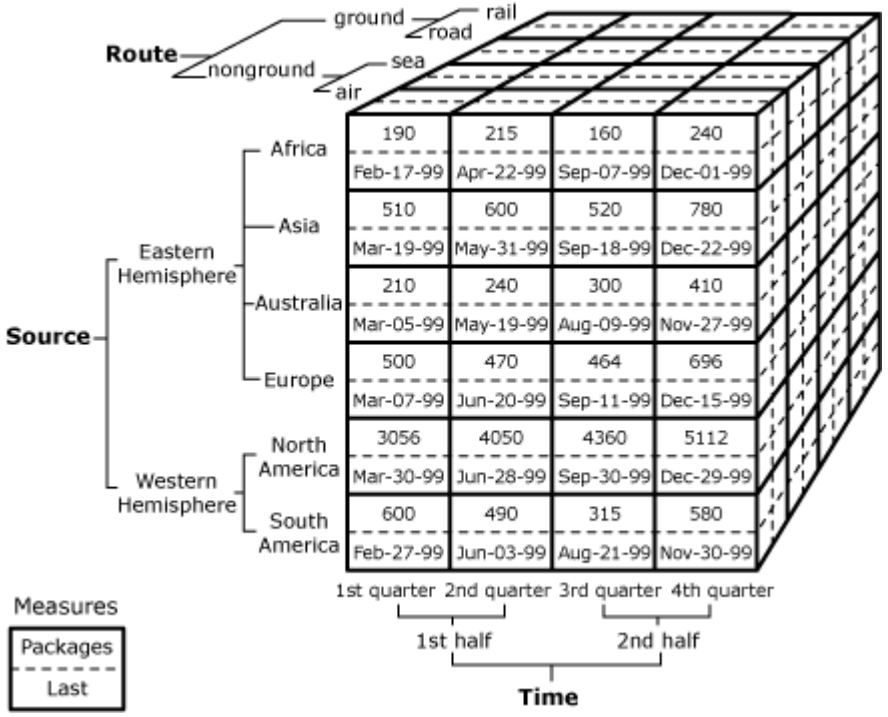
Διαδικασίας ETL

- Λήψη δεδομένων από πολλές, ετερογενείς ΒΔ ή αρχεία.
- Καθαρισμός Δεδομένων.
- Μετασχηματισμός Δεδομένων.
- Φόρτωση των δεδομένων στην αποθήκη δεδομένων.
- Ενημέρωση των κύβων
- Εκτέλεση διαδικασιών OLAP (online Analytical Processing)

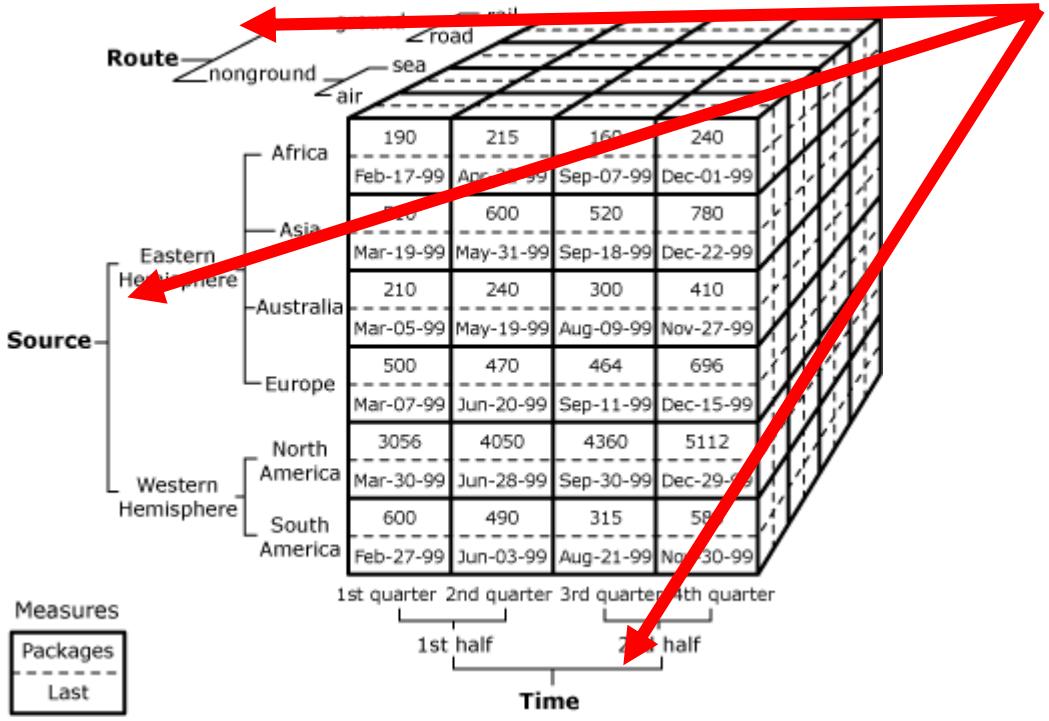


Κύβοι

- Ένας κύβος είναι μια δομή δεδομένων που αποτελείται από πολλαπλές Διαστάσεις και επιτρέπει την γρήγορη ανάλυση δεδομένων.
- Ένας πολυδιάστατος κύβος για την αναφορά Μεταφορές Προϊόντων μπορεί να αποτελείται, για παράδειγμα, από 3 διαστάσεις: Route (Διαδρομή), Source (Προορισμός), Time (HM/NIA).



Multi-Dimensional Database (Cube)



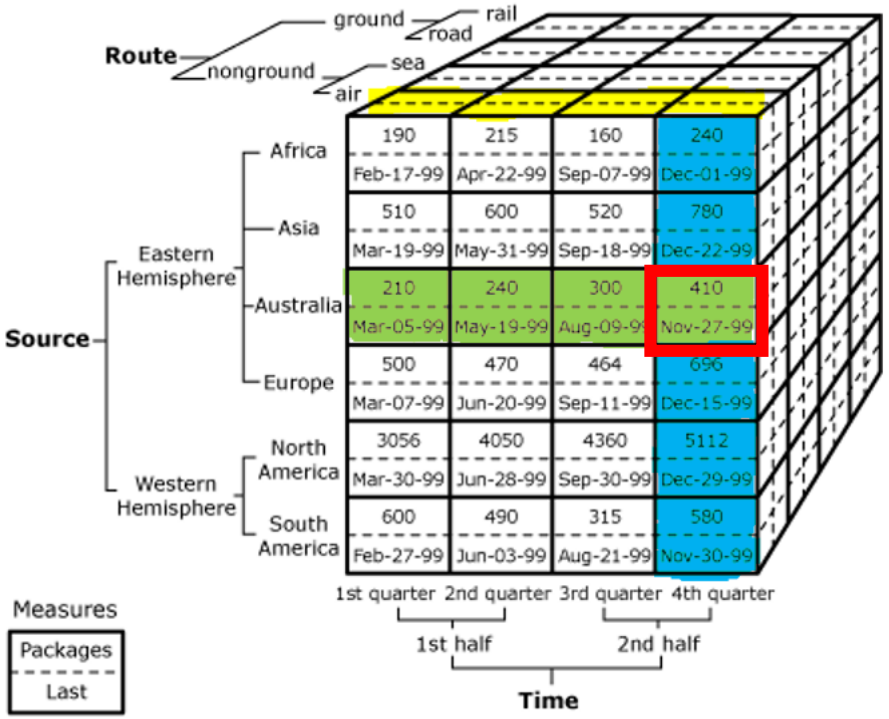
Διαστάσεις

Στο κύβο του παραδείγματος βλέπουμε τις 3 διαστάσεις.

- Route
- Source
- Time

Reference: http://technet.microsoft.com/en-us/library/ms174587(v=sql.90).aspx

Multi-Dimensional Database (Cube)



Οπότε η τιμή 410 πακέτα αναφέρεται σε :

Διαδρομή:
Non-Ground / Air

Προορισμός:
Eastern Hemisphere / Australia

HM/NIA:
2nd Half / 4th Quarter on November 27, 1999

Reference: http://technet.microsoft.com/en-us/library/ms174587(v=sql.90).aspx

Σχεδιασμός Data Mart

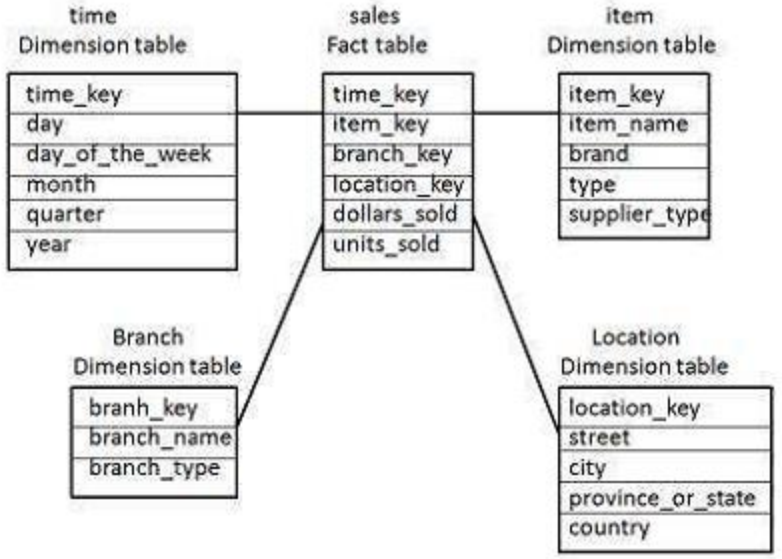
Για τον σχεδιασμό υπάρχουν 3 τεχνικές

- Σχήμα αστέρα (Star Schema).
- Σχήμα νιφάδας (Snowflakes Schema).
- Αστερισμοί γεγονότων (Fact Constellation Schema).

Στο δικό μας παράδειγμα θα χρησιμοποιήσουμε την τεχνική του αστέρα, όπου έχουμε ένα πίνακα γεγονότων στη μέση και πολλούς άλλους πίνακες διαστάσεων που συνδέονται με τον κεντρικό πίνακα.

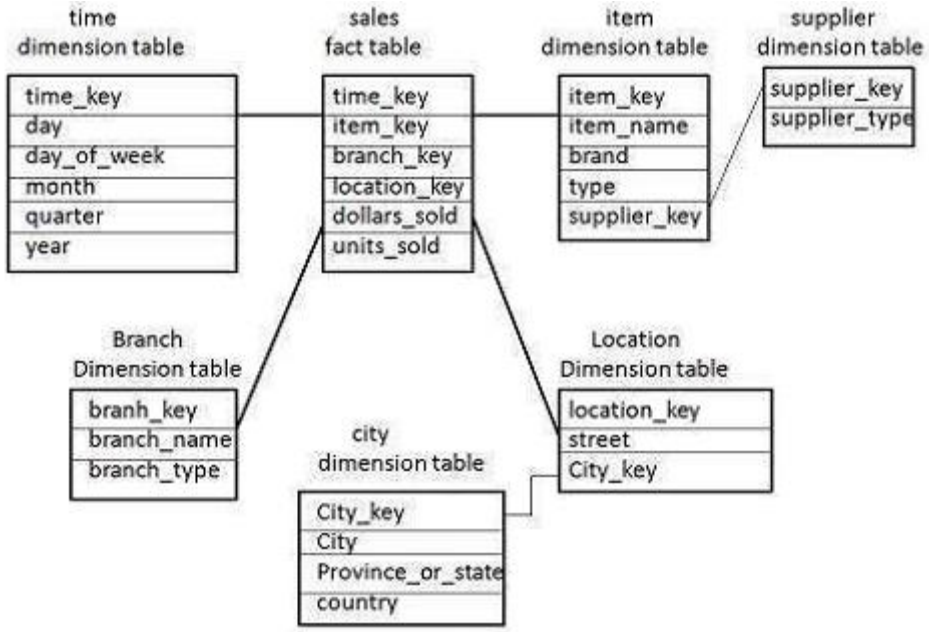
Star Schemas

- Έχουμε πολλούς πίνακες με τις βασικές οντότητες της εφαρμογής (dimension tables).
- Έχουμε έναν πίνακα που περιέχει τις συνδέσεις των πινάκων (fact table).
- Κάθε πίνακας Dimension συνδέεται μέσω ενός κλειδιού στον fact πίνακα .



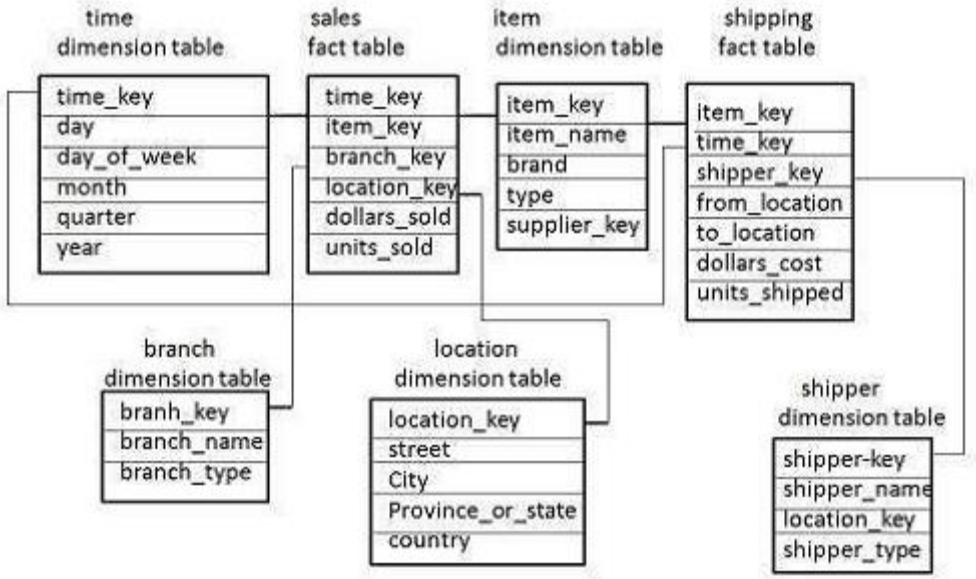
Snowflakes

- Μοιάζει με το Star Schema ,με την διαφορά ότι ένας dimension πίνακας σπάει και σε άλλους μικρότερους dimension πίνακες



Fact Constellation

- Σε αυτό το Schema έχουμε περισσότερους fact πίνακες (Sales &Shipping)



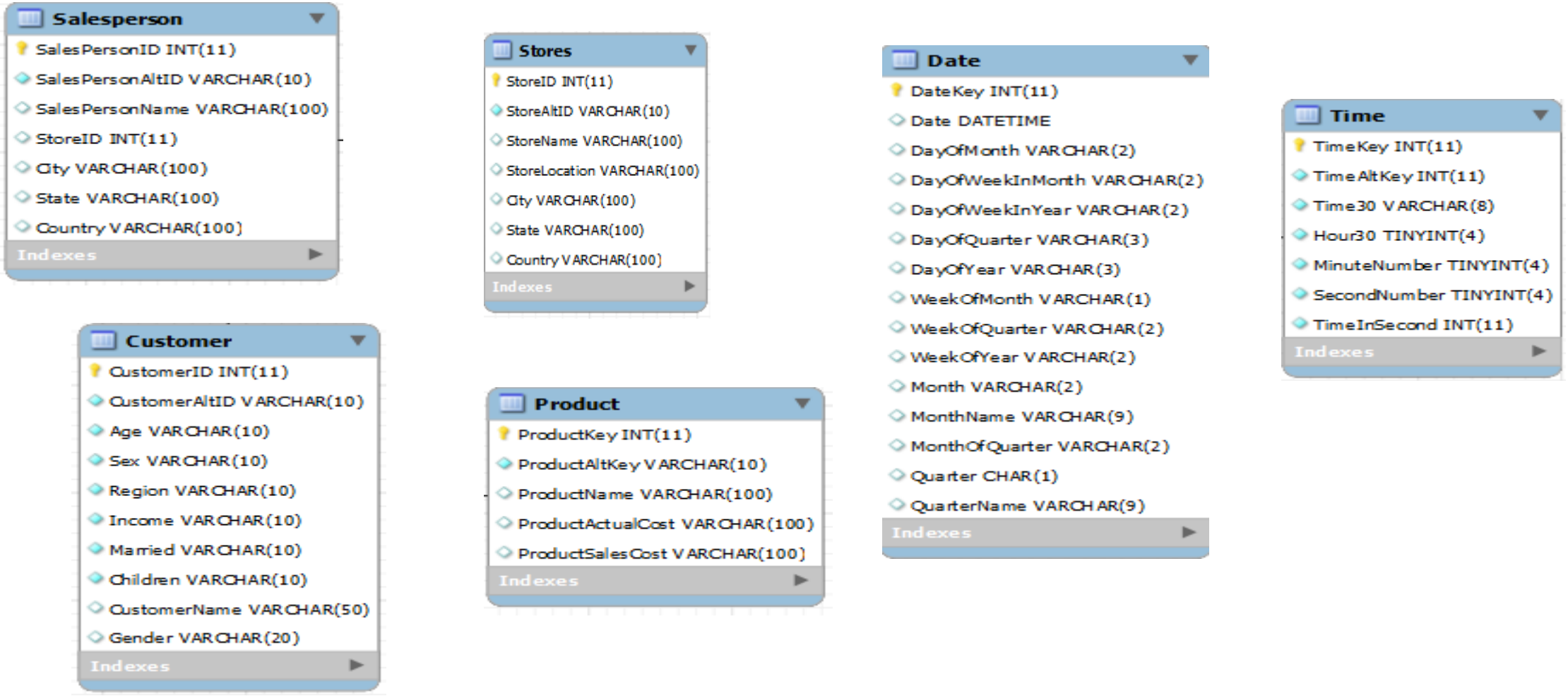
Παράδειγμα - Ηλεκτρονικό Κατάστημα

- Υπάρχουν 1 ηλεκτρονικό κατάστημα το οποίο διαθέτει 7 σημεία πώλησης σε όλη την Ευρώπη.
- Το κατάστημα υποστηρίζει 3 διαφορετικά είδη προϊόντων: κινητά τηλέφωνα, ηλεκτρικές συσκευές και είδη κάμπινγκ.
- Οι λειτουργίες που υποστηρίζει το κατάστημα είναι: α) πωλήσεις, β) αγορά μεταχειρισμένων προϊόντων.
- Μία από τις σημαντικότερες αποφάσεις στην επιχείρηση, αποτελεί η τιμολόγηση και η προώθηση των νέων προϊόντων.
- Δεδομένα λαμβάνονται κατά την αγορά προϊόντων και κατά την διάρκεια επίσκεψης στην ιστοσελίδα του καταστήματος.

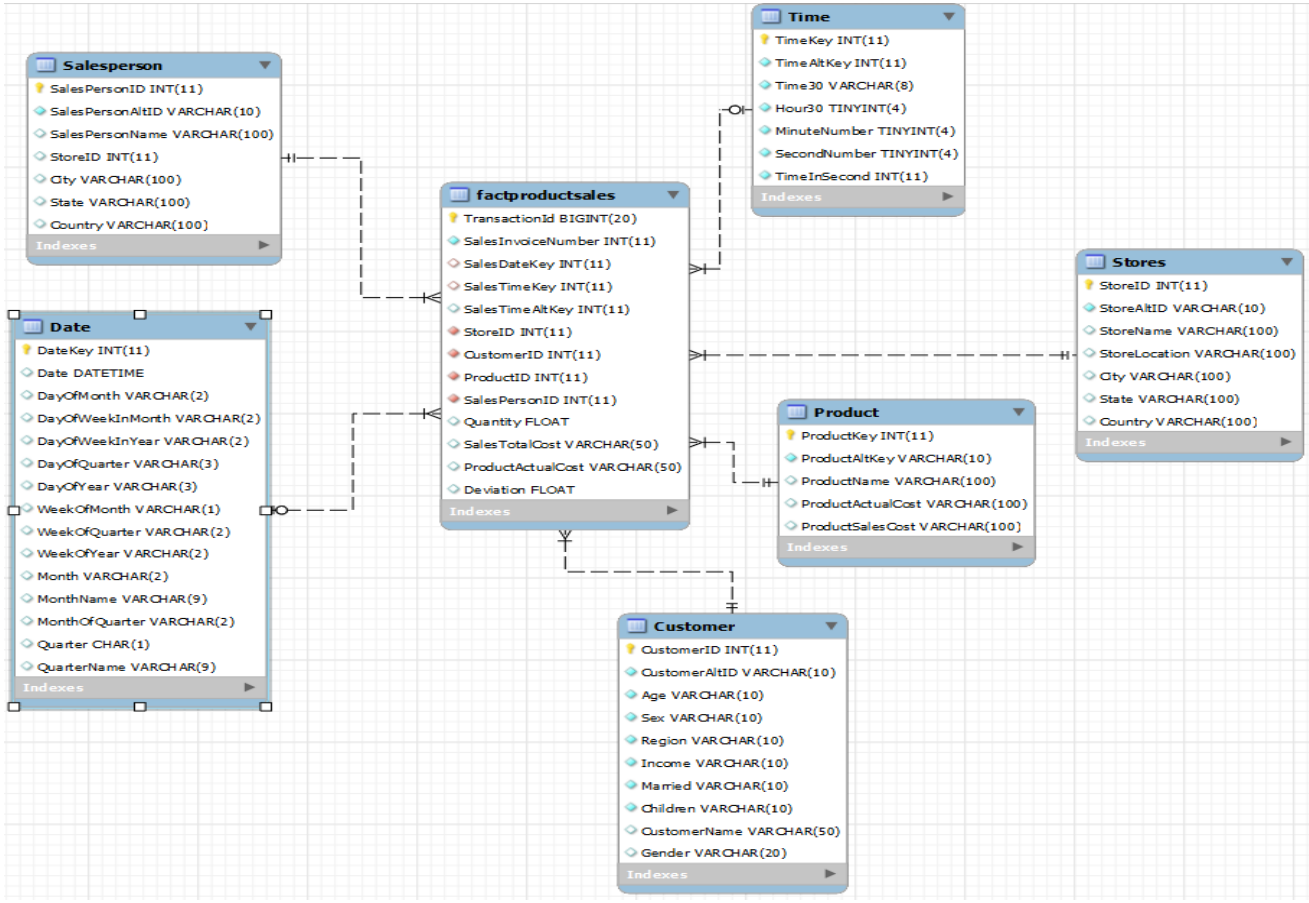
Τα 4 βήματα για το Ηλεκτρονικό Κατάστημα

- Επιλογή της διαδικασίας ή λειτουργίας που θέλουμε να αναλύσουμε π.χ. Προώθηση Προϊόντων
- Επιλογή επιπέδου λεπτομέρειας.
  - Το επίπεδο της λεπτομέρειας θα καθορίσει και το είδος των πληροφοριών που θα καταγράφουμε π.χ. κρατάμε πληροφορίες ξεχωριστά για κάθε αγορά πακέτου για κάθε πελάτη ή συνολικά ανά είδος.
- Επιλογή των διαστάσεων π.χ. είδος, ημερομηνία, προφίλ πελάτη,...
- Επιλογή των μεγεθών που θέλουμε να μετράμε π.χ. πωλήσεις, κόστος αντικειμένου,...

Dimensions - Ηλεκτρονικό Κατάστημα



Dimensions+Facts - Ηλεκτρονικό Κατάστημα



ΤΕΛΟΣ