ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ

Σ. ΛΙΓΟΥΔΙΣΤΙΑΝΟΣ

8.1 ΓΕΝΙΚΑ

Από τα μέσα της δεκαετίας του '70, η αλματώδης παραγωγή πολύ ισχυρών συστημάτων διαχείρισης βάσεων δεδομένων βοήθησε στην ανάπτυξη πληροφοριακών συστημάτων που καλύπτουν τις λειτουργικές ανάγκες οργανισμών και επιχειρήσεων. Τα μεγαλύτερα και ισχυρότερα συστήματα αναπτύχθηκαν με στόχο τον αυτοματισμό βασικών αναγκών των οργανισμών όπως η διεκπεραίωση των τραπεζικών εργασιών και τα λογιστικά συστήματα. Η λειτουργία αυτών των πληροφοριακών συστημάτων είναι πλέον κρίσιμη και πολύτιμη για τη ζωή των οργανισμών στους οποίους έχουν εγκατασταθεί, η δε βάση δεδομένων ενός τέτοιου συστήματος αποτελεί τον πυρήνα τους. Η ορθή σχεδίαση, ανάπτυξη και λειτουργία της βάσης είναι ο σημαντικότερος παράγοντας για την επιτυχία ενός πληροφοριακού συστήματος. Τα συστήματα αυτά παρέχουν τη δυνατότητα επεξεργασίας μεγάλου αριθμού δοσοληψιών που διαχειρίζονται τα δεδομένα του οργανισμού (On-line transaction processing - OLTP). Ένα άλλο είδος πληροφοριακών συστημάτων που αναπτύσσονται στους οργανισμούς είναι τα συστήματα στήριξης αποφάσεων που σκοπό έχουν να βοηθήσουν τα στελέχη των οργανισμών να σχεδιάσουν τις δραστηριότητές του. Η επιτυχία των συστημάτων αυτών είναι επίσης βασικός παράγοντας επιτυχίας του οργανισμού. Μία βασική απαίτηση των συστημάτων στήριξης αποφάσεων είναι η αποδοτική πρόσβαση στα δεδομένα των συστημάτων αυτοματισμού. Το πρόβλημα που προκύπτει, όμως, είναι ότι τα συστήματα αυτοματισμού έχουν ήδη πολύ σοβαρό υπολογιστικό φορτίο από μόνα τους και επιπλέον, είναι σχεδιασμένα για την εκτέλεση διαφορετικών λειτουργιών.

Ένας τηλεπικοινωνιακός οργανισμός, για παράδειγμα, συνήθως διαθέτει ένα μεγάλο πληροφοριακό σύστημα ελέγχου του τηλεφωνικού δικτύου του. Αυτό το σύστημα ελέγχει την ομαλή λειτουργία του δικτύου και παράλληλα των παροχή υπηρεσιών και την χρέωση των συνδρομητών του. Η βάση δεδομένων του συστήματος περιέχει όλα τα δεδομένα των παραπάνω εργασιών. Είναι σαφές ότι αυτό το σύστημα λειτουργεί συνεχώς (24 ώρες

ημερησίως) με μεγάλο όγκο δοσοληψιών (transactions) να εξυπηρετούνται στη βάση δεδομένων. Από αυτή τη βάση θα πρέπει να αντλήσει και ένα σύστημα στήριξης αποφάσεων τα απαραίτητα δεδομένα, για να μπορέσει να βοηθήσει στο σχεδιασμό της λειτουργίας του οργανισμού. Μελετώντας λίγο πιο προσεκτικά την περίπτωση αυτή, θα δούμε ότι είναι πρακτικά αδύνατο το σύστημα ελέγχου του δικτύου και το σύστημα στήριξης αποφάσεων να λειτουργούν, χρησιμοποιώντας την ίδια βάση δεδομένων. Διάφορα προβλήματα κάνουν αδύνατη την εφαρμογή αυτού του σεναρίου. Τα κυριότερα από αυτά τα προβλήματα είναι τα παρακάτω:

1. Τα δύο συστήματα αναπτύχθηκαν πιθανότατα από διαφορετικούς ανθρώπους και κυρίως με τη χρήση διαφορετικών τεχνολογιών. Είναι πιθανό η τεχνολογία του συστήματος αποφάσεων να αδυνατεί να επιτρέψει άμεση πρόσβαση (on-line) στη βάση δεδομένων του συστήματος ελέγχου του δικτύου. Πολύ συχνά, σε μεγάλα συστήματα, όπως στην προκειμένη περίπτωση, το σύστημα ελέγχου του δικτύου έχει αναπτυχθεί με τη χρήση παρωχημένης τεχνολογίας, όπως, για παράδειγμα, αρχεία COBOL. Εφαρμογές που χρησιμοποιούν μοντέρνα τεχνολογία αντιμετωπίζουν προβλήματα στο να διαχειριστούν πληροφορία που προέρχεται από μια βάση δεδομένων παλαιάς τεχνολογίας.
2. Η βάση δεδομένων του συστήματος ελέγχου του δικτύου σχεδιάστηκε με βάση αποκλειστικά τις απαιτήσεις αυτής της εφαρμογής. Βασικό χαρακτηριστικό σε εφαρμογές αυτού του είδους είναι η όσο το δυνατό αποδοτικότερη ικανοποίηση μικρών δοσοληψιών που εισάγουν ή τροποποιούν πολύ μικρό αριθμό εγγραφών της βάσης. Μία τυπική δοσοληψία που θα αφορούσε τη χρέωση μίας υπεραστικής συνδιάλεξης θα εισήγαγε μία εγγραφή με τον κωδικό του συνδρομητή και τη διάρκεια της συνδιάλεξης. Στη σχεδίαση μιας τέτοιας βάσης δεδομένων, με την εφαρμογή των κανόνων κανονικοποίησης, καταλήγουμε σε μεγάλο αριθμό από πίνακες που ο κάθε ένας έχει περιορισμένο αριθμό πεδίων. Σε αντίθεση με τα παραπάνω, μία εφαρμογή που αντλεί στοιχεία λειτουργίας του δικτύου για λόγους ανάλυσης και λήψης αποφάσεων, δεν κάνει καμία αλλαγή στη βάση του δικτύου αλλά απαιτεί αποδοτική απόκριση από το σύστημα στις ερωτήσεις που θέτει. Αυτές οι ερωτήσεις συνήθως απαιτούν πρόσβαση σε μεγάλο αριθμό δεδομένων, θέτοντας διαφορετικούς κανόνες σχεδίασης της βάσης δεδομένων του συστήματος. Για μία ερώτηση σχετική με τη στρατηγική του οργανισμού, που θα είχε πρόσβαση σε μεγάλο αριθμό δεδομένων το κόστος σε μία βάση με πολλούς πίνακες θα ήταν σημαντικό, καθώς θα έπρεπε να εκτελεστεί μεγάλο αριθμός από πράξεις JOIN μεταξύ των πινάκων αυτών.
3. Κάθε σύστημα στήριξης αποφάσεων εκτελεί μεγάλο αριθμό ερωτήσεων, θα δεσμεύσει μεγάλο αριθμό πόρων του συστήματος διαχείρισης της βάσης δεδομένων με αποτέλεσμα να μειώσει την απόδοση του συστήματος ελέγχου του δικτύου. Για παράδειγμα, μία ερώτηση σχετική με τις χρεώσεις των πελατών του δικτύου, που απαιτεί κάποιο χρονικό διάστημα για να εκτελεστεί, θα κλείδωνε τον πίνακα με τις χρεώσεις των πελατών εμποδίζοντας οποιαδήποτε μεταβολή από το σύστημα ελέγχου (π.χ. μια νέα χρέωση).

Από τα παραπάνω γίνεται σαφές ότι είναι εξαιρετικά δυσχερής η χρήση των βάσεων δεδομένων των πληροφοριακών συστημάτων των οργανισμών από τα συστήματα στήριξης αποφάσεων. Όμως, η αποδοτική χρήση των συστημάτων στήριξης αποφάσεων απαιτεί όπως προαναφέρθηκε, πρόσβαση σε αυτά τα δεδομένα. Η εισαγωγή των "Αποθηκών Δεδομένων" είναι η λύση στο κρίσιμο αυτό πρόβλημα.

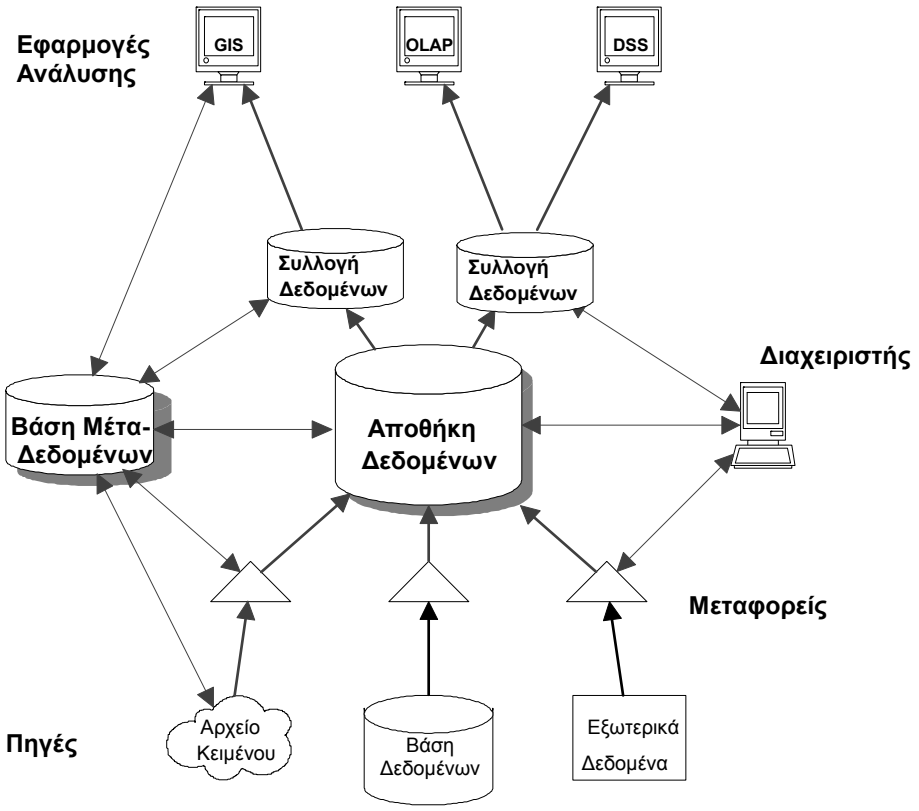
Με τον όρο *Αποθήκες Δεδομένων (Data Warehouses)* χαρακτηρίζουμε ένα σύνολο τεχνολογιών που επιτρέπει στους αναλυτές ενός οργανισμού στη σχεδίαση της πολιτικής του έχοντας αποδοτική πρόσβαση στα δεδομένα του οργανισμού. Μία Αποθήκη Δεδομένων διατηρεί δεδομένα που αντλεί από τις βάσεις δεδομένων των πληροφοριακών συστημάτων του οργανισμού αλλά και άλλες πηγές δεδομένων, όπως αρχεία του οργανισμού ή δεδομένα που προέρχονται από εξωτερικές πηγές. Αυτά τα δεδομένα οργανώνονται στην Αποθήκη Δεδομένων σε δομές κατάλληλες να απαντήσουν τις απαιτήσεις των αναλυτών - χρηστών των συστημάτων στήριξης αποφάσεων. Τα συστήματα στήριξης αποφάσεων αποκτούν πρόσβαση στα δεδομένα λειτουργίας του οργανισμού χωρίς την παρουσία των προαναφερθέντων προβλημάτων. Οι Αποθήκες Δεδομένων παρέχουν τη δυνατότητα για *Συνεχή Αναλυτική Επεξεργασία (On-Line Analytical Processing- OLAP)* των δεδομένων περιέχοντας συνήθως ιστορικά και συγκεντρωτικά δεδομένα που συνήθως αποδεικνύονται χρήσιμα για υποστήριξη αποφάσεων. Επίσης, παρέχουν μία ολοκληρωμένη εικόνα του σχήματος των δεδομένων του οργανισμού. Η σχεδίαση των Αποθηκών Δεδομένων έχει σαν στόχο την αποδοτική απάντηση των πολύπλοκων ερωτήσεων που θέτονται κατά την αναλυτική επεξεργασία δεδομένων από τις εφαρμογές στρατηγικού σχεδιασμού.

Η δημιουργία και η συντήρηση μίας Αποθήκης Δεδομένων είναι μία πολύπλοκη διαδικασία καθώς πολλές διαφορετικές προσεγγίσεις είναι εφικτές. Αρκετοί οργανισμοί επιδιώκουν να δημιουργήσουν μία Αποθήκη Δεδομένων που θα περιέχει αναλυτικά δεδομένα από όλες τις δραστηριότητες του οργανισμού. Πρόκειται για ένα πολύπλοκο εγχείρημα που απαιτεί μεγάλο κόστος για να επιτύχει. Μία άλλη λύση είναι η δημιουργία *Επιμέρους Συλλογών Δεδομένων (data marts)* με κριτήριο το αντικείμενο των εφαρμογών από τις οποίες προέρχονται ή το τμήμα του οργανισμού που τις χρησιμοποιεί. Πρόκειται για πιο ευέλικτα συστήματα στη δημιουργία τους, τα οποία όμως δεν παρέχουν ενιαία λύση, δημιουργώντας προβλήματα σε περίπτωση μακρόχρονης χρήσης τους.

Τα τελευταία χρόνια η ανάπτυξη και λειτουργία Αποθηκών Δεδομένων κρίνεται κρίσιμη για την λειτουργία των οργανισμών. Τεράστια ποσά επενδύονται σε αυτή τη δραστηριότητα ενώ τα οφέλη από τη λειτουργία τέτοιων συστημάτων κρίνονται ήδη ως ιδιαίτερα σημαντικά. Όπως είναι φυσικό, όλες οι μεγάλες εταιρείες του χώρου των Βάσεων Δεδομένων και των πληροφοριακών συστημάτων αναπτύσσουν και προτείνουν προϊόντα στο χώρο των Αποθηκών Δεδομένων. Τα επόμενα χρόνια αναμένονται ακόμα μεγαλύτερες επενδύσεις σε τεχνολογία αιχμής του χώρου. Για την παρουσίαση του κεφαλαίου αυτού, στηριχτήκαμε κυρίως στα [BS96], [CD 96], [CD97], [Coll96], [Inm96], [Kena95], [RedB97], [Wido95].

8.2 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΘΗΚΩΝ ΔΕΔΟΜΕΝΩΝ

Η επιλογή της αρχιτεκτονική μιας αποθήκης δεδομένων πρέπει να ικανοποιεί τις συγκεκριμένες ανάγκες του οργανισμού για τις οποίες δημιουργήθηκε και να εξασφαλίζει τη διαθεσιμότητα και την αποδοτικότητα του συστήματος. Το Σχήμα 8.1 παρουσιάζει μια γενική αρχιτεκτονική ενός συστήματος Αποθήκης Δεδομένων. Στο σχήμα σημειώνονται τα βασικά δομικά στοιχεία μίας Αποθήκης Δεδομένων, η διασύνδεση των στοιχείων τους, καθώς και η ροή των δεδομένων.



Σχήμα 8.1 : Γενική Αρχιτεκτονική Αποθήκης Δεδομένων

Τα δομικά μέρη της αρχιτεκτονικής ενός συστήματος Αποθήκης Δεδομένων είναι τα ακόλουθα:

Πηγές: Κάθε πηγή από την οποία η Αποθήκη Δεδομένων αντλεί δεδομένα.

Μεταφορείς - Μετατροπείς: Εφαρμογές που εκτελούν τις διαδικασίες μεταφοράς των δεδομένων από τις πηγές στην Αποθήκη Δεδομένων.

Αποθήκη Δεδομένων, Συλλογές Δεδομένων: Τα συστήματα που αποθηκεύονται τα δεδομένα που παρέχονται προς τους χρήστες.

Βάση Μετα-Δεδομένων: Σύστημα αποθήκευσης πληροφορίας σχετικά με τη δομή και λειτουργία του συστήματος.

Διαχειριστής: Εφαρμογή που παρέχει δυνατότητα διαχείρισης του συστήματος

Εφαρμογές Ανάλυσης: Εφαρμογές που έχουν πρόσβαση στην Αποθήκη Δεδομένων. Συνήθως είναι συστήματα στήριξης αποφάσεων.

8.2.1 Πηγές και Μεταφορείς - Μετατροπείς

Τα συστήματα διαχείρισης Αποθηκών Δεδομένων υποστηρίζουν άντληση δεδομένων από διάφορες κατηγορίες πηγών δεδομένων. Οι συνηθέστερες από αυτές είναι:

- Βάσεις Δεδομένων των συστημάτων του οργανισμού.
- Εξωτερικές πηγές πληροφοριών, όπως για παράδειγμα, πληροφορίες που παρέχονται από πληροφοριακά συστήματα στα οποία υπάρχει πρόσβαση από τον οργανισμό.
- Αρχεία Εφαρμογών και αρχεία κειμένου.

Οι *Μεταφορείς / Μετατροπείς δεδομένων (wrappers / loaders)* είναι εφαρμογές που εξάγουν δεδομένα από τις πηγές και τα μεταφέρουν στην Αποθήκη Δεδομένων. Η ύπαρξη διαφορετικών κατηγοριών (Σχεσιακές Βάσεις Δεδομένων, αρχεία COBOL, κείμενα MS-Word) που παρέχουν διαφορετική πρόσβαση στα δεδομένα τους οδηγεί στην ανάπτυξη διαφορετικών τύπων μεταφορέων. Συνήθως, για κάθε μία διαφορετική πηγή, ή κατηγορία πηγής, ένας διαφορετικός μεταφορέας αναλαμβάνει να αντλεί τα δεδομένα της,. Η λειτουργία αυτών των εφαρμογών κρίνεται ιδιαίτερα κρίσιμη για την επιτυχία του συστήματος, καθώς είναι υπεύθυνες για την αυτόματη μεταφορά, την επεξεργασία και τις αναγκαίες μετατροπές των δεδομένων από τις πηγές. Αναλυτικά, οι μεταφορείς αυτοματοποιούν τις παρακάτω διαδικασίες:

- Εξαγωγή δεδομένων από τις πηγές.
- Καθαρισμό των δεδομένων με την διάγνωση πιθανών ασυνεπειών και τη μεταφορά μόνο των πραγματικά χρήσιμων δεδομένων.
- Μετάδοση δεδομένων σε υψηλές ταχύτητες.
- Μετατροπή των δεδομένων μεταξύ διαφορετικών μοντέλων και προτύπων.
- Διάγνωση αλλαγών στα δεδομένα των πηγών και μεταφορά των νέων δεδομένων
- Εισαγωγή των δεδομένων στην Αποθήκη Δεδομένων.
- Δημιουργία αντιγράφων τμημάτων των πηγών στην Αποθήκη Δεδομένων.
- Ανάλυση των μεταφερόμενων δεδομένων για τη διάγνωση μη ορθής πληροφορίας.
- Έλεγχος πληρότητας Δεδομένων.

8.2.2 Αποθήκη Δεδομένων, Συλλογές Δεδομένων

Οι Αποθήκες Δεδομένων και οι Συλλογές Δεδομένων, όπως φαίνεται στο Σχήμα 8.1, υλοποιούνται με τη χρήση Σχεσιακών Συστημάτων Διαχείρισης Βάσεων Δεδομένων. Τα δεδομένα αποθηκεύονται σε σχεσιακές βάσεις δεδομένων, ενώ πρόσβαση σε αυτά παρέχεται από μία γλώσσα διαχείρισης δεδομένων που είναι επέκταση της SQL. Εναλλακτική της χρήσης σχεσιακών συστημάτων είναι η χρήση των *Πολυδιάστατων Συστημάτων Αναλυτικής Επεξεργασίας (Multidimensional OLAP servers),* που αποθηκεύουν και διαχειρίζονται δεδομένα με πολυδιάστατο τρόπο (βλ. και ενότητα 8.6). Η χρήση σχεσιακών ΣΔΒΔ εκμεταλλεύεται την ευελιξία και την ισχύ της τεχνολογίας των σύγχρονων συστημάτων. Κατά την αναλυτική επεξεργασία δεδομένων εκτελούνται πολύπλοκες ερωτήσεις που απαιτούν δυνατότητα διαχείρισης μεγάλου όγκου πληροφοριών. Τα πλεονεκτήματα των πολυδιάστατων συστημάτων βρίσκονται στη δυνατότητά τους να διαχειρίζονται δεδομένα, τα οποία είναι δομημένα με τρόπο που βρίσκεται πιο κοντά στις ανάγκες των εφαρμογών ανάλυσης (OLAP).

Η ύπαρξη των Συλλογών Δεδομένων είναι επιλογή του διαχειριστή του συστήματος. Οι Συλλογές Δεδομένων περιέχουν τμήματα των δεδομένων της Αποθήκης Δεδομένων. Ο καταμερισμός του περιεχομένου των Αποθηκών σε επιμέρους Συλλογές γίνεται με οργανωτικά κριτήρια και στόχο την πιο άμεση και αποδοτική πρόσβαση των εφαρμογών ανάλυσης στα δεδομένα της Αποθήκης καθώς και στον καταμερισμό των δεδομένων κατά αντικείμενο ή τμήμα. Παράλληλα, επιτυγχάνεται και η αποσυμφόρηση της Αποθήκης Δεδομένων.

8.2.3 Βάση Μετα-Δεδομένων

Τα *Μετα-Δεδομένα (metadata)*, έχουν ένα πολύ σημαντικό ρόλο στις Αποθήκες Δεδομένων. Η κατανόηση και η καταγραφή του περιεχομένου των δεδομένων και της οργάνωσης τους είναι απαραίτητη για τη αποδοτική λειτουργία και διαχείριση της Αποθήκης. Τα μετα-δεδομένα περιέχουν (ή καλύτερα, οφείλουν να περιέχουν):

- *Λεξικό Δεδομένων (Data Dictionary)* που περιέχει τον ορισμό και την περιγραφή των δεδομένων που αποθηκεύονται στην Αποθήκη Δεδομένων και τις μεταξύ τους συσχετίσεις.
- Περιγραφή της ροής των δεδομένων μέσα στο σύστημα.
- Περιγραφή των κανόνων μετατροπής των δεδομένων κατά τη μεταφορά τους.
- Δεδομένα ελέγχου των διαφόρων εκδοχών (versions) των δεδομένων.
- Στατιστικά χρήσης των δεδομένων.
- Πληροφορία σχετικά με τους κανόνες ελέγχου πρόσβασης στην Αποθήκη Δεδομένων.
- Διάφορα ψευδώνυμα (aliases).

Όπως φαίνεται και στη Σχήμα 8.1, τα μετα-δεδομένα αποθηκεύονται σε ένα σύστημα, όπου υπάρχει πρόσβαση από κάθε δομικό στοιχείο της αρχιτεκτονικής. Το γεγονός αυτό δημιουργεί την ανάγκη ύπαρξης ενός σταθερού προτύπου για τα μετα-δεδομένα, καθώς, όπως προαναφέρθηκε, τα διάφορα δομικά στοιχεία που συμμετέχουν στην αρχιτεκτονική των Αποθηκών Δεδομένων είναι εφαρμογές ανεπτυγμένες ανεξάρτητα από τις πηγές και τις εφαρμογές ανάλυσης. Ένα τέτοιο πρότυπο έχει προταθεί από μια ομάδα εταιρειών του χώρου και ονομάζεται *Metadata Interchange Specification (MDIS)*.

Οι Αποθήκες Δεδομένων, σε μερικές περιπτώσεις, είναι κατανεμημένες ώστε να πετυχαίνεται καταμερισμός του φορτίου, επεκτασιμότητα και διαθεσιμότητα του συστήματος. Σε αυτές τις κατανεμημένες αρχιτεκτονικές, υπάρχει συχνά ένα αντίγραφο του συστήματος των μετα-δεδομένων σε κάθε ένα από τους κατανεμημένους κόμβους της Αποθήκης, ενώ η όλη διαχείριση του συστήματος γίνεται από μία κεντρική εφαρμογή.

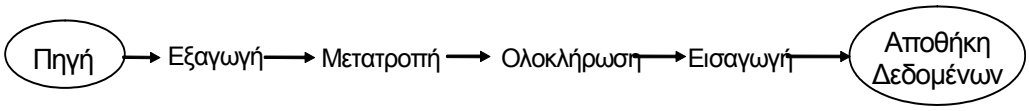
8.2.4 Σχεδίαση αρχιτεκτονικής Αποθηκών Δεδομένων

Η σχεδίαση μίας Αποθήκης Δεδομένων είναι μία πολύπλοκη διαδικασία που αποτελείται συνήθως από τις παρακάτω ενέργειες:

- Ορισμός της αρχιτεκτονικής και των απαιτούμενων στοιχείων του συστήματος. Επιλογή του κατάλληλου εξοπλισμού σε μηχανήματα, συστήματα Βάσεων Δεδομένων και εργαλείων λογισμικού.
- Εγκατάσταση επικοινωνίας μεταξύ των servers και των εργαλείων ανάλυσης
- Σχεδίαση του σχήματος της Αποθήκης Δεδομένων
- Δημιουργία της φυσικής οργάνωσης της Αποθήκης Δεδομένων, υλοποίηση των σχετικών δομών και των μεθόδων πρόσβασης στην Αποθήκη.
- Σχεδίαση και ανάπτυξη των προγραμμάτων που εκτελούν τη μεταφορά δεδομένων.
- Εγκατάσταση των μεταφορέων και σύνδεση με τις πηγές δεδομένων.
- Δημιουργία της Βάσης των Μετα-δεδομένων.
- Ολοκλήρωση των εφαρμογών ανάλυσης.

8.3 ΜΕΤΑΦΟΡΑ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΙΣ ΠΗΓΕΣ ΣΤΗΝ ΑΠΟΘΗΚΗ ΔΕΔΟΜΕΝΩΝ

Βασικός παράγοντας για την επιτυχία των Αποθηκών Δεδομένων είναι η ορθή τροφοδοσία της Αποθήκης Δεδομένων από τις πηγές. Η διαδικασία μεταφοράς δεδομένων από τις πηγές στην Αποθήκη δεδομένων είναι αρκετά πολύπλοκη καθώς πολλά προβλήματα πρέπει να αντιμετωπισθούν. Τα βήματα που ακολουθούνται κατά τη μεταφορά των δεδομένων παρουσιάζονται στο Σχήμα 8.2



Σχήμα 8.2 : Διαδικασία μεταφοράς δεδομένων

8.3.1 Εξαγωγή και Μετατροπή Δεδομένων

Η Εξαγωγή και η Μετατροπή δεδομένων εκτελούνται από τους Μεταφορείς / Μετατροπείς του Σχήματος 8.1. Για κάθε πηγή που χρησιμοποιούμε στο σύστημα εγκαθιστούμε λογισμικό που αντλεί τα δεδομένα από την πηγή, τα "καθαρίζει", κρατώντας μόνο αυτά που είναι πραγματικά χρήσιμα και τα μετασχηματίζει με βάση ένα καθορισμένο πρότυπο. Οι μετατροπές που γίνονται στα δεδομένα αφορούν τόσο τη δομή όσο και την τιμή τους. Για παράδειγμα, το πεδίο "Ημερομηνία" ενός πίνακα μπορεί να μετασχηματιστεί στα πεδία "Χρόνος", "Μήνας" και "Ημέρα", ενώ οι τιμές του πεδίου "Χαρακτηρισμός" είναι πιθανόν να μετατραπούν από "Α", "Β" κλπ σε "1", "2" κλπ αντίστοιχα. Αυτό το λογισμικό υλοποιείται με βάση τα ιδιαίτερα χαρακτηριστικά κάθε πηγής και εγκαθίσταται σε υπολογιστές με άμεση πρόσβαση στα δεδομένα της πηγής. Οι Αποθήκες Δεδομένων χρησιμοποιούν ποικίλα εργαλεία για εξαγωγή. Η εξαγωγή δεδομένων από τις απομακρυσμένες πηγές συχνά υλοποιείται μέσω πυλών (gateways) και καθιερωμένων προτύπων διασύνδεσης εφαρμογών (όπως ODBC, Oracle Open Connect, Information Builders EDA/SQL κλπ).

Εξωτερικά εργαλεία που εγκαθίστανται για κάθε διαφορετική πηγή δεδομένων αναλαμβάνουν την εξαγωγή των δεδομένων από τις πηγές. Παράλληλα εκτελούν και μία πρώτη επεξεργασία των δεδομένων αυτών. Καθώς οι Αποθήκες Δεδομένων χρησιμοποιούνται για στρατηγικές αποφάσεις, επιβάλλεται να περιέχουν σωστά δεδομένα. Στις διάφορες πηγές όπου υπάρχει μεγάλος όγκος δεδομένων είναι πολύ πιθανόν να υπάρχουν λάθη ή ανωμαλίες. Διάφορα εργαλεία βοηθούν στη διάγνωση των ανωμαλιών των δεδομένων και στη διόρθωσή τους όπου αυτό είναι εφικτό. Ως περιπτώσεις όπου ο καθαρισμός των δεδομένων είναι σημαντικός αναφέρονται: ασυνέπειες στο μήκος των πεδίων διαφορετικών πηγών, ασυνέπειες σχετικά με την περιγραφή των δεδομένων, ασυνεπείς τιμές δεδομένων, απουσίες εγγραφών και παραβίαση περιορισμών ακεραιότητας.

8.3.2 Ολοκλήρωση

Η διαδικασία της *ολοκλήρωσης (integration)* των δεδομένων είναι αρκετά πολύπλοκη και περιλαμβάνει τη δημιουργία και συντήρηση ενός καθολικού ιδεατού σχήματος των δεδομένων των πηγών. Αυτό το σχήμα περιλαμβάνει κάθε οντότητα που παρέχει δεδομένα από οποιαδήποτε πηγή της Αποθήκης Δεδομένων. Η Βάση των μετα-δεδομένων ενημερώνεται και συντηρεί το καθολικό σχήμα. Με βάση το καθολικό σχήμα, κάθε ποσότητα δεδομένων που έρχεται από τις πηγές πρέπει να μετασχηματιστεί ώστε να εισαχθεί στην Αποθήκη Δεδομένων.

Για παράδειγμα, σε ένα τηλεπικοινωνιακό οργανισμό, είναι πιθανόν δύο συστήματα να διαχειρίζονται χρεώσεις από διαφορετικούς πελάτες. Στις βάσεις και των δύο πληροφοριακών συστημάτων υπάρχει πίνακας "Χρέωση" με πιθανά διαφορετικό ορισμό στο κάθε σύστημα.. Μετά τη διαδικασία ολοκλήρωσης των πηγών στο καθολικό σχήμα, υπάρχει επίσης η οντότητα "Χρέωση", ο ορισμός της οποίας προκύπτει μεν από τους επιμέρους ορισμούς, αλλά πιθανά δεν ακολουθεί επακριβώς κάποιον από τους δύο ή και τους δύο. Οι εγγραφές που αντιστοιχούν σε χρεώσεις που γίνονται στα δύο συστήματα θα πρέπει να τροποποιηθούν για να εισαχθούν στην Αποθήκη Δεδομένων.

8.3.3 Εισαγωγή δεδομένων

Τελευταίο στάδιο στη μεταφορά των δεδομένων από τις πηγές στην Αποθήκη Δεδομένων είναι η διαδικασία εισαγωγής. Κατά τη διάρκεια της εισαγωγής, τα δεδομένα επεξεργάζονται ώστε να ελεγχθούν οι περιορισμοί ακεραιότητας της Αποθήκης Δεδομένων και να γίνουν υπολογισμοί πάνω στα δεδομένα όπως αθροιστικές πράξεις και ομαδοποιήσεις. Από το αποτέλεσμα αυτών των πράξεων θα προκύψουν τα δεδομένα που θα καταγραφούν στην Αποθήκη Δεδομένων, ενώ παράλληλα ενημερώνονται τα ευρετήρια της βάσης της Αποθήκης. Κατά τη διάρκεια της εισαγωγής των δεδομένων, πρέπει να παρέχεται η δυνατότητα στο διαχειριστή της Αποθήκης δεδομένων να παρακολουθεί και να επεμβαίνει στην όλη διαδικασία. Καθώς η διαδικασία εισαγωγής δεδομένων έχει μεγάλο υπολογιστικό κόστος και στις πηγές, αλλά και στην αποθήκη δεδομένων, η διαδικασία αυτή γίνεται μαζικά σε περιοδικά χρονικά διαστήματα όπου δεν υπάρχει φορτίο στο σύστημα.

8.3.4 Ενημέρωση

Η ενημέρωση της Αποθήκης Δεδομένων είναι η διαδικασία που μεταφέρει τις αλλαγές που συμβαίνουν στα δεδομένα των πηγών εκτελώντας αντίστοιχες αλλαγές στα δεδομένα της Αποθήκης. Η διαδικασία αυτή ακολουθεί όλα τα παραπάνω βήματα (εξαγωγή, μετατροπή, ολοκλήρωση, εισαγωγή). Υπάρχουν όμως και μερικά επιπλέον ζητήματα που προκύπτουν από τη δυνατότητα διάγνωσης των μεταβολών στις πηγές, καθώς και από τον όγκο των δεδομένων που τροποποιούνται. Συνήθως, οι Αποθήκες Δεδομένων ενημερώνονται περιοδικά. Υπάρχουν όμως και περιπτώσεις όπου εφαρμογές ανάλυσης απαιτούν άμεση πρόσβαση σε τρέχοντα δεδομένα, οπότε επιβάλλεται η άμεση ενημέρωση των Αποθηκών για κάθε μεταβολή στις πηγές. 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.

Οι τεχνικές ενημέρωσης εξαρτώνται από τα χαρακτηριστικά των πηγών. Πολλές φορές είναι δυνατή μόνο η εξαγωγή ενός ολόκληρου αρχείου ή μία βάσης δεδομένων από μία πηγή. Σε αυτή την περίπτωση, η ενημέρωση της Αποθήκης θα ισοδυναμούσε με διαγραφή όλων των δεδομένων που σχετίζονται με τη πηγή και επαναεισαγωγή των εξαγχθέντων δεδομένων. Πρόκειται για μία καθόλου αποδοτική λύση, που όμως πολλές φορές είναι και η μοναδική, όταν η πηγή αδυνατεί να μας δώσει πληροφορίες για τις μεταβολές που συντελούνται σε αυτή.

Δεδομένου ότι οι Αποθήκες Δεδομένων συσσωρεύουν μεγάλη ποσότητα δεδομένων, οι εφαρμογή της παραπάνω μεθόδου ενημέρωσης καθίσταται απαγορευτική. Γιαυτό και είναι αναγκαία η διάγνωση των μεταβολών που συμβαίνουν στις πηγές (εισαγωγές, διαγραφές και τροποποιήσεις εγγραφών), ώστε σε κάθε διαδικασία ενημέρωσης να μην γίνονται περιττές διαγραφές και εισαγωγές δεδομένων που στην πραγματικότητα παραμένουν αναλλοίωτα. Σύμφωνα με τις τεχνικές προοδευτικής ενημέρωσης και συντήρησης των δεδομένων, εισάγονται στην Αποθήκη νέες εγγραφές που προκύπτουν αποκλειστικά από αντίστοιχες εισαγωγές δεδομένων στις πηγές. Ομοίως, και οι διαγραφές και τροποποιήσεις εγγραφών προκύπτουν από αντίστοιχες πράξεις δεδομένων στις πηγές.

Για να γίνει εφικτή η εφαρμογή της προοδευτικής ενημέρωσης πρέπει οι πηγές να μας δίνουν τη δυνατότητα διάγνωσης των μεταβολών που συντελούνται στα δεδομένα τους. Στις περιπτώσεις που η πηγή είναι ένα σύγχρονο σύστημα βάσης δεδομένων, υπάρχουν τρεις βασικές τεχνικές με τις οποίες είναι εφικτή η διάγνωση των μεταβολών αυτών:

- **Στιγμιότυπα**: Αρκετά συστήματα βάσεων δεδομένων είναι σε θέση να εξάγουν, όταν τους ζητηθεί, στιγμιότυπα (snapshots) από πίνακες της βάσης τους. Από τα στιγμιότυπα

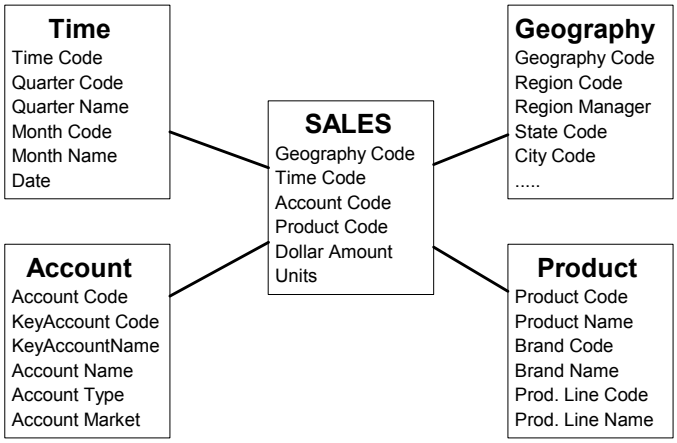
αυτά με διάφορες μεθόδους αποδοτικής σύγκρισης μπορούμε να διαγνώσουμε τις τροποποιήσεις που συνέβηκαν στην πηγή και να ενημερωθεί σχετικά η Αποθήκη.

- **Μηχανισμός καταγραφής (log)**: Τα περισσότερα σύγχρονα συστήματα βάσεων δεδομένων καταγράφουν όλες τις μεταβολές των δεδομένων τους και τις πράξεις που τις προκαλούν ώστε να μπορούν να παρέχουν ομαλή εκτέλεση των δοσοληψιών. Οι μεταφορείς που εξάγουν τα δεδομένα από τέτοιες πηγές μπορούν να έχουν άμεση πρόσβαση στις μεταβολές που συντελούνται, αν τους δοθεί η δυνατότητα πρόσβασης στο αρχείο (log file) που καταγράφονται αυτές οι μεταβολές.
- **Triggers**: Σε περίπτωση που μία πηγή είναι ένα μοντέρνο σύστημα που παρέχει τη δυνατότητα δημιουργίας triggers, μπορούμε για κάθε πίνακα της πηγής να δημιουργήσουμε έναν trigger που θα μας ενημερώνει για οποιαδήποτε μεταβολή συμβαίνει στον πίνακα αυτόν.

8.4 ΣΧΕΔΙΑΣΗ ΑΠΟΘΗΚΩΝ ΔΕΔΟΜΕΝΩΝ

Καθώς οι Αποθήκες Δεδομένων χρησιμοποιούνται αποκλειστικά για την απάντηση των ερωτήσεων των εφαρμογών ανάλυσης, η σχεδίαση και η οργάνωση των δεδομένων είναι διαφορετική από τις κλασικές βάσεις δεδομένων. Τα διαγράμματα Οντοτήτων - Συσχετίσεων και οι τεχνικές κανονικοποίησης είναι οι κλασικές μέθοδοι για τη σχεδίαση των βάσεων δεδομένων των συστημάτων επεξεργασίας δοσοληψιών (OLTP). Αυτές οι μέθοδοι αποδεικνύονται συχνά ακατάλληλες για τη σχεδίαση των Αποθηκών Δεδομένων, καθώς ο στόχος τους είναι να αντιμετωπίσουν προβλήματα, όπως ο πλεονασμός (redundacy) ή η ανανέωση των δεδομένων. Επιπλέον, αυτές οι μέθοδοι οδηγούν στη δημιουργία πολλών πινάκων με μικρό αριθμό πεδίων, σχήμα που έχει σαν αποτέλεσμα την εκτέλεση μεγάλου αριθμού από πράξεις JOIN, στην περίπτωση που θέλουμε να αντλήσουμε μεγάλο όγκο αναλυτικών πληροφοριών.

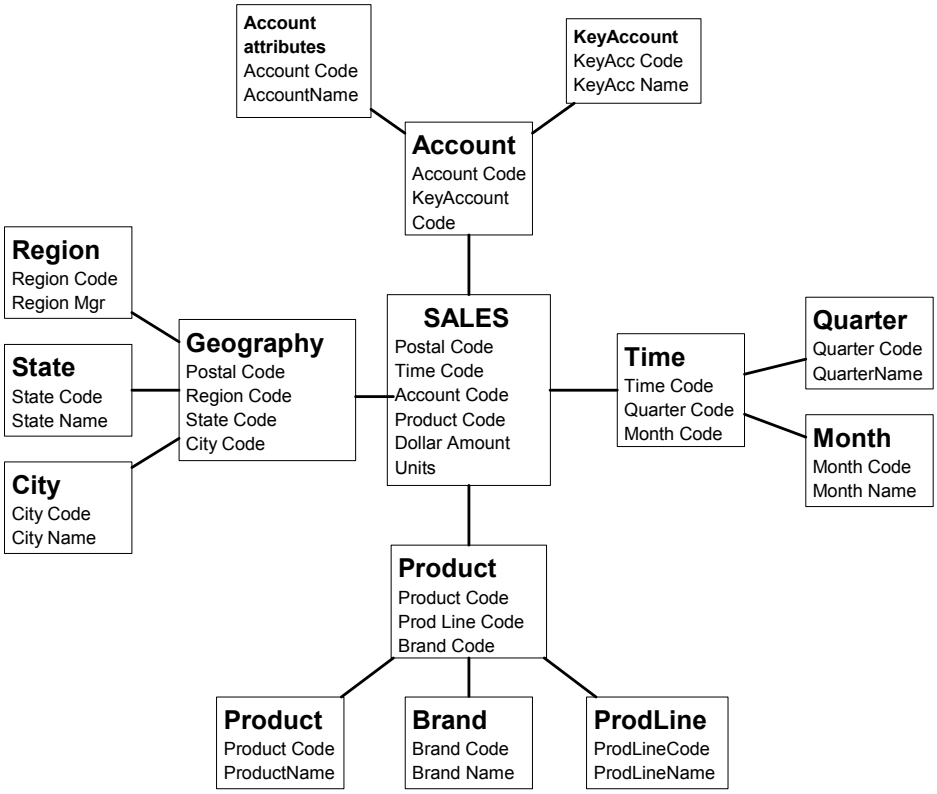
Οι πιο κατάλληλες τεχνικές για τη σχεδίαση των βάσεων των Αποθηκών Δεδομένων είναι τα *αστεροειδή σχήματα (star schemata)* και τα *σχήματα χιονονιφάδας (snowflake schemata)*. Το αστεροειδές σχήμα είναι πιο κοντά στο πολυδιάστατο χαρακτήρα των δεδομένων. Σε μία αστεροειδή βάση, υπάρχει ένας βασικός πίνακας που χαρακτηρίζεται *πίνακας συμβάντων (fact table)*. Υπάρχει επίσης ένας πίνακας για κάθε μία διάσταση. Κάθε εγγραφή του πίνακα συμβάντων αποτελείται από ένα δείκτη (ξένο κλειδί) σε μία εγγραφή κάθε ενός από τους πίνακες διαστάσεων. Κάθε *πίνακας διάστασης (dimension table)* περιλαμβάνει εγγραφές που αντιστοιχούν σε τιμές των διαστάσεων.



Σχήμα 8.3 : Παράδειγμα αστεροειδούς σχήματος

Στο Σχήμα 8.3 παρουσιάζεται ένα παράδειγμα αστεροειδούς σχήματος. Πρόκειται για το σχήμα Αποθήκης Δεδομένων που περιλαμβάνει δεδομένα σχετικά με τις πωλήσεις προϊόντων σε διάφορες πόλεις. Ο πίνακας των πωλήσεων (SALES) είναι στην προκειμένη περίπτωση ο πίνακας συμβάντων. Παρατηρούμε -για παράδειγμα- ότι μπορεί να εκτελεστεί οποιαδήποτε ερώτηση που συσχετίζει πωλήσεις, με τους λογαριασμούς (accounts) που παράγονται τα προϊόντα και με τις γεωγραφικές περιοχές που αυτά πωλούνται, εκτελώντας μόνο δύο πράξεις JOIN.

Η κύρια αδυναμία των αστεροειδών σχημάτων εντοπίζεται στον τρόπο με τον οποίο εκφράζουν τις ιεραρχίες των διαστάσεων. Για παράδειγμα, στη διάσταση χρόνος υπάρχει μία προφανής ιεραρχία μεταξύ ημερών, μηνών ετών κλπ. Μια εναλλακτική μοντελοποίηση των ιεραρχιών γίνεται από τα σχήματα χιονονιφάδας. Στο Σχήμα 8.4 παρουσιάζεται το παράδειγμα βάσης ίδιου περιεχομένου με τη βάση του Σχήματος 8.3, αλλά οργανωμένη σύμφωνα με το σχήμα χιονονιφάδας. Πρόκειται για μία βελτίωση του αστεροειδούς σχήματος, όπου η ιεραρχία των διαστάσεων αναπαριστάται κανονικοποιώντας τους πίνακες των διαστάσεων.



Σχήμα 8.4 Παράδειγμα σχήματος χιονονιφάδας.

Σε περιπτώσεις σχεδίασης Αποθηκών Δεδομένων με δεδομένα πολύπλοκης δομής είναι πιθανό, περισσότεροι του ενός πίνακες συμβάντων να έχουν κοινούς πίνακες διαστάσεων. Για παράδειγμα, οι παραγγελίες και οι πωλήσεις έχουν κοινές τις περισσότερες διαστάσεις.

Εκτός από τους πίνακες συμβάντων και διαστάσεων, είναι πιθανόν να υπάρχουν και επιπρόσθετοι πίνακες με συγκεντρωτικά, προ-υπολογισμένα δεδομένα στην Αποθήκη Δεδομένων. Στην πιο απλή περίπτωση, τα συγκεντρωτικά δεδομένα αντιστοιχούν στην ομαδοποίηση των εγγραφών των πινάκων συμβάντων στη βάση συνδυασμού διαστάσεων. Στη βάση του παραδείγματος των σχημάτων 8.3 και 8.4 μπορεί να προστεθούν πίνακες οι οποίοι να περιέχουν τις συνολικές πωλήσεις προϊόντων ανά γεωγραφική περιοχή και μονάδα χρόνου. Στην πραγματικότητα, δηλαδή, πρόκειται για πίνακες που περιέχουν πληροφορία που προκύπτει από τα δεδομένα του πίνακα συμβάντων. Η σκοπιμότητα ύπαρξης αυτών των πινάκων κρίνεται από την άφιξη σχετικών ερωτήσεων στη βάση από τις εφαρμογές ανάλυσης. Εναλλακτική της δημιουργίας τέτοιων πινάκων, είναι η εισαγωγή στους πίνακες συμβάντων εγγραφών που θα περιέχουν συγκεντρωτικές πληροφορίες για μερικές από τις διαστάσεις. Στην περίπτωση που θέλουμε να εισάγουμε στον πίνακα των πωλήσεων (SALES) συγκεντρωτικές τιμές ανά περιοχή και μονάδα χρόνου, θα εισάγουμε εγγραφές που

- τα πεδία που αντιστοιχούν στις διαστάσεις θα έχουν τις αντίστοιχες τιμές,
- τα πεδία που αντιστοιχούν στη τιμή που μετρά κάθε εγγραφή (units) θα υπάρχει η συγκεντρωτική τιμή, και

- στα πεδία των διαστάσεων που αθροίζονται (που δεν μας ενδιαφέρουν πλέον, δηλαδή) θα υπάρχουν τιμές NULL.

Η παρακάτω εγγραφή αντιστοιχεί στις συνολικές πωλήσεις του Ιουνίου στην περιοχή με κωδικό 16232 που παρέχει η αποθήκη δεδομένων του σχήματος 8.3 και 8.4.

Postal code	**Time Code**	**Account Code**	**Product Code**	**Dollar Amount**	**Units**
16232	Ιούλιος 1997	*null*	*null*	1654000	6520

8.5 ΕΙΔΙΚΑ ΘΕΜΑΤΑ ΤΩΝ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΘΗΚΩΝ ΔΕΔΟΜΕΝΩΝ

Οι Αποθήκες Δεδομένων συνήθως περιέχουν εξαιρετικά μεγάλες ποσότητες δεδομένων. Η αποδοτική απάντηση των ερωτήσεων απαιτεί ευφυείς μεθόδους πρόσβασης και τεχνικές επεξεργασίας των ερωτήσεων. Οι συνήθεις λύσεις που δίνονται σχετίζονται με την ευρεία χρήση *ευρετηρίων (indexes)*. Η επιλογή των κατάλληλων ευρετηρίων που θα δημιουργηθούν είναι πολύ σημαντικό πρόβλημα της σχεδίασης της Αποθήκης Δεδομένων. Το επόμενο βήμα αφορά τη σωστή διαχείριση των παραπάνω δομών. Η βελτιστοποίηση των ερωτήσεων είναι επίσης ένα σημαντικό ζήτημα. Καθώς αρκετές από τις ερωτήσεις που θέτονται στο σύστημα δεν είναι αποδοτικό να απαντηθούν με τη χρήση των ευρετηρίων, είναι αναγκαία η βελτιστοποίηση ακόμα και των μεθόδων που εκτελούν σειριακή αναζήτηση. Οι δυνατότητες που παρέχουν τα παράλληλα συστήματα αποδεικνύονται συχνά αποτελεσματικές.

8.5.1 Δομές Ευρετηρίων και η χρήση τους

Υπάρχουν πολλές τεχνικές επεξεργασίας ερωτήσεων που εκμεταλλεύονται αποδοτικά τα ευρετήρια. Για παράδειγμα ερωτήσεις με πολλαπλές συνθήκες μπορούν να απαντηθούν με τη χρήση της τομής και ένωσης των δεδομένων ευρετηρίων. Αυτές οι πράξεις μπορούν να χρησιμοποιηθούν για σημαντική μείωση του κόστους απάντησης των ερωτήσεων, ενώ συχνά παρακάμπτεται η πρόσβαση στους πίνακες με τα δεδομένα.

Τα συστήματα Αποθηκών Δεδομένων χρησιμοποιούν bitmap ευρετήρια, που υποστηρίζουν αποδοτικά πράξεις όπως ένωση ή τομή. Θεωρήστε μια σελίδα σε φύλλο ενός ευρετηρίου που αντιστοιχεί στην τιμή Α. Μία τέτοια σελίδα περιέχει μία λίστα από διευθύνσεις εγγραφών που περιέχουν τη τιμή Α. Τα bitmap ευρετήρια δομούν τη λίστα των διευθύνσεων ως ένα διάνυσμα από δυαδικές τιμές (0,1), που έχει μία δυαδική μεταβλητή (bit) για κάθε εγγραφή. Η μεταβλητή αυτή παίρνει τιμή 1 ανν η εγγραφή στην οποία αντιστοιχεί περιέχει τη τιμή Α. Η αποδοχή των bitmap ευρετηρίων στηρίζεται στο γεγονός ότι οι αναπαράσταση της λίστας των διευθύνσεων των εγγραφών σε διάνυσμα από bits επιταχύνει πράξεις όπως σύνδεση, τομή, ένωση και ομαδοποίηση, καθώς αυτές μετατρέπονται σε λογικές πράξεις πάνω σε πίνακες από bits και εκτελούνται γρήγορα.

Εκτός από τα ευρετήρια τιμών σε ένα πίνακα, η δομή των αστεροειδών σχημάτων επιβάλει την χρήση των *ευρετηρίων σύνδεσης (join indices)*. Τα ευρετήρια αυτού του είδους παρέχουν τη συσχέτιση τη τιμής ενός ξένου κλειδιού ενός πίνακα με την αντίστοιχη τιμή του κλειδιού του πίνακα στον οποίο αναφέρεται. Σε μία βάση με αστεροειδές σχήμα μπορούμε να συσχετίσουμε τον πίνακα συμβάντων με τους πίνακες των διαστάσεων με τη χρήση των ευρετηρίων σύνδεσης. Για παράδειγμα, στη βάση του Σχήματος 8.3 μπορεί να υπάρχει ένα ευρετήριο

σύνδεσης στον κωδικό Postal Code που κρατά για κάθε διαφορετική πόλη τις εγγραφές στον πίνακα των συμβάντων που αντιστοιχούν στην πόλη αυτή.

8.5.2 Μετατροπή Πολύπλοκων Ερωτήσεων

Η εύρεση κατάλληλου μετασχηματισμού των ερωτήσεων ώστε να απαντιούνται αποδοτικά αποδεικνύεται επίσης αρκετά σημαντική. Στη περιοχή των Αποθηκών Δεδομένων συναντάμε κλασικά θέματα, όπως αυτό της επεξεργασίας των φωλιασμένων ερωτήσεων. Οι ερωτήσεις που περιέχουν φωλιασμένες υποερωτήσεις καταναλώνουν γενικά πολύ χρόνο για να απαντηθούν. Υπάρχουν αρκετές τεχνικές που μετασχηματίζουν τις φωλιασμένες και ερωτήσεις πολλών συνθηκών.

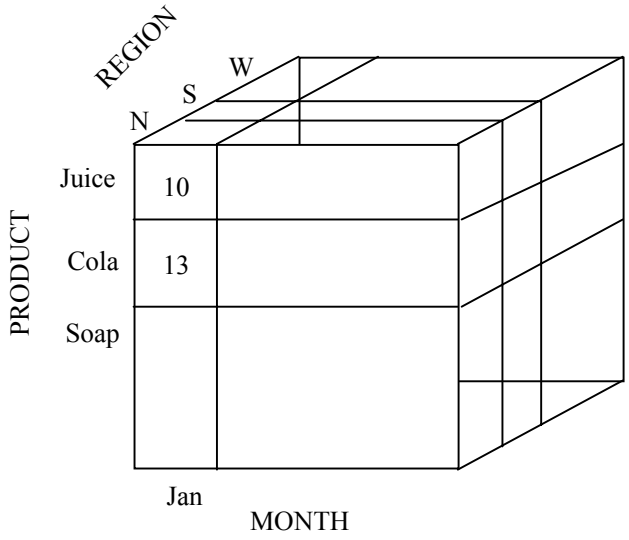
Η αποδοτική εκτέλεση των ερωτήσεων που περιλαμβάνουν πράξεις join μεταξύ πινάκων είναι επίσης αναγκαία. Ειδικές λύσεις προτείνονται για βάσεις με αστεροειδή σχήματα ή σχήματα χιονονιφάδας, καθώς η εκτέλεση ερωτήσεων στην Αποθήκη δεδομένων περιλαμβάνει, σε όλες σχεδόν τις περιπτώσεις, join μεταξύ του πίνακα συμβάντων και πινάκων διαστάσεων. Σε αρκετές περιπτώσεις εκτέλεσης join μεταξύ του πίνακα συμβάντων και περισσότερων του ενός πινάκων διαστάσεων, ακολουθείται η παρακάτω στρατηγική. Η αποθήκη εκτελεί ερωτήσεις και υπολογίζει το πλήρες καρτεσιανό γινόμενο μεταξύ των πινάκων των διαστάσεων (έχοντας πιθανά περιορίσει το εύρος των τιμών που λαμβάνουν μέρος με βάση τη συνθήκη επιλογής της ερώτησης). Κατόπιν εκτελεί μία απλή πράξη join μεταξύ του καρτεσιανού γινομένου και του πίνακα συμβάντων. Η παραπάνω μέθοδος χρησιμοποιείται για να αποφευχθεί η εκτέλεση πολλαπλών join στα οποία θα συμμετέχει ο πίνακας συμβάντων, ο οποίος συνήθως περιέχει συγκριτικά με τους πίνακες διαστάσεων, πολλαπλάσιο αριθμό εγγραφών.

8.6 ΜΟΝΤΕΛΑ ΣΥΣΤΗΜΑΤΩΝ ΑΝΑΛΥΤΙΚΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ

Η αναλυτική επεξεργασία δεδομένων είναι τμήμα των εφαρμογών στήριξης αποφάσεων και των στρατηγικών πληροφοριακών συστημάτων. Η λειτουργία της αναλυτικής επεξεργασίας χαρακτηρίζεται από δυναμική πολυδιάστατη ανάλυση των δεδομένων του οργανισμού. Οι εφαρμογές που εκτελούν συνεχή αναλυτική επεξεργασία (OLAP), χαρακτηρίζονται από τις συνεχής ερωτήσεις που εκτελούν πάνω στα δεδομένα του οργανισμού. Οι ερωτήσεις αυτές έχουν συγκεκριμένη και πολύπλοκη δομή, ενώ η πληροφορία που αντλούν έχει πολυδιάστατο χαρακτήρα. Για παράδειγμα, μια εφαρμογή που εκτελεί αναλυτική επεξεργασία στα δεδομένα των πωλήσεων ενός οργανισμού, συνεχώς εκτελεί ερωτήσεις για να μπορεί να έχει συγκεντρωτικά δεδομένα για τις πωλήσεις ανά προϊόν, ανά μήνα και ανά περιοχή. Η παρουσίαση των αποτελεσμάτων των πωλήσεων μπορεί να προκαλέσει τον χρήστη να εκτελέσει μία πιο συγκεντρωτική ερώτηση, ώστε να πάρει δεδομένα των ετήσιων πωλήσεων ανά προϊόν και περιοχή, ή να εκτελέσει μια πιο λεπτομερή ερώτηση παίρνοντας τις μηνιαίες πωλήσεις κάθε προϊόντος ανά συγκεκριμένο πελάτη.

8.6.1 Πολυδιάστατα Μοντέλα Δεδομένων

Οι πίνακες των σχεσιακών βάσεων δεδομένων περιέχουν εγγραφές οι οποίες αποτελούνται από πεδία. Σε φυσιολογικές σχεσιακές βάσεις δεδομένων, ένα υποσύνολο των πεδίων ενός πίνακα συνθέτουν το κλειδί του. Αντίθετα τα πολυδιάστατα μοντέλα δεδομένων περιέχουν ν-διάστατους πίνακες που συχνά αποκαλούνται *υπερκύβοι (cubes ή hypercubes)*. Κάθε διάσταση έχει μία ιεραρχία επιπέδων. Για παράδειγμα, η διάσταση "Γεωγραφική τοποθεσία" έχει τα επίπεδα πόλη, περιοχή, χώρα. Οι τιμές (μετρικές) που περιέχουν οι υπερκύβοι αντιστοιχούν στις στήλες των σχεσιακών πινάκων.



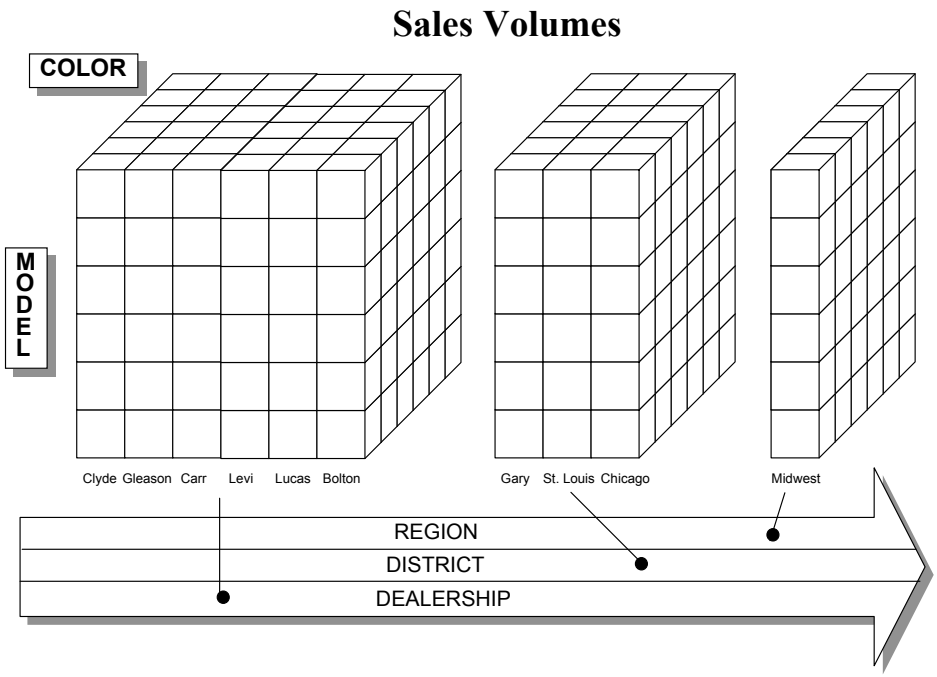
Σχήμα 8.5 Παράδειγμα υπερκύβου

Στο Σχήμα 8.5 παρουσιάζεται το μοντέλο των δεδομένων ενός υπερκύβου που παρέχει δεδομένα για τις πωλήσεις των προϊόντων. Σύμφωνα με το παράδειγμα οι πωλήσεις του προϊόντος "Cola" το μήνα Ιανουάριο στην βόρεια περιοχή ήταν 13. Οι τιμές "Cola", "JAN" και "N" στο συγκεκριμένο παράδειγμα είναι οι τιμές των διαστάσεων "PRODUCT", "MONTH" και "REGION" αντίστοιχα, ενώ το 13 αποτελεί τιμή των δεδομένων του υπερκύβου. Είναι πιθανόν η τιμή σε μία θέση του υπερκύβου να περιέχει συγκεντρωτική τιμή μίας μετρικής. Στο συγκεκριμένο παράδειγμα, ο αριθμός 13 αφορά τις πωλήσεις όλων των ημερών του μήνα, σε όλες τις πόλεις στης Βόρειας περιοχής. Προφανώς, αυτός ο αριθμός έχει προκύψει από αναλυτικότερα δεδομένα τα οποία μπορεί και να υπάρχουν σε άλλον υπερκύβο. Οι διαστάσεις του υπερκύβου λειτουργούν ως δείκτες στα δεδομένα του. Σε κάθε ερώτηση μπορούμε να έχουμε πρόσβαση στις τιμές των δεδομένων του κύβου με τη χρήση τιμών των διαστάσεων. Με τη χρήση τιμών για όλες τις διαστάσεις παίρνουμε μία απλή τιμή από τον υπερκύβο. Μπορούμε να χρησιμοποιήσουμε τιμές για μερικές από τις διαστάσεις του υπερκύβου. Αν κάνουμε μια ερώτηση χρησιμοποιώντας τιμές έστω από δύο διαστάσεις του κύβου, τότε θα πάρουμε συγκεντρωτική πληροφορία για αυτές τις δύο διαστάσεις. Αν στο παράδειγμα του Σχήματος 8.5 εκτελέσουμε μία αναζήτηση στα δεδομένα του κύβου με βάση μόνο συγκεκριμένες τιμές για δυο διαστάσεις (PRODUCT = "Cola" και MONTH = "JAN") τότε θα πάρουμε το άθροισμα των πωλήσεων του προϊόντος "Cola" το μήνα Ιανουάριο σε όλες τις περιοχές. Αν εκτελέσουμε μία αναζήτηση στα δεδομένα του κύβου για την τιμή PRODUCT = "Cola" θα πάρουμε τις συνολικές πωλήσεις του συγκεκριμένου προϊόντος. Επίσης, μας δίνεται η δυνατότητα αναζητήσεων χωρίς να δίνουμε συγκεκριμένες τιμές στις διαστάσεις ή να δίνουμε περιοχή τιμών. Μπορούμε, για παράδειγμα, να πάρουμε τις συγκεντρωτικές πωλήσεις όλων των προϊόντων ανά προϊόν και μήνα για τους μήνες Ιανουάριος έως Μάρτιο.

8.6.2 Πράξεις στους υπερκύβους

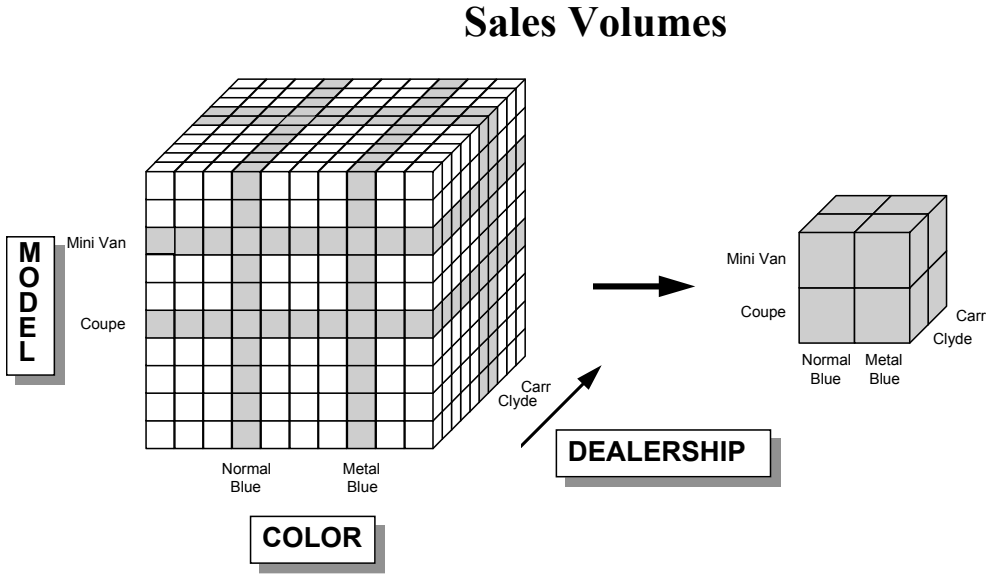
Οι υπερκύβοι μας δίνουν τη δυνατότητα πλοήγησης στις ιεραρχίες των διαστάσεών τους. Η πλοήγηση είναι δυνατή από τις πράξεις οι οποίες μας παρέχονται. Οι πράξεις που συνήθως γίνονται στους υπερκύβους είναι οι παρακάτω:

- **Roll-up**: Πρόκειται για πράξη με την οποία εκτελούμε ένα βήμα ανόδου στην ιεραρχία μιας διάστασης. Στο παράδειγμα του Σχήματος 8.6, έχουμε αρχικά ένα κύβο που αποτελείται από τρεις διαστάσεις: *Χρώμα (Color), Μοντέλο (Model)* και *Γεωγραφία (Geography)*. Η διάσταση *Γεωγραφία* έχει τρία επίπεδα: *κατάστημα (dealership)*, *περιοχή (district)* και *περιφέρεια (region)*. Μία πράξη roll-up στη διάσταση *Γεωγραφία* θα μας έδινε έναν νέο κύβο που θα περιείχε αθροιστικές πωλήσεις προϊόντων ανά *περιοχή*, *χρώμα* και *μοντέλο*. Ο κύβος που προκύπτει από την πράξη περιέχει πιο ομαδοποιημένα δεδομένα, με βάση τη διάσταση στην οποία έγινε η ομαδοποίηση. Η ανάβαση στην ιεραρχία μπορεί να συνεχιστεί με όμοιο τρόπο.
- **Drill-down**: Είναι η αντίστροφη πράξη του roll-up, όπου πάμε από ένα υψηλότερο επίπεδο ιεραρχίας μίας διάστασης σε ένα χαμηλότερο. Στον πίνακα του Σχήματος 8.6, μία πράξη drill-down στη διάσταση *Γεωγραφία*, από το επίπεδο περιφέρειας, στον τελευταίο κύβο, στο επίπεδο καταστήματος, θα μας έδινε τον αρχικό κύβο.



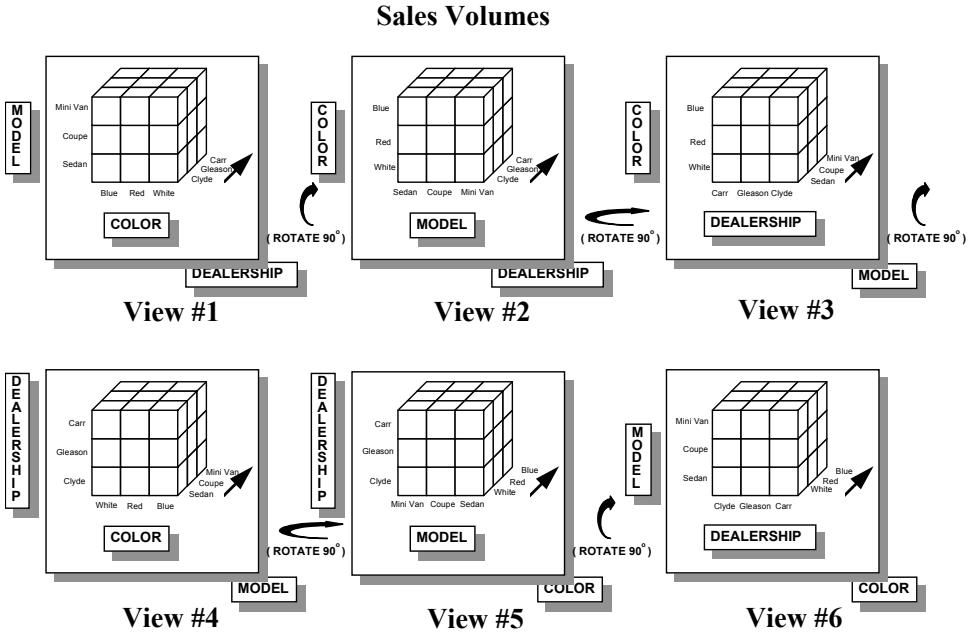
Σχήμα 8.6 Πλοήγηση σε μια ιεραρχία διαστάσεων

- **Slicing**: Πρόκειται για πράξη επιλογής δεδομένων σε μία συγκεκριμένη διάσταση. Ένα *επίπεδο (slice)* είναι ένα υποσύνολο ενός υπερκύβου σύμφωνα με μία περιοχή τιμών ή μια συγκεκριμένη τιμή ενός επιπέδου διάστασης. Στο παράδειγμα του σχήματος 8.7 κάνουμε τις εξής επιλογές:
  - στην διάσταση *Μοντέλου* κρατάμε μόνο τις τιμές SPORTS COUPE και MINI VAN.
  - στην διάσταση *Γεωγραφίας* κρατάμε μόνο τα καταστήματα CARR και CLYDE.
  - στην διάσταση *Χρώματος* κρατάμε μόνο τις τιμές METAL BLUE και NORMAL BLUE



Σχήμα 8.7 Slicing

- **Pivoting**: Πρόκειται για πράξη αλλαγής της διάταξης των διαστάσεων ώστε να διευκολυνθεί η ανάλυση. Κατά το pivoting, δεν μεταβάλλονται ούτε μειώνονται τα δεδομένα του υπερκύβου. Απλά αλλάζει ο τρόπος παρουσίασής τους στην εφαρμογή ανάλυσης. Στο σχήμα 8.8 φαίνονται οι διαφορετικοί τρόποι παρουσίασης ενός κύβου.



Σχήμα 8.8 Pivoting

8.7 ΑΝΑΦΟΡΕΣ

[BS96] J.Byard and D. Schneider. The Ins & Outs (and everything in between) of Data Warehousing. *Tutorials of ACM SIGMOD International Conference on Management of Data.* Montreal, Canada, 1996.

[CD 96] S. Chaudhuri, U.Dayal, 'Data warehousing and OLAP for Decision Support', Tutorials of *22nd International Conference on Very large Data Bases*, Mumbai, India, September 1996

[CD97] S. Chaudhuri, U. Dayal. *An Overview of Data Warehousing and OLAP Technology*. SIGMOD Record, Vol. 26, No. 1, March 1997

[Coll96] G. Colliat. *OLAP, Relational, and Multidimensional Database Systems*. SIGMOD Record, Vol. 25, No.3, September 1996.

[Inm96] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, second edition, 1996.

[Kena95] An Introduction to Multidimensional Database Technology. Kenan Technologies, 1995. Available at: http://www.kenan.com/content/compinfo/whitepapers/white.htm

[RedB97] Red Brick Systems, Inc.. *Red Brick Warehouse 5.0*. http://www.redbrick.com/rbs-g/html/whouse50.html, 1997.

[Wido 95] J. Widom, 'Research Problems in Data Warehousing.' *Proceedings of the 4th Int'l Conference on Information and Knowledge Management* (CIKM), November 1995.