ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΣΧΕΔΙΑΣΗ –ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΗΣ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ (DATA WAREHOUSE) ΣΤΗΝ ACCESS GIA ΤΑ ΟΙΚΟΝΟΜΙΚΑ ΣΤΟΙΧΕΙΑ ΜΙΑΣ ΕΠΙΧΕΙΡΗΣΗΣ ΩΣ ΣΥΣΤΗΜΑ ΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ (DECISION SUPPORT SYSTEM)

ΝΤΖΟΥΜΑΝΙΚΑ Ε. ΣΠΥΡΙΔΟΥΛΑ

ΣΑΛΑΠΠΑΣ Κ.ΑΛΕΞΙΟΣ

ΧΡΙΣΤΟΦΙΔΗΣ Γ. ΣΤΥΛΙΑΝΟΣ

ΕΙΣΗΓΗΤΗΣ ΠΑΠΑΣΤΕΡΓΙΟΥ ΘΩΜΑΣ

ΜΕΣΟΛΟΓΓΙ 2015

ΠΕΡΙΛΗΨΗ

Οι αποθήκες δεδομένων αποτελούν τον πυρήνα της αρχιτεκτονικής των πληροφοριακών συστημάτων των σύγχρονων επιχειρήσεων. Η ανάγκη για αποδοτική και ταχεία αξιοποίηση ενός μεγάλου όγκου δεδομένων, έδωσε ώθηση στην ανάπτυξη συστημάτων που εξειδικεύονται στην πολυδιάστατη ανάλυση της συσσωρευόμενης πληροφορίας. Κεντρικό ρόλο προς αυτή την κατεύθυνση έχουν οι εφαρμογές Online Analytical Processing στις βάσεις δε-δομένων, που επιτρέπουν τη διαχείριση της πληροφορίας σε πολυδιάστατο επίπεδο, για την υποστήριξη του σχεδιασμού και της λήψης αποφάσεων.

Στόχος της παρούσας εργασίας είναι η παρουσίαση των δυνατοτήτων πολυδιάστατης ανάλυσης, ορισμένων ευρέως χρησιμοποιούμενων συστημάτων διαχείρισης βάσεων δεδομένων.

Abstract

Data warehouses constitute the core of the information systems design of modern businesses. The need for an effective and fast development of the huge volume of data motivated the development of systems specialized in the multidimensional analysis of accumulated knowledge. Online Analytical Processing applications in databases play a central role towards this direction; they allow the management of information on a multidimensional level and support planning and decision-making.

The aim of the present thesis is the demonstration of the possibilities of multidimensional analysis of some widely used systems of managing databases.

Εισαγωγή

Το βασικό χαρακτηριστικό της εποχής της πληροφορίας είναι ο ραγδαίος ρυθμός αύξησης του είδους και του όγκου των δεδομένων που συγκεντρώνονται από οργανισμούς και εταιρείες. Η πρόκληση που καλούνται να αντιμετωπίσουν οι σχετιζόμενες τεχνολογίες είναι η βέλτιστη αξιοποίηση της διαρκώς αυξανόμενης συσσώρευσης πληροφορίας. Ο κλάδος των βάσεων δεδομένων είναι προφανώς αυτός που επηρεάζεται και εμπλέκεται πιο άμεσα από τις νέες τάσεις και αυξημένες απαιτήσεις για αποθήκευση και ανάκτηση δεδομένων. Η κλασική αντίληψη ενός συστήματος βάσεως δεδομένων στην οποία ένας τελικός χρήστης δύναται να εκτελεί ένα κάθε φορά ερώτημα για να λάβει μια συγκεκριμένη απάντηση, έχει υποχρεωτικά διευρυνθεί καθώς η νέα απαίτηση είναι ο χρήστης να είναι σε θέση να αξιοποιεί το σύνολο των δεδομένων του με σκοπό την παρακολούθηση διαχρονικών τάσεων και τον εντοπισμό συσχετίσεων μεταξύ τους. Επιπλέον έχει διευρυνθεί και ο ορισμός του χρήστη ενός συστήματος βάσεων δεδομένων, πλέον σε επαφή με τις αποθήκες δεδομένων δεν έρχονται μόνο εξειδικευμένα στελέχη επιχειρήσεων με τεχνικές γνώσεις, αλλά και ομάδες διευθυντικών στελεχών και υπαλλήλων που δεν έχουν κατά ανάγκη το ανάλογο τεχνικό υπόβαθρο.

Οι προαναφερθείσες εξελίξεις έχουν οδηγήσει στην συνεχή εξέλιξη των υπαρχόντων συστημάτων διαχείρισης βάσεων δεδομένων, αφενός σε επίπεδο βελτίωσης των επιδόσεων τους, όσο και επίπεδο απλούστευσης της προσβασιμότητας τους για χρήστες διαφορετικών αναγκών. Η σημερινή τάση είναι πρωτίστως η ανάπτυξη τεχνολογιών με κύριο στόχο την ταχεία και αποδοτική επεξεργασία μεγάλου όγκου δεδομένων, σε παραλληλία με τη δημιουργία διαδραστικά ελκυστικών εφαρμογών για την προβολή των αποτελεσμάτων και την δημιουργία εκθέσεων και αναφορών που έχουν πρακτικό ενδιαφέρον για ένα ευρύτερο φάσμα τελικών χρηστών.

Η ανάγκη για αξιοποίηση δεδομένων που προέρχονται από διαφορετικές πηγές ώστε να είναι εφικτή η ταυτόχρονη επεξεργασία τους ικανοποιήθηκε από την σημαντική πρόοδο στην τεχνολογία των αποθηκών δεδομένων (data warehouses). Παράλληλα με τις αποθήκες δεδομένων αναπτύχθηκαν και οι εφαρμογές σύγχρονης αναλυτικής επεξεργασίας δεδομένων (OLAP: Online Analytical Processing). Ο όρος OLAP χρησιμοποιείται για να περιγραφεί η ανάλυση σε πολυδιάστατο επίπεδο πολύπλοκων δεδομένων που προέρχονται από αποθήκες δεδομένων. Η ερευνητική δραστηριότητα σε θέματα σχετικά με την μοντελοποίηση και αποθήκευση πολυδιάστατων δεδομένων

και το Online Analytical Processing υπήρξε ιδιαίτερα έντονη την τελευταία δεκαετία με την παρουσίαση διαφόρων προσεγγίσεων σε ότι αφορά τη δημιουργία του πολυδιάστατου μοντέλου.

Η δομή της πτυχιακής εργασίας αποτελείται από τα εξής κεφάλαια :

Στο κεφάλαιο 1 γίνεται αναφορά για την ιστορική εξέλιξη της πληροφορικής και για την τεχνολογική των υπολογιστών. Στο κεφάλαιο 2, αναφερόμαστε στο σχεδιασμό, τη δομή και την οργάνωση των βάσεων δεδομένων. Συνεχίζοντας στο κεφάλαιο 3, αναλύουμε τις αποθήκες δεδομένων και την εξόρυξη γνώσης από τα δεδομένα. Τέλος, στο κεφάλαιο 4 αναφερόμαστε στην αναλυτική επεξεργασία δεδομένων (OLAP ανάλυση) και για την χρήση της Microsoft Access πάνω σε παραδείγματα OLAP ανάλυσης.

Πίνακας περιεχομένων

ΚΕΦΑΛΑΙΟ 1: ΕΞΕΛΙΞΗ ΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

1.1 Ιστορική εξέλιξη της πληροφορικής

Οι υπολογιστές με την σύγχρονη μορφή τους είναι κοντά μας περίπου πέντε δεκαετίες αλλά ο συνδυασμός των τεχνικών γνώσεων που προέρχονται από τις εμπλεκόμενες επιστήμες για την κατασκευή τους, πηγάζει από την απαρχή των επιστημών αυτών και είναι αποτέλεσμα αιώνων.

Η επιστήμη των υπολογιστών γεννήθηκε από την ανάγκη να πραγματοποιήσει ο άνθρωπος σύνθετους υπολογισμούς που δεν μπορούσε να συγκρατήσει με την σκέψη του. Σε αυτή του την προσπάθεια τα πρώτα εργαλεία μέτρησης που χρησιμοποίησε ήταν ο άβακας και το ινδοαραβικό αριθμητικό σύστημα.

Στις αρχές του 19^{ου} αιώνα ξεκινάει η ιστορική εξέλιξη της πληροφορικής ως επιστήμης όπως την γνωρίζουμε σήμερα. Ο Charles Babbage και η Augusta Ada King, Δούκισσα του Λοβλείς, οραματίστηκαν την κατασκευή μηχανικών υπολογιστών που θα μπορούσαν να εκτελούν υπολογισμούς αυτόματα και αξιόπιστα. Καθώς μέχρι τότε η σύνθετοι υπολογισμοί γίνονταν από ανθρώπους με την βοήθεια μαθηματικών πινάκων. Λόγο της ανθρώπινης φύσης στους μαθηματικούς πίνακες παρουσιαζόταν συχνά λάθη και τους καθιστούσαν αναξιόπιστους.

Για την υλοποίηση αυτού του οράματος χρειάστηκαν περίπου 125 χρόνια και οι σημαντικότεροι σταθμοί αυτής της πορείας είναι οι παρακάτω :

- Το 1939 ένας νεαρός Γερμανός μηχανικός , ο Konrad Zuse, ολοκλήρωσε τον Z1, τον πρώτο προγραμματιζόμενο γενικής χρήσης ψηφιακό υπολογιστή. Το 1941 ο Zuse και ένας φίλος του ζήτησαν από την Γερμανική κυβέρνηση να επιδοτήσει την κατασκευή ενός ταχύτερου ηλεκτρονικού υπολογιστή που θα βοηθούσε στην αποκωδικοποίηση των εχθρικών μηνυμάτων κατά την διάρκεια του β' παγκοσμίου πολέμου. Η ηγεσία του στρατού του ναζιστικού καθεστώτος απέρριψε την πρόταση του Zuse, επιδεικνύοντας υπέρμετρη εμπιστοσύνη στην αεροπορία τους πιστεύοντας ότι τα αεροπλάνα τους αρκούν

για να κερδίσουν τον πόλεμο, χωρίς την βοήθεια προηγμένων υπολογιστικών μηχανών.

- Περίπου τον ίδιο καιρό, η βρετανική κυβέρνηση δημιούργησε μια απόρρητη ομάδα μαθηματικών και μηχανικών, η οποία είχε σκοπό την παραβίαση των Γερμανικών στρατιωτικών κωδικών. Το 1943 αυτή η ομάδα με ηγέτη τον μαθηματικό Alan Turing, ολοκλήρωσε την κατασκευή του υπολογιστή Κολοσσός. Αυτός ο ειδικής χρήσης υπολογιστής έδωσε την δυνατότητα στις Βρετανικές μυστικές υπηρεσίες να υποκλέψουν τα πλέον απόρρητα μηνύματα των Γερμανών μέχρι το τέλος του πολέμου.
- Το 1939 το πανεπιστήμιο της πολιτείας της Αϊόβα και συγκεκριμένα, ο καθηγητής John Atanasoff και ο διπλωματούχος φοιτητής του Clifford Berry δημιούργησαν τον υπολογιστή Atanasoff- Berry Computer (ABC), ο οποίος είχε την δυνατότητα να λύνει συστήματα γραμμικών εξισώσεων. Όταν ο Atanasoff προσέγγισε την IBM για να ζητήσει χρηματοδότηση, του απάντησαν ότι « η IBM δεν θα ενδιαφερθεί ποτέ για μια ηλεκτρονική υπολογιστική μηχανή ».
- Ο καθηγητής του Χάρβαρντ Howard Aiken είχε μεγαλύτερη επιτυχία στη χρηματοδότηση του αυτόματου υπολογιστή γενικής χρήσης που ανέπτυξε. Χάρη στη χορηγία ενός εκατομμυρίου δολαρίων από την IBM, ολοκλήρωσε τον υπολογιστή Mark 1 το 1944. Αυτό το τέρας με διαστάσεις 15 μέτρα μήκος και 2,5 μέτρα ύψος μπορεί να χρησιμοποιούσε θορυβώδη ηλεκτρομαγνητικά ρελέ, ωστόσο απέδειξε την αξία του με τον αριθμό βαλλιστικών πινάκων για το ναυτικό των ΗΠΑ.
- Αφού συνεργάστηκε με τον Atanasoff και μελέτησε τον υπολογιστή ABC, ο John Mauchly ένωσε τις δυνάμεις του με τον J. Presper Eckert, προκειμένου να βοηθήσεις στις προσπάθειες των ΗΠΑ στο β' παγκόσμιο πόλεμο κατασκευάζοντας ένα μηχάνημα που θα μπορούσε να υπολογίσει βαλλιστικούς πίνακες για το ναυτικό των ΗΠΑ. Αυτό το μηχάνημα ήταν ο Ηλεκτρονικός Αριθμητικός Υπολογιστής Electronic Integrator and Computer (ENIAC). Ένας τεραστίων διαστάσεων υπολογιστής με βάρος δύο τόνων, ο οποίος χαλούσε κατά μέσω όρο κάθε επτά λεπτά. Όταν λειτουργούσε μπορούσε να εκτελέσει υπολογισμούς 500 φορές πιο γρήγορα από τους τότε υπάρχοντες ηλεκτρομαγνητικούς υπολογιστές, με την ίδια περίπου ταχύτητα των σημερινών αριθμομηχανών τσέπης. Ο ENIAC ολοκληρώθηκε δύο μήνες μετά

το τέλος του β' παγκοσμίου πολέμου το 1945, αλλά έπεισε τους δημιουργούς του ότι οι υπολογιστές μεγάλων δυνατοτήτων ήταν κάτι που μπορούσε να παραχθεί ευρέως στην αγορά. Μετά τον πόλεμο οι Mauchly και Eckert ίδρυσαν μια ιδιωτική εταιρεία και σχεδίασαν τον UNIVAC 1, τον πρώτο γενικής χρήσεως υπολογιστή της αγοράς που κατασκευάστηκε στις ΗΠΑ. Οι Mauchly και Eckert ήταν καλύτεροι μηχανικοί από επιχειρηματίες. Η κατασκευάστρια εταιρεία αριθμομηχανών Remington Rand τους εξαγόρασε το 1950, ολοκλήρωσε το UNIVAC 1 και τον παρέδωσε στην υπηρεσία απογραφής των ΗΠΑ το 1951.

1.2 Τεχνολογική Εξέλιξη των υπολογιστών

Οι πρώτοι υπολογιστές ήταν αρκετά ογκώδης , ακριβοί και με συχνά τεχνικά προβλήματα. Τους αγόραζαν μόνο μεγάλοι οργανισμοί όπως τράπεζες ή κυβερνήσεις. Απαιτούσαν μεγάλες εγκαταστάσεις για την στέγασή τους και πληθώρα εξειδικευμένου εργατικού δυναμικού για την λειτουργία τους.

Η πρώτη μεγάλη αλλαγή όσον αφορά την τεχνολογία των υπολογιστών ήταν το 1948 όταν εφευρέθηκε το τρανζίστορ ως υποκατάστατο της λυχνίας και εμφανίστηκε στους υπολογιστές οκτώ χρόνια αργότερα. Σημαντικά πλεονεκτήματα της νέας αυτής τεχνολογίας ήταν η μείωση του όγκου των μηχανημάτων καθώς και η αυξημένη αξιοπιστίας τους σε σχέση με τον προκάτοχό τους. Παράλληλα με την είσοδο των τρανζίστορ, παρατηρήθηκε κι η βελτίωση των λογισμικών, κάνοντας τους υπολογιστές πιο γρήγορους και εύκολους στη χρήση. Αποτέλεσμα αυτής της τεχνολογικής αλλαγής ήταν η ραγδαία είσοδος των υπολογιστών στις επιχειρήσεις, την επιστήμη και την μηχανική.

Η είσοδος στην εποχή του διαστήματος και τις προσπάθειες του ανθρώπου να κάνει τα πρώτα του βήματα εκτός γης με τον σχετικό αναβρασμό που επικρατούσε στις ισχυρές κυβερνήσεις της εποχής, (Σοβιετική Ένωση – ΗΠΑ), ώθησαν τους υπολογιστές σε περαιτέρω εξέλιξη.

Οι νέες ανάγκες απαιτούσαν υπολογιστικά συστήματα μικρότερα και ποιο ισχυρά. Τα ολοκληρωμένα κυκλώματα ήταν η απάντηση στις ανάγκες αυτές. Ένα ολοκληρωμένο κύκλωμα είναι ένα μικρό τσιπ σιλικόνης που περιέχει εκατοντάδες τρανζίστορ. Μέχρι τα μέσα της δεκαετίας του 60 αυτή η νέα τεχνολογία είχε παραγκωνίσει την παλιά για τους ίδιους λόγους που τα τρανζίστορ είχαν αντικαταστήσει τις λυχνίες.

Πλεονεκτήματα των ολοκληρωμένων κυκλωμάτων :

- Αξιοπιστία. Οι μηχανές που κατασκευάζονται με ολοκληρωμένα κυκλώματα είναι λιγότερο ευάλωτες σε βλάβες επειδή τα τσιπ μπορούν να ελεγχθούν πιο αυστηρά πριν την εγκατάστασή τους.
- Μέγεθος. Ένα μόνο τσιπ μπορούσε να αντικαταστήσει ολόκληρες πλακέτες με εκατοντάδες τρανζίστορ και άλλα ηλεκτρονικά στοιχεία, καθιστώντας εφικτή την κατασκευή πολύ μικρών μηχανημάτων.
- Ταχύτητα. Επειδή ο ηλεκτρισμός έπρεπε να διανύσει μικρότερες αποστάσεις, οι μικρότερες συσκευές ήταν πολύ ταχύτερες από τους προκατόχους τους.
- Αποδοτικότητα. Επειδή τα τσιπ ήταν πολύ μικρά κατανάλωναν πολύ λιγότερη ηλεκτρική ενέργεια. Το αποτέλεσμα ήταν χαμηλότερη θερμότητα.
- Κόστος. Οι τεχνικές μαζικής παραγωγής έκαναν εύκολη την κατασκευή οικονομικών τσιπ.

Η καινοτομία που είχε το μεγαλύτερο αντίκτυπο στην μέχρι τώρα εξέλιξη της πληροφορικής, στην κοινωνία, ήταν ο μικροεπεξεργαστής. Το 1971 μηχανικοί της Intel ανέπτυξαν τον πρώτο μικροεπεξεργαστή. Το κόστος έρευνας και ανάπτυξής του ήταν τεράστιο αλλά όταν εμφανίστηκαν οι γραμμές παραγωγής, τα τσιπ σιλικόνης για υπολογιστές μπορούσαν να παραχθούν μαζικά χωρίς μεγάλο κόστος. Λόγω του χαμηλού κόστους παραγωγής, κυκλοφόρησαν στην αγορά πληθώρα προϊόντα που βασίζονταν σε φθηνούς μικροεπεξεργαστές. Η περιοχή Σαν Χοσέ της Καλιφόρνια απέκτησε το όνομα Σίλικον Βάλεϋ (κοιλάδα της σιλικόνης) όταν δεκάδες εταιρείες ημιαγωγών γεννήθηκαν και ανδρώθηκαν εκεί.

Στη δεκαετία του 1970 οι υπολογιστές γίνονται προσιτοί στο ευρύ κοινό χάρη στην κατασκευή του προσωπικού υπολογιστή. Εταιρείες όπως οι Apple, Commondore, Tandy παρουσίασαν οικονομικούς υπολογιστές στο μέγεθος γραφομηχανής οι οποίοι

βασίζονται σε μικροεπεξεργαστές και ήταν το ίδιο ισχυροί με τους τεράστιους υπολογιστές μεγέθους δωματίου που υπήρχαν μέχρι τότε. Οι προσωπικοί υπολογιστές (PC) εντάσσονται στην καθημερινότητα του ανθρώπου και οι μικροεπεξεργαστές βρίσκουν εφαρμογή σε ένα μεγάλο εύρος ηλεκτρονικών συσκευών.

ΚΕΦΑΛΑΙΟ 2: ΣΧΕΔΙΑΣΜΟΣ ΔΟΜΗ ΚΑΙ ΟΡΓΑΝΩΣΗ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

2.1 Τι είναι σύστημα βάσης δεδομένων

Σύστημα βάσης δεδομένων είναι ένα ηλεκτρονικό σύστημα τήρησης εγγράφων. Είναι ένα σύστημα με υπολογιστές που ο γενικός σκοπός του είναι να τηρεί πληροφορίες και να δίνει αυτές τις πληροφορίες όταν του ζητούνται. Οι πληροφορίες που τηρούνται σε ένα τέτοιο σύστημα μπορεί να είναι οτιδήποτε έχει σημασία για το άτομο ή τον οργανισμό για την υποβοήθηση των εργασιών του ατόμου ή του οργανισμού.

2.1.1. Δεδομένα

Τα συστήματα βάσεως δεδομένων αναλόγως το πώς διαχειρίζονται τα δεδομένα που εμπεριέχονται σε αυτά χωρίζονται σε δύο μεγάλες κατηγορίες : α) τα συστήματα ενός χρήστη β) τα συστήματα πολλών χρηστών. Στα συστήματα ενός χρήστη μόνο ένας χρήστης έχει πρόσβαση μια δεδομένη στιγμή, ενώ αντίθετα στα συστήματα πολλών χρηστών (multiuser system) πολλοί χρήστες έχουν πρόσβαση στα δεδομένα ταυτόχρονα.

Τα δεδομένα που εμπεριέχονται σε μια βάση δεδομένων χαρακτηρίζονται ως **ενοποιημένα** (integrated) και **μεριζόμενα** (shared).

Με τον όρο **ενοποίηση δεδομένων** (data integration) εννοούμε ότι η βάση δεδομένων μπορεί να θεωρείτε η συνένωση πολλών αρχείων δεδομένων, που από κάθε άλλη άποψη είναι ξεχωριστά το ένα με το άλλο, ενώ κάθε πλεονασμός εξαιτίας της επανάληψης δεδομένων μεταξύ αυτών των αρχείων έχει εξαλειφθεί εντελώς ή κατά ένα μέρος.

Με τον όρο μερισμός δεδομένων (data sharing) εννοούμε ότι τα μεμονωμένα στοιχεία δεδομένων της βάσης δεδομένων μπορούν να μοιράζονται πολλοί διαφορετικοί

χρήστες και ο καθένας από αυτούς τους χρήστες να μπορεί να έχει πρόσβαση στο ίδιο στοιχείο δεδομένων. Η ταυτόχρονη προσπέλαση δηλαδή διάφοροι χρήστες να μπορούν να έχουν πρόσβαση στο ίδιο στοιχείο δεδομένων την ίδια στιγμή, είναι συνέπια της ενοποίησης και του μερισμού των δεδομένων.

2.1.2 Υλικό

Για την δημιουργία μιας βάσης δεδομένων χρειάζεται κάποιο συγκεκριμένο μηχανολογικό υλικό (hardware) που αποτελείτε από τα παρακάτω κύρια μέρη:

- Τα μέσα δευτερεύουσας αποθήκευσης. Στα οποία περιλαμβάνονται μαγνητικοί δίσκοι, συσκευές εισόδου – εξόδου και ελεγκτές συσκευών.
- Ο επεξεργαστής ή οι επεξεργαστές και η κύρια μνήμη που χρησιμοποιούνται για την εκτέλεση του λογισμικού μιας βάσης δεδομένων.

2.1.3 Λογισμικό

Λογισμικό είναι το μέσο, το οποίο προσδίδει διαδραστικότητα μεταξύ των χρηστών μιας βάσης δεδομένων και της φυσικής βάσης δεδομένων. Είναι το σύνολο των προγραμμάτων εκείνων που δίνουν την δυνατότητα στο χρήστη ενός πληροφοριακού συστήματος να επεξεργαστεί τα δεδομένα σε ένα σύστημα βάσεων δεδομένων. Ο διαχειριστής βάσεων δεδομένων (database manager) ή όπως είναι ευρύτερα γνωστό, το σύστημα διαχείρισης βάσεων δεδομένων (database management system, DBMS), διαχειρίζεται όλες τις απαιτήσεις των χρηστών για προσπέλαση της βάσης δεδομένων. Λειτουργίες όπως προσθήκη και αφαίρεση αρχείων ή πινάκων, ανάκληση και ενημέρωση δεδομένων που αποθηκεύονται σε αρχεία ή πίνακες είναι όλες υπηρεσίες που παρέχονται από το σύστημα διαχείρισης βάσεων δεδομένων DBMS. Επίσης μια γενικότερη υπηρεσία που παρέχεται από ένα DBMS είναι η απομόνωση των χρηστών της βάσης δεδομένων από τις λεπτομέρειες που αφορούν το υλικό. Το DBMS παρέχει στους χρήστες μια άποψη της βάσης δεδομένων ανυψωμένη πάνω το επίπεδο του υλικού υποστηρίζοντας τις πράξεις των τελικών χρηστών.

Σε ένα σύστημα βάσεων δεδομένων το DBMS είναι το σημαντικότερο στοιχείο λογισμικού ολόκληρου του συστήματος. Επίσης συναντάμε και άλλα στοιχεία λογισμικού όπως βοηθητικά προγράμματα, εργαλεία ανάπτυξης εφαρμογών, σχεδιαστικά βοηθήματα και εργαλεία σύνταξης αναφορών.

2.1.4 Χρήστες

Οι χρήστες των συστημάτων βάσεων δεδομένων χωρίζονται σε τρείς γενικές κατηγορίες.

Στην πρώτη κατηγορία ανήκουν οι προγραμματιστές εφαρμογών (application programmers). Είναι οι χρήστες που είναι υπεύθυνοι για τι γράψιμο προγραμμάτων και εφαρμογών που χρησιμοποιούν τη βάση δεδομένων. Αυτό γίνεται μέσο μιας γλώσσας προγραμματισμού όπως η C ή Pascal. Μέσω αυτών των προγραμμάτων μπορούν να ανακληθούν, να προστεθούν, να διαγραφούν και να αλλάξουν πληροφορίες. Τα προγράμματα μπορεί να είναι εφαρμογές ομαδικής επεξεργασίας (batch applications) ή εφαρμογές άμεσης επεξεργασίας (online application) και υποστηρίζουν τον τελικό χρήστη που προσπελάζει τη βάση δεδομένων από κάποιο σταθμό εργασίας ή τερματικό.

Στην δεύτερη κατηγορία χρηστών συναντάμε τους τελικούς χρήστες. Οι τελικοί χρήστες αλληλεπιδρούν με το σύστημα μέσω συνδεδεμένων σταθμών εργασίας ή τερματικών. Η προσπέλαση μιας βάσης δεδομένων από ένα τελικό χρήστη μπορεί να γίνει από τις εφαρμογές άμεσης επεξεργασίας ή χρησιμοποιώντας κάποια σύνδεση που είναι οργανικό μέρος του λογισμικού του συστήματος βάσης δεδομένων. Όσο αφορά τις διασυνδέσεις ενός συστήματος βάσεων δεδομένων, είναι ενσωματωμένες (build in) και όχι γραμμένες από τον τελικό χρήστη. Τα περισσότερα συστήματα διαθέτουν ενσωματωμένες εφαρμογές δηλαδή έναν επεξεργαστή γλώσσας ερωτημάτων (interactive query language processor). Με την βοήθεια των ενσωματωμένων εφαρμογών ο τελικός χρήστης έχει την δυνατότητα να δίνει σε ένα DBMS εντολές υψηλού επιπέδου. Επίσης τα περισσότερα συστήματα παρέχουν στους χρήστες τους πρόσθετες ενσωματωμένες διασυνδέσεις που δουλεύουν επιλέγοντας στοιχεία από ένα μενού ή συμπληρώνοντας πλαίσια σε μια φόρμα. Οι διασυνδέσεις που οδηγούνται από μενού ή από φόρμες είναι πιο εύχρηστες για τους χρήστες και δεν απαιτούν κάποια

ιδιαίτερη εκπαίδευση ή τεχνική κατάρτιση, υστερούν σε σχέση με τις διασυνδέσεις που οδηγούνται από διαταγές καθώς είναι λιγότερο ευέλικτες.

Στην τρίτη κατηγορία χρηστών ανήκει ο υπεύθυνος διαχείρισης βάσεων δεδομένων (database administrator, DBA). Καθώς και ο υπεύθυνος διαχείρισης δεδομένων (data administrator, DA). Ο υπεύθυνος διαχείρισης δεδομένων είναι ένα ανώτερο διοικητικό στέλεχος που κατανοεί τα δεδομένα και τις ανάγκες της επιχείρησης όσο αφορά τα δεδομένα καθώς είναι συνήθως από τα πολυτιμότερα περιουσιακά στοιχεία της επιχείρησης. Ο υπεύθυνος διαχείρισης βάσεων δεδομένων είναι ο τεχνικός που έχει την ευθύνη για την υλοποίηση των αποφάσεων του υπευθύνου διαχείρισης δεδομένων. η εργασία του κατά κύριο λόγο τεχνική και χαρακτηρίζεται από υψηλή κατάρτιση. Στα πλαίσια της εργασίας του είναι η δημιουργία της βάσης δεδομένων, η υλοποίηση τεχνικών ελέγχων και η εν γένει εξασφάλιση της εύρυθμης λειτουργίας του συστήματος.

2.2 Δομή συστημάτων βάσεων δεδομένων

Ένα σύστημα βάσης δεδομένων χωρίζεται σε λειτουργικές μονάδες που αναλαμβάνουν τις ευθύνες του συστήματος. Τα λειτουργικά συστατικά ενός συστήματος βάσης δεδομένων διαιρούνται στα συστατικά του διαχειριστή αποθήκευσης και στα συστατικά του επεξεργαστή ερωτημάτων.

2.2.1 Διαχειριστής αποθήκευσης

Οι βάσεις δεδομένων απαιτούν μεγάλη ποσότητα χώρου αποθήκευσης λόγω του τεράστιου όγκου δεδομένων που καλούνται να διαχειριστούν. Οι ανάγκες τους για αποθηκευτικό χώρο αναδεικνύουν την σημαντικότητα του διαχειριστή αποθήκευσης. Η κύρια μνήμη των υπολογιστών δεν μπορεί να αποθηκεύει μεγάλο όγκο πληροφοριών, έτσι οι πληροφορίες αποθηκεύονται σε σκληρούς δίσκους και τα δεδομένα μετακινούνται μεταξύ δίσκων και κύριας μνήμης. Η μετακίνηση δεδομένων από ένα δίσκο είναι αργή σε σχέση με την ταχύτητα της μονάδας επεξεργασίας έτσι

προκύπτει η ανάγκη η δομή του συστήματος να δομεί τα δεδομένα ώστε να ελαχιστοποιεί την μετακίνηση δεδομένων μεταξύ δίσκου και κύριας μνήμης. Ο διαχειριστής αποθήκευσης είναι μια λειτουργική μονάδα προγράμματος η οποία διασυνδέει τα δεδομένα που έχουν αποθηκευτεί στην βάση δεδομένων και των ερωτημάτων που στέλνονται στο σύστημα. Τα δεδομένα αποθηκεύονται στο δίσκο μέσω ενός συμβατικού λειτουργικού συστήματος. Ο διαχειριστής αποθήκευσης μεταφράζει τις εντολές μίας γλώσσας χειρισμού δεδομένων (DML) σε χαμηλού επιπέδου εντολές του συστήματος αρχείων.

Γλώσσα χειρισμού δεδομένων (data- manipulation language- DML) είναι η γλώσσα που επιτρέπει στους χρήστες να έχουν πρόσβαση ή να χειρίζονται δεδομένα όπως είναι οργανωμένα από το κατάλληλο μοντέλο δεδομένων.

Ο διαχειριστής αποθήκευσης, αποθηκεύει, ανακαλεί και ενημερώνει τα δεδομένα μιας βάσης δεδομένων και αποτελείται από τις παρακάτω συστατικές λειτουργίες :

- Τον διαχειριστή ελέγχου ταυτότητας και ακεραιότητας που ελέγχει την ακεραιότητα των δεδομένων καθώς και την προσβασιμότητα των χρηστών μιας βάσης δεδομένων σε αυτά.
- Τον διαχειριστή συναλλαγών που διασφαλίζει την ορθή λειτουργία μιας βάσης δεδομένων ανεξάρτητα από τις αποτυχίες του συστήματος και οι ταυτόχρονες εκτελέσεις προχωρούν χωρίς διενέξεις.
- Τον διαχειριστή αρχείων όπου διαχειρίζεται τον δεσμευμένο χώρο στο δίσκο και τις δομές δεδομένων.
- Τον διαχειριστή buffer όπου διαχειρίζεται την τροφοδοσία των δεδομένων από τον δίσκο στην κύρια μνήμη και αποφασίζει ποια δεδομένα θα εισέλθουν στην κύρια μνήμη. Η σημαντικότητά του έγκειται στον χειρισμό μεγεθών δεδομένων που είναι πολύ μεγαλύτερα από την κύρια μνήμη.

Μέσο των λειτουργιών ενός συστήματος βάσης δεδομένων ο διαχειριστής αποθήκευσης καλείται να χειριστεί διάφορες δομές δεδομένων όπως αρχεία δεδομένων που αποθηκεύονται στην ίδια βάση δεδομένων, λεξικό δεδομένων και ευρετήρια που παρέχουν γρήγορη πρόσβαση σε στοιχεία δεδομένων.

2.2.2 Επεξεργαστής ερωτημάτων

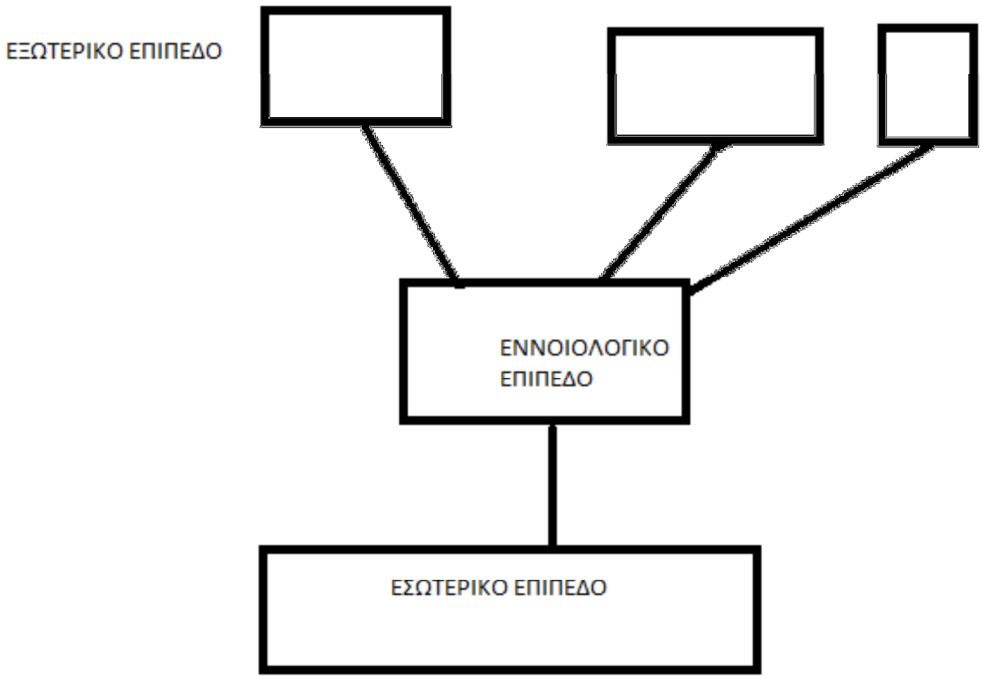
Ο επεξεργαστής ερωτημάτων σε ένα σύστημα βάσης δεδομένων απλοποιεί και διευκολύνει την πρόσβαση στα δεδομένα. Με την βοήθεια των προβολών υψηλού επιπέδου οι χρήστες του συστήματος δεν χρειάζεται να κατανοήσουν τις λεπτομέρειες του χειρισμού του συστήματος για να αλληλεπιδράσουν με αυτό. Ο επεξεργαστής ερωτημάτων μεταφράζει τις ενημερώσεις και τα ερωτήματα που έχουν γραφτεί σε μία μη διαδικαστική γλώσσα σε μια αποτελεσματική σειρά από λειτουργίες σε φυσικό επίπεδο και αποτελείται από τα παρακάτω στοιχεία:

- Τον DDL μεταφραστή ο οποίος μεταφράζει τις DDL εντολές και εγγραφές στο λεξικό δεδομένων. DDL ονομάζουμε τη γλώσσα ορισμού των δεδομένων (data – definition language) με την οποία καθορίζεται η διάταξη της βάσης δεδομένων
- Τον DML μεταγλωττιστή, που μεταφράζει τις εντολές μιας γλώσσας ερωτημάτων σε ένα πλάνο από εντολές χαμηλού επιπέδου, που καταλαβαίνει η μηχανή υπολογισμού των ερωτημάτων.
- Την μηχανή υπολογισμού ερωτημάτων, που εκτελεί εντολές χαμηλού επιπέδου που δημιουργούνται από τον μεταγλωττιστή DML

2. 3 Αρχιτεκτονική συστημάτων βάσεων δεδομένων

Η αρχιτεκτονική ενός συστήματος βάσεων δεδομένων χωρίζεται σε τρία επίπεδα:

- Το εσωτερικό επίπεδο (internal level). Που είναι αυτό που αφορά τον τρόπο φυσικής αποθήκευσης των δεδομένων.
- Το εξωτερικό επίπεδο (external level). Πού είναι αυτό που αφορά τον τρόπο που βλέπουν οι μεμονωμένοι χρήστες τα δεδομένα.
- Το εννοιολογικό επίπεδο (conceptual level). Που είναι το επίπεδο που εξασφαλίζει την σύνδεση των δύο προηγούμενων επιπέδων.



Σχήμα 1. Αρχιτεκτονική Συστημάτων Βάσεων Δεδομένων

2. 3. 1 Το εξωτερικό επίπεδο

Το εξωτερικό επίπεδο είναι το επίπεδο που διαδρά με το σύστημα ένας μεμονωμένος χρήστης. Μεμονωμένος χρήστης μπορεί να είναι ο οποιοσδήποτε τελικός χρήστης ανεξάρτητα από τον βαθμό εξειδίκευσής του και παρεμβατικότητας του σε ένα πληροφοριακό σύστημα βάσεων δεδομένων. Σε αυτό το επίπεδο μπορούμε να συναντήσουμε από έναν απλό χρήστη καθώς και έναν υπεύθυνο διαχείρισης βάσεων δεδομένων.

Κάθε χρήστης του εξωτερικού επιπέδου ενός πληροφοριακού συστήματος χρησιμοποιεί κάποιες γλώσσες αναλόγως τη φύση της εργασίας του. Ένας διαχειριστής εφαρμογών συνήθως χρησιμοποιεί γλώσσες προγραμματισμού όπως C, Cobol, ή pl/i, καθώς και αποκλειστικές γλώσσες προγραμματισμού που εξειδικεύονται στο συγκεκριμένο σύστημα. Σε αντίθεση με ένα τελικό χρήστη που συνήθως

χρησιμοποιεί είτε μία γλώσσα ερωτημάτων (query language) είτε κάποια γλώσσα ειδικής χρήσης οδηγημένη από φόρμες ή μενού προσαρμοσμένα στις απαιτήσεις του συγκεκριμένου χρήστη.

Ένας μεμονωμένος χρήστης έχει μια εξωτερική άποψη σε ένα σύστημα βάσεων δεδομένων. Αυτό προκύπτει γιατί ο μεμονωμένος χρήστης ενδιαφέρεται για ένα συγκεκριμένο περιεχόμενο και όχι για το σύνολο των δεδομένων ενός πληροφοριακού συστήματος.

2. 3. 2 Το εννοιολογικό επίπεδο

Το εννοιολογικό επίπεδο είναι μια αναπαράσταση ολόκληρου του πληροφοριακού περιεχομένου της βάσης δεδομένων με κάπως πιο αφηρημένη μορφή σε σχέση με το πώς αποθηκεύονται πραγματικά τα δεδομένα. Η εννοιολογική άποψη αποτελείται από εννοιολογικές εγγραφές διαφόρων τύπων και ορίζεται από το εννοιολογικό σχήμα που περιλαμβάνει τους ορισμούς των διαφόρων τύπων εννοιολογικών εγγραφών.

2. 3.3Το εσωτερικό επίπεδο

Το εσωτερικό επίπεδο είναι μια αναπαράσταση χαμηλού επιπέδου για ολόκληρη τη βάση δεδομένων και αποτελείται από διάφορους τύπους εσωτερικών εγγραφών. Η εσωτερική άποψη περιγράφεται με το εσωτερικό σχήμα που ορίζει τους διάφορους τύπους εγγραφών καθώς και ποια ευρετήρια υπάρχουν και πως αναπαρίστανται τα αποθηκευμένα πεδία.

2.4 Πλεονεκτήματα και Μειονεκτήματα των Συστημάτων Διαχείρισης Βάσεων

Τα πλεονεκτήματα των συστημάτων διαχείρισης βάσεων δεδομένων είναι :

- Έλεγχος του πλεονασμού δεδομένων : Όπως αναφέρθηκε παραπάνω τα παραδοσιακά συστήματα αρχείων σπαταλούσαν αρκετό χώρο με το να αποθηκεύουν τα ίδια δεδομένα σε περισσότερα από ένα αρχεία. Αντιθέτως , τα συστήματα βάσεων προσπαθούν να εξαλείψουν τον πλεονασμό τελείως ενσωματώνοντας τα αρχεία έτσι ώστε να μην υπάρχουν πολλά αντίγραφα των ίδιων δεδομένων. Παρ' όλα αυτά οι βάσεις δεδομένων δεν εξαφανίζουν τελείως τον πλεονασμό των δεδομένων, αφού σε πολλές περιπτώσεις χρειάζεται να έχουμε επανάληψη των ίδιων δεδομένων όπως για παράδειγμα στην υλοποίηση σύνθετων σχέσεων (relationships) ανάμεσα στα στοιχεία της βάσης.
- Συνεκτικότητα των δεδομένων: Με την εξαφάνιση ή τον έλεγχο του πλεονασμού των δεδομένων ελαττώνουμε τον κίνδυνο εμφάνισης μη συνεκτικών δεδομένων. Εάν τα δεδομένα είναι αποθηκευμένα μονάχα μία φορά στη βάση, οποιαδήποτε ενημέρωση στις τιμές τους εκτελείτε μία φορά και η νέα τιμή είναι κατευθείαν διαθέσιμη σε όλους τους τελικούς χρήστες. Εάν πάλι τα ίδια δεδομένα είναι αποθηκευμένα περισσότερες από μία φορές στη βάση και το σύστημα διαχείρισης είναι ενήμερο, μπορεί να εγγυηθεί ότι όλα τα αντίγραφα θα κρατηθούν ενήμερα. Δυστυχώς όμως μέχρι και σήμερα δεν μπορούν όλα τα υπάρχοντα στο εμπόριο συστήματα διαχείρισης βάσεων να εγγυηθούν αυτή τη συνεκτικότητα των δεδομένων.
- Επιπλέον πληροφορίες από τα ίδια δεδομένα: Μέσω της ενσωμάτωσης των δεδομένων καθίσταται για έναν οργανισμό να αντλήσει από τα δεδομένα της βάσης επιπλέον πληροφορίες, είτε μέσω συναρτήσεων στατιστικών του συστήματος διαχείρισης της βάσης, είτε μέσω της συνένωσης πινάκων.
- Κοινοποίηση δεδομένων: Τυπικά, τα αρχεία ανήκουν σε όλους τους εξουσιοδοτημένους χρήστες και έτσι οι περισσότεροι χρήστες μπορούν να μοιραστούν τα δεδομένα. Επιπλέον οι εφαρμογές μπορούν να επεκτείνουν τα υπάρχοντα δεδομένα προσθέτοντας απλά τα νέα δεδομένα στη βάση, χωρίς να χρειάζεται να ορίσουν ξανά όλα τα δεδομένα. Οι εφαρμογές επίσης μπορούν

να βασίζονται στις συναρτήσεις του συστήματος διαχείρισης χωρίς να χρειάζεται να έχουν τις δικές τους συναρτήσεις.

- Βελτιωμένη ακεραιότητα δεδομένων: Η ακεραιότητα εκφράζει συνήθως τους διάφορους περιορισμούς, οι οποίοι είναι στην ουσία κανόνες, τους οποίους η βάση δεν πρέπει να παραβαίνει. Οι περιορισμοί αυτοί μπορεί να εφαρμόζονται στα δεδομένα ενός πεδίου (γνωρίσματος), ενός πίνακα, ή μπορεί να εφαρμόζονται και στα σχέσεις μεταξύ των πινάκων. Για παράδειγμα, στο πεδίο (γνώρισμα) μιας email διευθύνσεως θα θέλαμε να υπάρχει το σύμβολο @υποχρεωτικά.
- Βελτιωμένη ασφάλεια: Η ασφάλεια μιας βάσης δεδομένων αποτελεί την προστασία της απέναντι σε μη εξουσιοδοτημένους χρήστες. Χωρίς τα απαραίτητα μέτρα η συνένωση ων αρχείων κάνει τα δεδομένα ακόμα πιο επιρρεπή και ευάλωτα σε σχέση με τα συστήματα αρχείων. Έτσι τα συστήματα διαχείρισης βάσεων επιτρέπουν στον administrator να ορίσει και να επιβάλλει την ασφάλεια της βάσης. Αυτό μπορεί να γίνει με τη μορφή ονόματος χρήστη και κωδικού έτσι ώστε να ορισθούν οι εξουσιοδοτημένοι χρήστες. Επιπλέον ορίζονται και τα δικαιώματα που μπορεί να έχει ένας χρήστης στους διάφορους πίνακες της βάσης.
- Βελτιωμένη διαθεσιμότητα και απόκριση: Σαν αποτέλεσμα της ενσωμάτωσης των αρχείων τα δεδομένα είναι απευθείας προσβάσιμα από τον τελικό χρήστη. Τα περισσότερα συστήματα βάσεων παρέχουν στον τελικό χρήστη γλώσσες υποβολής ερωτήσεων στη βάση, έτσι ώστε ο κάθε χρήστης να μπορεί να λάβει τα στοιχεία που αυτός θέλει, χωρίς να είναι απαραίτητη η παρουσία κάποιου προγραμματιστή ο οποίος θα γράψει κάποια εφαρμογή για την εξαγωγή στοιχείων από τη βάση.
- Αυξημένη παραγωγικότητα: Όπως αναφέρθηκα και πριν τα διάφορα συστήματα διαχείρισης παρέχουν έτοιμες συναρτήσεις στους προγραμματιστές εφαρμογών ώστε να μην χρειάζεται να ανησυχούν για πολύ χαμηλού επιπέδου λεπτομέρειες. Αυτό έχει ως αποτέλεσμα την ανάπτυξη της παραγωγικότητας των προγραμματιστών και την μείωση του χρόνου ανάπτυξης των διαφόρων εφαρμογών με τελικό αποτέλεσμα και την μείωση του κόστους μιας τέτοιας εφαρμογής.
- Βελτιωμένη συντήρηση: Στα παλαιότερα συστήματα αρχείων η περιγραφή των δεδομένων ήταν ενσωματωμένη μέσα σε κάθε εφαρμογή, κάνοντας έτσι την

κάθε εφαρμογή να εξαρτάται από τα δεδομένα. Μια οποιαδήποτε αλλαγή στη δομή των δεδομένων απαιτούσε και την ανάλογη αλλαγή και στα προγράμματα εφαρμογών που επηρεάζονταν από αυτήν. Αντίθετα στα συστήματα διαχείρισης απομονώνεται η περιγραφή των δεδομένων από τις εφαρμογές με αποτέλεσμα αυτές να μένουν απρόσβλητες από οποιαδήποτε αλλαγή.

- Αυξημένος συγχρονισμός: Σε πολλά από τα παλιά συστήματα αρχείων όταν δύο ή περισσότεροι χρήστες προσπαθούσαν να έχουν πρόσβαση στο ίδιο αρχείο συγχρόνως ήταν πιθανό οι προσβάσεις αυτές να ανακατεύονταν με αποτέλεσμα την απώλεια των πληροφοριών ή ακόμα και την απώλεια της ακεραιότητας. Τα σημερινά συστήματα διαχείρισης όμως εξασφαλίζουν ότι κάτι τέτοιο δεν θα συμβεί.

Τα μειονεκτήματα των συστημάτων διαχείρισης βάσεων δεδομένων είναι:

- Πολυπλοκότητα: Η παροχή όλων των λειτουργιών που απαιτούσε ένα καλό σύστημα διαχείρισης γίνεται από ένα πολύ σύνθετο πρόγραμμα. Οι σχεδιαστές, οι προγραμματιστές, οι διαχειριστές, ακόμα και οι τελικοί χρήστες θα πρέπει να αντιληφθούν τις λειτουργίες του συστήματος διαχείρισης για να μπορέσουν να το εκμεταλλευθούν. Αποτυχία στο να μπορέσουν να αντιληφθούν τις λειτουργίες του συστήματος διαχείρισης θα μπορούσε να οδηγήσει σε λανθασμένες αποφάσεις σχεδίασης με πολλαπλές συνέπειες.
- Μέγεθος: Η πολυπλοκότητα και το εύρος των λειτουργιών του συστήματος διαχείρισης το κάνουν ένα πολύ μεγάλο πρόγραμμα με αρκετές απαιτήσεις σε αποθηκευτικό χώρο και μνήμης για να τρέξει ικανοποιητικά.
- Επιδόσεις συστήματος: Τυπικά ένα παλιό σύστημα αρχείων είναι γραμμένο για μία συγκεκριμένη εφαρμογή με αποτέλεσμα να έχει καλές αποδόσεις. Αντιθέτως ένα σύστημα διαχείρισης είναι γραμμένο πιο γενικά με σκοπό να καλύπτει τις ανάγκες πολλών εφαρμογών και όχι μίας μονάχα. Αυτό έχει ως αποτέλεσμα οι εφαρμογές να μην τρέχουν τόσο γρήγορα όπως θα έτρεχαν σε ένα σύστημα αρχείων.
- Μεγαλύτερες επιπτώσεις σε αποτυχία: Η συγκέντρωση όλων των πόρων έχει ως αποτέλεσμα να γίνεται το σύστημα πιο ευάλωτο. Από τη στιγμή που όλοι οι χρήστες και οι εφαρμογές βασίζονται στη διαθεσιμότητα του συστήματος η αποτυχία οποιουδήποτε μέρους μπορεί να οδηγήσει το σύστημα σε προσωρινή παύση.

ΚΕΦΑΛΑΙΟ 3: ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

Η σύγχρονη εποχή συχνά αναφέρεται ως εποχή της πληροφορίας λόγω του μεγάλου όγκου των δεδομένων που παράγονται και κατ' επέκταση αποθηκεύονται χάρη στη γρήγορη και φθηνή τεχνολογία αποθήκευσης. Η έλευση των υπολογιστών και των μέσων μαζικής αποθήκευσης ψηφιακής πληροφορίας κατέστησε δυνατή τη συλλογή και αποθήκευση κάθε είδους δεδομένων όχι όμως και εύκολη τη διαχείριση των δεδομένων αυτών. Το πρόβλημα της διαχείρισης της πληθώρας δεδομένων που προέρχονται από ετερογενείς πηγές αντιμετωπίστηκε με την δημιουργία δομημένων βάσεων δεδομένων καθώς και συστημάτων διαχείρισης βάσεων δεδομένων (DBMS). Η παραγωγή ισχυρών συστημάτων διαχείρισης βάσεων δεδομένων βοήθησε σημαντικά στην ανάπτυξη πληροφοριακών συστημάτων που καλύπτουν τις λειτουργικές ανάγκες οργανισμών και επιχειρήσεων. Ο πυρήνας κάθε πληροφοριακού συστήματος είναι η βάση δεδομένων του και η σωστή σχεδίαση, ανάπτυξη και λειτουργία της βάσης εξασφαλίζει την επιτυχία του πληροφοριακού συστήματος. Τα πληροφορικά συστήματα διακρίνονται σε συστήματα μεγάλου αριθμού δοσοληψιών των δεδομένων ενός οργανισμού (on-line transaction processing – OLTP) και σε συστήματα στήριξης αποφάσεων (Decision Support Systems- DSS) που βοηθούν τα στελέχη των οργανισμών στη λήψη αποφάσεων.

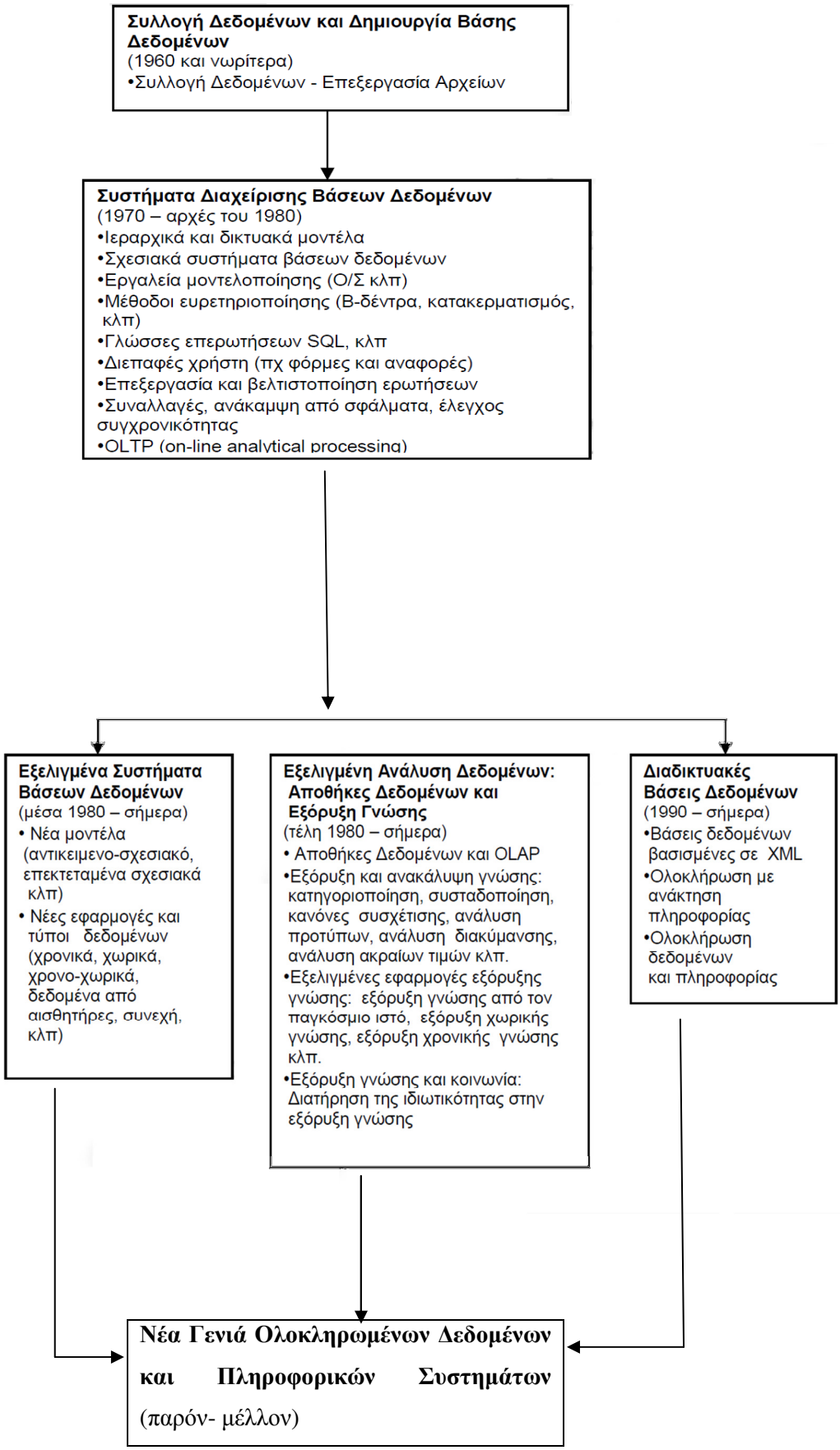
Μια από τις βασικές απαιτήσεις των συστημάτων στήριξης αποφάσεων είναι η αποδοτική πρόσβαση στα δεδομένα. Ωστόσο, αυτό δεν είναι πάντοτε εφικτό καθώς οι βάσεις δεδομένων έχουν μεγάλο υπολογιστικό φορτίο και έχουν σχεδιαστεί για την εκτέλεση συγκεκριμένων λειτουργιών. Επίσης, πολλές φορές οι βάσεις έχουν υλοποιηθεί με την χρήση τεχνολογίας που θεωρείται πλέον παρωχημένη (π.χ. αρχεία COBOL) και επομένως εφαρμογές χρησιμοποιούν μοντέρνα τεχνολογία δεν μπορούν να διαχειριστούν πληροφορία που προέρχεται από βάση δεδομένων παλαιάς τεχνολογίας. Τέλος, κάθε σύστημα στήριξης αποφάσεων εκτελεί μεγάλο αριθμό ερωτήσεων με αποτέλεσμα να δεσμεύει αρκετούς από του πόρους του συστήματος διαχείρισης της βάσης δεδομένων. αυτό έχει ως συνέπεια την μείωση της απόδοσης του συστήματος το οποίο αρχικά σχεδιάστηκε για να λειτουργεί συνεχώς και η βάση του να εξυπηρετεί μεγάλο όγκο δοσοληψιών. Επομένως, γίνεται σαφές ότι είναι

εξαιρετικά δύσκολη ή και πρακτικά αδύνατη η χρήση βάσεων δεδομένων πληροφορικών συστημάτων από συστήματα στήριξης αποφάσεων. Η λύση στο πρόβλημα δόθηκε με την ανάπτυξη ΄΄ Αποθηκών Δεδομένων ΄΄.

΄΄ 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 ΄΄. Μια Αποθήκη Δεδομένων αντλεί δεδομένα από βάσεις δεδομένων πληροφοριακών συστημάτων αλλά και από άλλες πηγές δεδομένων όπως αρχεία και δεδομένα που προέρχονται από εξωτερικές πηγές. Στη συνέχεια τα δεδομένα αυτά οργανώνονται στην Αποθήκη με κατάλληλες δομές που ανταποκρίνονται στις απαιτήσεις των αναλυτών – χρηστών των συστημάτων στήριξης αποφάσεων και παρέχουν αποδοτική πρόσβαση στα δεδομένα, χωρίς την παρουσία των προαναφερθέντων προβλημάτων. Οι Αποθήκες Δεδομένων παρέχουν τη δυνατότητα για συνεχή Αναλυτική Επεξεργασία (On- line Analytical Processing- OLAP) συγκεντρωμένης ιστορικής πληροφορίας χρήσιμης για την υποστήριξη αποφάσεων και τις εφαρμογές στρατηγικού σχεδιασμού.

Ωστόσο, η εύκολη ανάκτηση της πληροφορίας που εξασφαλίζουν οι Αποθήκες Δεδομένων δεν είναι αρκετή για την λήψη αποφάσεων. Οι τεράστιες συλλογές ποικίλων δεδομένων, όπως απλές αριθμητικές μετρήσεις, κείμενα αλλά και πιο σύνθετη πληροφορία όπως χωρικά δεδομένα, χρονικά δεδομένα και δεδομένα που αντλούνται από τον Παγκόσμιο Ιστό, δημιουργούν νέες ανάγκες για καλύτερες επιλογές διαχείρισης δεδομένων. Τέτοιες ανάγκες είναι : η αυτόματη σύνοψη των δεδομένων, η εξαγωγή της ουσίας της αποθηκευμένης πληροφορίας καθώς και η ανακάλυψη προτύπων από ακατέργαστα δεδομένα. Οι όροι ανακάλυψης γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases - KDD) και εξόρυξη γνώσης από δεδομένα (Data Mining- DM) συχνά αναφέρονται στην ίδια έννοια, δηλαδή στη διαδικασία ανακάλυψης χρήσιμων, συνήθως κρυμμένων προτύπων από τα δεδομένα.

Σε σχέση με την ιστορία των βάσεων δεδομένων όπως φαίνεται στο παρακάτω σχήμα, η διαδικασία ανακάλυψης γνώσης που περιλαμβάνει τον σχεδιασμό Αποθηκών Δεδομένων, τη συλλογή και την προεπεξεργασία των δεδομένων, την εξόρυξη γνώσης, την επιλογή μοντέλου ή συνδυασμού μοντέλων, την αξιολόγηση και τελικά την ενοποίηση και χρησιμοποίηση της εξαγόμενης γνώσης για την λήψη αποφάσεων, είναι πολύ καινούργια.



Σχήμα 2: Η εξέλιξη της τεχνολογίας διαχείρισης συστημάτων βάσεων δεδομένων

3.1 Αποθήκες Δεδομένων

Οι Αποθήκες Δεδομένων γενικεύουν και ενοποιούν τα δεδομένα σε πολυδιάστατο χώρο. Ουσιαστικά αποτελούν ένα σύνολο τεχνολογιών που παρέχει τη δυνατότητα στους αναλυτές ενός οργανισμού- επιχείρησης να σχεδιάσουν την πολιτική του έχοντας αποδοτική πρόσβαση στα δεδομένα του οργανισμού – επιχείρησης. Η υλοποίηση μιας Αποθήκης Δεδομένων περιλαμβάνει το σχεδιασμό μιας κεντρικής βάσης δεδομένων με σκοπό τη συγκέντρωση ετερογενών πηγών πληροφοριών σε μια τοποθεσία και παράλληλα την αποφυγή σύγκρουσης μεταξύ συστημάτων επεξεργασίας συναλλαγών (OLTP) και συστημάτων αναλυτικής επεξεργασίας δεδομένων (OLAP). Η σχεδίαση Αποθηκών Δεδομένων έχει σαν στόχο την αποδοτική απάντηση πολύπλοκών ερωτήσεων που δημιουργούνται κατά την αναλυτική επεξεργασία των δεδομένων και συντελεί στην αύξηση της αποδοτικότητας των εφαρμογών για την λήψη αποφάσεων και την χάραξη στρατηγικού σχεδιασμού. Η δημιουργία και η συντήρηση μιας Αποθήκης Δεδομένων είναι μια πολύπλοκη διαδικασία και εξαρτάται από τους στόχους που θέτει κάθε οργανισμός κατά την αναλυτική επεξεργασία των δεδομένων του. Πολλοί οργανισμοί επιδιώκουν τη δημιουργία Αποθήκης Δεδομένων στην οποία να συγκεντρώνετε η αναλυτική πληροφορία από τις δραστηριότητες του οργανισμού γεγονός που αυξάνει σημαντικά το κόστος υλοποίησης της αποθήκης. Ενίοτε, η Αποθήκη Δεδομένων ενός οργανισμού συμπληρώνεται από εξειδικευμένα θεματικά υποσύνολα – επιμέρους συλλογές δεδομένων (data marts) για περαιτέρω απόδοση των OLAP εφαρμογών, καθώς πρόκειται για πιο ευέλικτα συστήματα στη δημιουργία τους που όμως δεν παρέχουν ενιαία λύση, ενώ η μακροχρόνια χρήση τους δημιουργεί προβλήματα

Στη συνέχεια, θα μελετήσουμε τι ακριβώς είναι μια Αποθήκη Δεδομένων, την αρχιτεκτονική, το σχήμα της και τις λειτουργικές της διαδικασίες, καθώς όλο και περισσότεροι οργανισμοί υλοποιούν Αποθήκες Δεδομένων για την ανάλυση των δεδομένων τους . Ιδιαίτερα, θα μελετήσουμε τον κύβο δεδομένων (data cube), δηλαδή το πολυδιάστατο μοντέλο δεδομένων για τις Αποθήκες Δεδομένων και την αναλυτική επεξεργασία των δεδομένων καθώς επίσης και τις βασικές λειτουργίες OLAP (τεμαχισμός – slice, κομμάτιασμα – dice, συσσώρευση – roll-up, εμβάθυνση – drill-down). Η αναλυτική αναφορά στην τεχνολογία που έχει αναπτυχθεί γύρο από τις

Αποθήκες Δεδομένων θα συνεχιστεί στις επόμενες παραγράφους με την αναλυτική παρουσίαση των τεχνικών εξόρυξης γνώσης.

3.1.1 Η Αποθήκη Δεδομένων – Τι είναι και ποιες οι διαφορές της από τις λειτουργικές βάσεις δεδομένων

Η χρησιμοποίηση της τεχνολογίας των Αποθηκών Δεδομένων παρέχει στους αναλυτές και τα διευθυντικά στελέχη των επιχειρήσεων τεχνικές και εργαλεία για την συστηματική οργάνωση, κατανόηση και τελικά χρήση των δεδομένων για την χάραξη στρατηγικού σχεδιασμού. Τα συστήματα Αποθηκών Δεδομένων αποτελούν χρήσιμα εργαλεία στο σύγχρονο ανταγωνιστικό και γρήγορα εξελισσόμενο κόσμο. Τα τελευταία χρόνια η ανάπτυξη και λειτουργία Αποθηκών Δεδομένων είναι πολύ σημαντική για την λειτουργία οργανισμών και επιχειρήσεων καθώς οι περισσότεροι πιστεύουν ότι στη σύγχρονη ανταγωνιστική βιομηχανία τέτοια συστήματα παρέχουν σημαντικά οφέλη, με αποτέλεσμα να επενδύονται τεράστια ποσά σε αντίστοιχες δραστηριότητες. Αυτός είναι και ο λόγος που όλες οι μεγάλες εταιρείες του χώρου των Βάσεων Δεδομένων και των πληροφοριακών συστημάτων αναπτύσσουν και προτείνουν προϊόντα στο χώρο των Αποθηκών Δεδομένων και στα επόμενα χρόνια αναμένονται κόμη μεγαλύτερες επενδύσεις σε αντίστοιχη τεχνολογία.

Οι Αποθήκες Δεδομένων έχουν ορισθεί με ποικίλους τρόπους γεγονός που καθιστά δύσκολη τη διατύπωση ενός αυστηρού ορισμού. Γενικά, με τον όρο **Αποθήκες Δεδομένων** (Data Warehouses) χαρακτηρίζεται ένα σύνολο τεχνολογιών που υποστηρίζουν την αποδοτική πρόσβαση στα δεδομένα ενός οργανισμού και την επεξεργασία τους με σκοπό τη σχεδίαση της πολιτικής του. Ουσιαστικά, μια Αποθήκη Δεδομένων είναι μια βάση δεδομένων που υποστηρίζει αποφάσεις και συντηρείται ξεχωριστά από την λειτουργική βάση δεδομένων (Operational Database) ενός οργανισμού.

Ένας γενικός ορισμός δίδεται ως εξής: ΄΄ Οι Αποθήκες Δεδομένων (Data Warehouses) αποτελούν θεματο- κεντρικά (subject- oriented), συγκεντρωμένα (integrated), με χρονική διάσταση (time- variant), μη ευμετάβλητα (non- volatile) συστήματα διαχείρισης πληροφοριακών δεδομένων για την υποστήριξη των διαδικασιών λήψης

αποφάσεων ΄΄. Ο σύντομος αλλά ωστόσο πολύ περιεκτικός αυτός ορισμός συνοψίζει τα κύρια χαρακτηριστικά μιας Αποθήκης Δεδομένων. Οι τέσσερις λέξεις κλειδιά, θεματο- κεντρικά, συγκεντρωμένα, με χρονική διάσταση και μη ευμετάβλητα, διαφοροποιούν τις Αποθήκες Δεδομένων από άλλα συστήματα αποθήκευσης δεδομένων, όπως για παράδειγμα τα σχεσιακά συστήματα βάσεων δεδομένων, τα συστήματα επεξεργασίας συναλλαγών και τα συστήματα αρχείων. Ειδικότερα:

- Θεματο- κεντρικά: Μια Αποθήκη Δεδομένων οργανώνεται γύρω από συγκεκριμένα θέματα, όπως πελάτες, προμηθευτές, προϊόντα, πωλήσεις και επικεντρώνεται στην μοντελοποίηση και στην ανάλυση των δεδομένων για την λήψη αποφάσεων. Δεδομένα ή πλευρές θεμάτων που χρησιμεύουν στη διαδικασία υποστήριξης αποφάσεων δεν συμπεριλαμβάνονται στην Αποθήκη.
- Συγκεντρωμένα: Μια Αποθήκη Δεδομένων συνήθως κατασκευάζεται με συγκέντρωση πολλαπλών ετερογενών πηγών, όπως σχεσιακές βάσεις δεδομένων, αρχεία κλπ. Στη συνέχεια εφαρμόζονται τεχνικές καθαρισμού και ολοκλήρωσης δεδομένων (π.χ. δομές κωδικοποίησης, μέτρα των χαρακτηριστικών, συμβάσεις ονοματολογίας κλπ) για την εξασφάλιση συνέπειας των ετερογενών δεδομένων.
- Με χρονική διάσταση: Τα δεδομένα αποθηκεύονται για αν παρέχουν ιστορική πληροφορία (π.χ. για τα τελευταία 5-10 χρόνια). Γενικά, η έννοια του χρόνου είναι αναπόσπαστο τμήμα μιας Αποθήκης Δεδομένων.
- Μη ευμετάβλητα: Μια Αποθήκη Δεδομένων αποθηκεύεται ξεχωριστά από τη λειτουργική βάση δεδομένων και τα δεδομένα της δεν υπόκεινται σε τροποποιήσεις (π.χ. επεξεργασία συναλλαγών, ανάνηψη, έλεγχος συνδρομικότητας – concurrency control). Στις Αποθήκες Δεδομένων υπάρχει μόνο η λειτουργία της φόρτωσης είτε πλήρως (full loading) είτε αυξητικά (incremental loading).

Η κατασκευή και η συντήρηση μιας Αποθήκης Δεδομένων είναι μια πολύπλοκη διαδικασία και εξαρτάται από τις ανάγκες κάθε οργανισμού ή επιχείρησης. Πολλοί οργανισμοί επιδιώκουν να δημιουργήσουν μια Αποθήκη Δεδομένων που θα περιέχει αναλυτικά δεδομένα από όλες τις δραστηριότητες του οργανισμού. Ένα τέτοιο εγχείρημα απαιτεί πολύ μεγάλο κόστος για επιτύχει. Συχνά, μια Αποθήκη Δεδομένων συμπληρώνεται από εξειδικευμένα θεματικά υποσύνολα ή αλλιώς από Επιμέρους Συλλογές Δεδομένων (data marts) για επιπλέον απόδοση

των OLAP εφαρμογών. Οι Συλλογές Δεδομένων είναι πιο ευέλικτα συστήματα στη δημιουργία τους, τα οποία έχουν ως στόχο την ολοκλήρωση (integration) ετερογενών πηγών πληροφοριών με τη συγκέντρωση όλης της ενδιαφέρουσας πληροφορίας σε μια τοποθεσία και την αποφυγή της σύγκρουσης μεταξύ OLTP (on-line transaction processing) και OLAP (on-line analytical processing) συστημάτων ώστε να εξασφαλίζεται η απόδοση των εφαρμογών και η διαθεσιμότητα του συστήματος. Ωστόσο, η μακρόχρονη χρήση του δημιουργεί προβλήματα.

Επειδή οι περισσότεροι άνθρωποι είναι εξοικειωμένοι με τα εμπορικά συστήματα σχεσιακών βάσεων δεδομένων, είναι αρκετά εύκολο να κατανοήσουμε τι είναι μια Αποθήκη Δεδομένων συγκρίνοντας τα συστήματα OLTP/OLAP.

Ένα Σύστημα Επεξεργασίας Συναλλαγών (OLTP) παρέχει ένα πλήρες σύστημα που περιέχει εργαλεία για τον προγραμματισμό των εφαρμογών, την εκτέλεση και τη διαχείριση των συναλλαγών. Είναι μια εφαρμογή που δουλεύει συνεχώς, είναι συνήθως κατανεμημένη και περιλαμβάνει μια βάση δεδομένων, κάποιο δίκτυο και τα αντίστοιχα προγράμματα για την εφαρμογή. Από την άλλη πλευρά ένα Σύστημα Αναλυτικής Επεξεργασίας Συναλλαγών (OLAP) παρέχει ευέλικτη, υψηλής απόδοσης πρόσβαση και ανάλυση μεγάλου όγκου σύνθετων δεδομένων από διαφορετικές εφαρμογές, συμμετοχή αθροιστικών και ιστορικών δεδομένων σε πολύπλοκες ερωτήσεις, μεταβολή της ΄΄ οπτικής γωνίας ΄΄ παρουσίασης των δεδομένων (π.χ. από πωλήσεις ανά περιοχή σε πωλήσεις ανά τμήμα κλπ), συμμετοχή πολύπλοκων υπολογισμών (π.χ. στατιστικές συναρτήσεις) και γρήγορες απαντήσεις σε οποιαδήποτε χρονική στιγμή τεθεί ένα ερώτημα (On-Line). Στη συνέχεια παρουσιάζονται αναλυτικά τα βασικά χαρακτηριστικά που διαφοροποιούν τα OLTP συστήματα από τα OLAP ενώ στον πίνακα 1 που ακολουθεί γίνεται μια σύνοψη των διαφορών.

- Χρήστες και προσανατολισμός του συστήματος : Ένα OLTP σύστημα προσανατολίζεται στις απαιτήσεις του πελάτη και χρησιμοποιείται από διοικητικούς υπαλλήλους και διαχειριστές της βάσης δεδομένων του

οργανισμού. Ένα OLAP σύστημα προσανατολίζεται στις απαιτήσεις της αγοράς και χρησιμοποιείται από διευθυντικά στελέχη και αναλυτές.

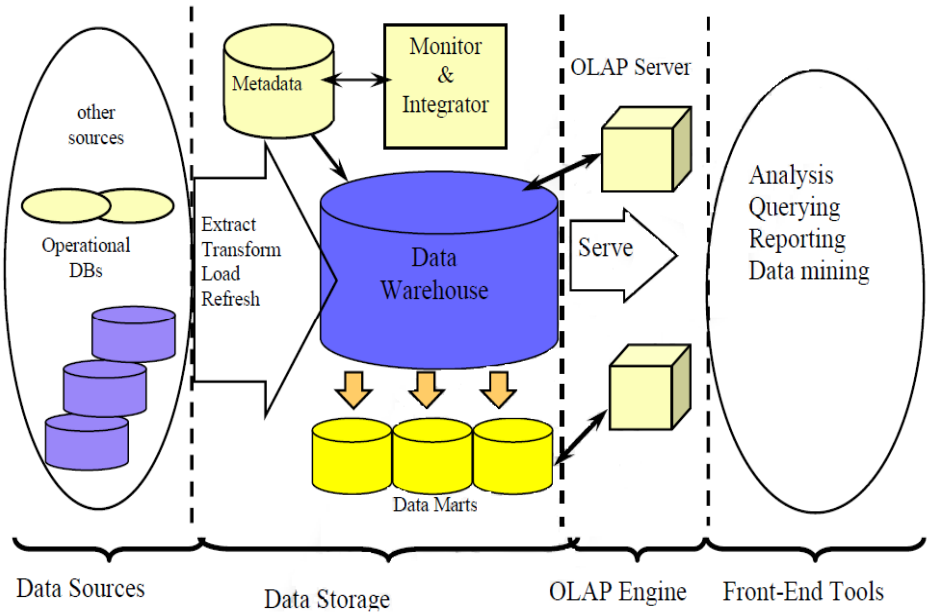
- Περιεχόμενα δεδομένων: Ένα OLTP σύστημα διαχειρίζεται τρέχοντα – καθημερινά δεδομένα μεγάλης λεπτομέρειας τα οποία μπορούν εύκολα να αναζητηθούν και απαντούν σε απλές ερωτήσεις. Ένα OLAP σύστημα διαχειρίζεται μεγάλες ποσότητες ιστορικής πληροφορίας και παρέχει αποδοτική πρόσβαση στα δεδομένα για λήψη αποφάσεων.
- Σχεδιασμός βάσης δεδομένων: Ένα OLTP σύστημα σχεδιάζεται για να διατηρεί την ακεραιότητα των δεδομένων και να εξασφαλίζει ταχύτητα στην αποθήκευση των καθημερινών συναλλαγών του οργανισμού, επομένως η βάση δεδομένων του συστήματος είναι κανονικοποιημένη βάσει κάποιου μοντέλου Οντοτήτων – Συσχετίσεων (Entity- Relationship model). Ένα OLAP σύστημα σχεδιάζεται για να παρέχει ταχύτητα στην ανάλυση και η βάση δεδομένων του συστήματος είναι από-κανονικοποιημένη βάσει κάποιου μοντέλου αστέρα ή χιονονιφάδας (star/ snowflake schema αντίστοιχα) καθώς οι εφαρμογές OLAP επιταχύνονται αν τα δεδομένα οργανωθούν με μη παραδοσιακούς τρόπους.
- Πρότυπα πρόσβασης (access patterns): Τα δεδομένα ενός OLTP συστήματος υπόκεινται σε λειτουργίες τροποποίησης (π.χ. επεξεργασία συναλλαγών, ανάνηψη, έλεγχος συνδρομικότητας). Από την άλλη πλευρά, τα OLAP συστήματα περιέχουν ιστορική πληροφορία που δεν μεταβάλλεται και επομένως η πρόσβαση σε αυτά επιτρέπει λειτουργίες μόνο για ανάγνωση (read- only).

Πίνακας 1. Σύγκριση συστημάτων OLTP/OLAP

	OLTP	**OLAP**
Δομή	Files /DBMS's	RDBMS
Πρόσβαση	SQL/COBOL/…	SQL και επεκτάσεις
Ανάγκες που καλύπτουν	Αυτοματισμός καθημερινών εργασιών	Άντληση και επεξεργασία πληροφορίας για χάραξη στρατηγικής
Τύπος Δεδομένων	Λεπτομερή, λειτουργικά	Συνοπτικά, αθροιστικά
Όγκος Δεδομένων	Από 100MB έως GB	Από 100GB έως TB
Φύση Δεδομένων	Δυναμικά, τρέχοντα	Στατικά, ιστορικά
I/O Τύποι	Περιορισμένο I/O συχνές αναζητήσεις στο δίσκο	Εκτεταμένο I/O συχνές σαρώσεις του δίσκου
Τροποποιήσεις	Συνεχείς	Περιοδικές ενημερώσεις
Μέτρηση Απόδοσης	Μέσος Ρυθμός Αποθήκευσης Εγγραφών-Throughput	Χρόνος απόκρισης
Φόρτος	Συναλλαγές με πρόσβαση λίγων εγγραφών	Ερωτήσεις που σαρώνουν εκατομμύρια εγγραφών
Σχεδίασης Βάσης Δεδομένων	Κατευθυνόμενη από εφαρμογή	Κατευθυνόμενη από περιεχόμενο
Τυπικοί Χρήστες	Χαμηλόβαθμοι υπάλληλοι, π.χ. διοικητικοί υπάλληλοι, διαχειριστές βάσης δεδομένων	Υψηλόβαθμοι υπάλληλοι π.χ. διευθυντικά στελέχη, αναλυτές
Χρήση	Μέσω προκατασκευασμένων φορμών	Ad –hoc
Αριθμός Χρηστών	Χιλιάδες	Δεκάδες
Εστίαση	Εισαγωγή δεδομένων	Εξαγωγή πληροφοριών

3.1.2 Η αρχιτεκτονική της αποθήκης δεδομένων

Η επιλογή της αρχιτεκτονικής για μια αποθήκη δεδομένων πρέπει να ικανοποιεί τις συγκεκριμένες ανάγκες του οργανισμού για τις οποίες δημιουργήθηκε ώστε να εξασφαλίζεται η διαθεσιμότητα και η αποδοτικότητα του συστήματος. Γενικά, η αρχιτεκτονική μιας αποθήκης δεδομένων είναι όπως παρουσιάζεται στο σχήμα 3 όπου σημειώνονται τα βασικά δομικά στοιχεία της αποθήκης, η διασύνδεση των στοιχείων τους και η ροή των δεδομένων.



Σχήμα 3. Αρχιτεκτονική Αποθήκευση Δεδομένων

Τα κύρια δομικά μέρη της αρχιτεκτονικής μιας Αποθήκης Δεδομένων είναι τα παρακάτω:

Πηγές δεδομένων (Data sources): Κάθε πηγή από την οποία η Αποθήκη Δεδομένων αντλεί δεδομένα. Τα συστήματα διαχείρισης Αποθηκών Δεδομένων αντλούν από διάφορες ετερογενείς πηγές, όπως για παράδειγμα:

- Βάσεις δεδομένων των συστημάτων του οργανισμού.
- Εξωτερικές πηγές πληροφοριών, δηλαδή, πληροφορίες που προέρχονται από τα πληροφοριακά συστήματα και στα οποία ο οργανισμός έχει πρόσβαση.
- Αρχεία εφαρμογών και αρχεία κειμένου.

ETL (Extract – Transform- Load) εφαρμογές: εφαρμογές που εκτελούν τις διαδικασίες εξαγωγής, μεταφοράς, μετασχηματισμού, καθαρισμού και φόρτωσης των δεδομένων από τις πηγές στην Αποθήκη Δεδομένων. Πιο αναλυτικά οι παρακάτω εφαρμογές αυτοματοποιούν διαδικασίες όπως:

- 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 από τις πηγές 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.

Τέλος, η Ενημέρωση (Refresh) της Αποθήκης Δεδομένων είναι η διαδικασία που μεταφέρει τις αλλαγές που συμβαίνουν στα δεδομένα των πηγών εκτελώντας τις αντίστοιχες αλλαγές στα δεδομένα της Αποθήκευσης. Συνήθως, οι Αποθήκες Δεδομένων ενημερώνονται περιοδικά (π.χ. ανά εβδομάδα, μήνα κλπ). Υπάρχουν όμως και περιπτώσεις που για τις ανάγκες της ανάλυσης απαιτείται άμεση πρόσβαση σε τρέχοντα δεδομένα, με αποτέλεσμα να επιβάλλεται και η άμεση ενημέρωση των

Αποθηκών για κάθε μεταβολή στις πηγές. Ωστόσο, επειδή οι Αποθήκες Δεδομένων συσσωρεύουν μεγάλη ποσότητα δεδομένων η εφαρμογή της διαδικασίας ενημέρωσης καθίσταται απαγορευτική. Γι' αυτό και είναι αναγκαία η τροποποιήσεις των μεταβολών που συμβαίνουν στις πηγές (εισαγωγές, διαγραφές και τροποποιήσεις εγγράφων), ώστε σε κάθε διαδικασία ενημέρωσης να μη γίνονται περιττές διαγραφές και εισαγωγές δεδομένων που στην πραγματικότητα παραμένουν αναλλοίωτα. Πάντως, κάθε φορά η πολιτική ενημέρωσης καθορίζεται από το διαχειριστή της Αποθήκης Δεδομένων βάσει των αναγκών των εφαρμογών ανάλυσης, τη διαθεσιμότητα των πηγών και την κατάσταση του δικτύου που συνδέει την Αποθήκη με τις πηγές.

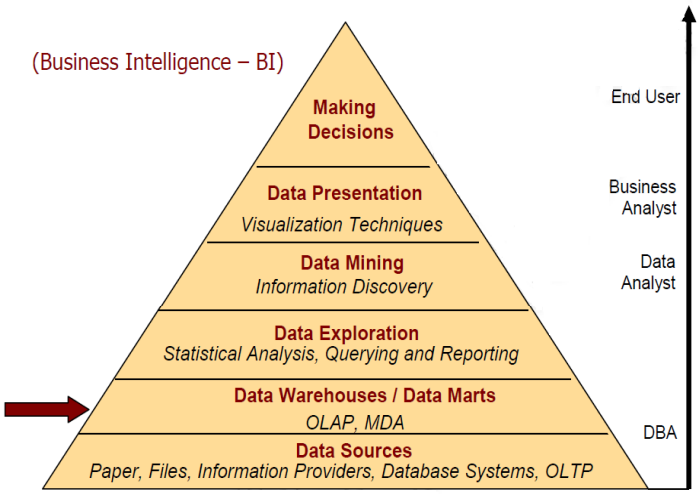
***Αποθήκη Δεδομένων (Data Warehouse), Συλλογές Δεδομένων (Data Marts*)**: είναι τα συστήματα που αποθηκεύονται τα δεδομένα που παρέχονται προς τους χρήστες, τα οποία υλοποιούνται με τη χρήση Σχεσιακών Συστημάτων Διαχείρισης Βάσεων Δεδομένων. Τα δεδομένα αποθηκεύονται σε σχεσιακές βάσεις δεδομένων και η πρόσβαση σε αυτά γίνεται μέσω μιας γλώσσας διαχείρισης δεδομένων που είναι επέκταση της SQL. Εναλλακτικά χρησιμοποιούνται και Πολυδιάστατα Συστήματα Αναλυτικής Επεξεργασίας (Multidimensional OLAP servers), που αποθηκεύουν και διαχειρίζονται δεδομένα με πολυδιάστατο τρόπο. Το κυριότερο πλεονέκτημα των πολυδιάστατων συστημάτων σε σύγκριση με την ευελιξία που χαρακτηρίζει τα Σχεσιακά Συστήματα Διαχείρισης βάσεων Δεδομένων είναι η δυνατότητά τους να διαχειρίζονται δεδομένα, τα οποία είναι δομημένα με τρόπο που βρίσκεται πιο κοντά στις ανάγκες των εφαρμογών ανάλυσης (OLAP). Οι Συλλογές Δεδομένων περιέχουν τμήματα των δεδομένων της Αποθήκης Δεδομένων και η ύπαρξη τους είναι επιλογή του διαχειριστή του συστήματος. Ο καταμερισμός των δεδομένων της Αποθήκης σε επιμέρους Συλλογές ανά αντικείμενο ή τμήμα γίνεται κυρίως με οργανωτικά κριτήρια και έχει ως στόχο την άμεση και αποδοτική πρόσβαση των εφαρμογών ανάλυσης στα δεδομένα της αποθήκης.

***Βάση Μετά- Δεδομένων (Metadata Repository)*:** Είναι το υποσύστημα αποθήκευσης πληροφορίας σχετικά με την δομή και λειτουργία όλου του συστήματος και όπως φαίνεται στο σχήμα 4 από το υποσύστημα αυτό υπάρχει πρόσβαση σε όλα τα δομικά στοιχεία της αρχιτεκτονικής της Αποθήκης Δεδομένων. η κατανόηση και η καταγραφή του περιεχομένου των δεδομένων και της οργάνωσής τους είναι απαραίτητα για την αποδοτική λειτουργία και διαχείριση της Αποθήκης. Τα μετά- δεδομένα πρέπει να περιέχουν :

- Λεξικό δεδομένων (Data Dictionary) που περιέχει τον ορισμό και την περιγραφή των δεδομένων που αποθηκεύονται στην Αποθήκη Δεδομένων και της 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).

Οι Αποθήκες Δεδομένων συγκεντρώνουν μεγάλο όγκο ετερογενών δεδομένων σε πολυδιάστατο χώρο. Η αρχιτεκτονική τους περιλαμβάνει μεταξύ άλλων τον καθαρισμό, την ολοκλήρωση, την μετατροπή και την εισαγωγή των δεδομένων από τις πηγές στην Αποθήκη. Η σχεδίαση της αρχιτεκτονικής ενός συστήματος αποθήκης δεδομένων αποτελεί μια πολύπλοκη διαδικασία και μπορεί να θεωρηθεί ως το κύριο βήμα προεπεξεργασίας των δεδομένων για την εξόρυξη γνώσης που κατ' επέκταση αποτελεί το στάδιο που προηγείται της ολοκλήρωσης της διαδικασίας ανακάλυψης γνώσης. Ουσιαστικά, πάνω στην αρχιτεκτονική της Αποθήκης Δεδομένων βασίζονται οι εφαρμογές ανάλυσης, όπως για παράδειγμα οι εφαρμογές παραγωγής αναφορών (Reporting), η αναλυτική επεξεργασία δεδομένων με σύνθετα ερωτήματα (OLAP Querying), η ανάλυση για λήψη αποφάσεων (Analysis) και η Εξόρυξη Γνώσης (Data Mining). Όπως φαίνεται στο σχήμα 4, οι αποθήκες δεδομένων βρίσκονται στην βάση της « Πυραμίδας » της Επιχειρηματικής Ευφυΐας που μπορεί να αποκομίσει ο τελικός

χρήστης (End User) από πρωτογενή δεδομένα με την βοήθεια της σύγχρονης τεχνολογίας.



Σχήμα 4. « Πυραμίδα » Επιχειρηματικής Ευφυΐας

3.1.3 Η Εννοιολογική Σχεδίαση της Αποθήκης Δεδομένων

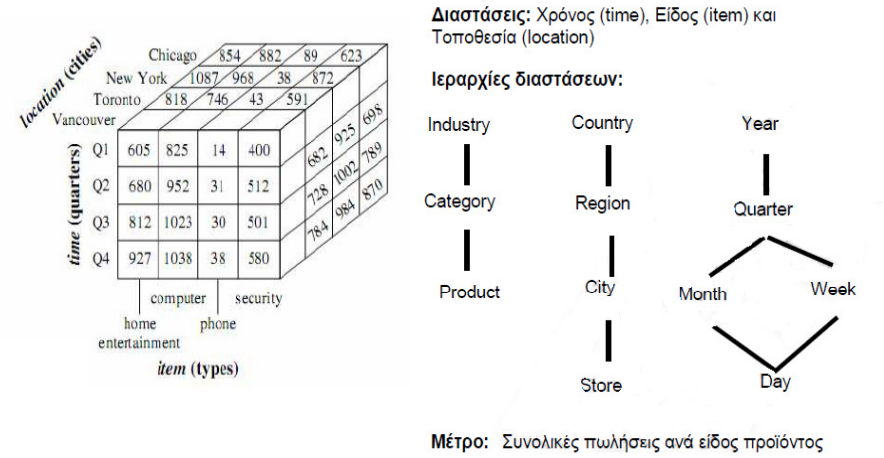
Οι Αποθήκες Δεδομένων χρησιμοποιούνται για την αναλυτική επεξεργασία μεγάλου όγκου δεδομένων με σκοπό την απάντηση πολύπλοκων ερωτήσεων σε αποδεκτούς χρόνους. Αυτός είναι και ο λόγος που τόσο η σχεδίαση όσο και η οργάνωση των δεδομένων τους είναι διαφορετική από τις παραδοσιακές σχεσιακές βάσεις δεδομένων. Τα διαγράμματα Οντοτήτων – Συσχετίσεων (Entity – Relationship) και οι τεχνικές κανονικοποιήσης των OLTP συστημάτων αποδεικνύονται ακατάλληλα για τη

σχεδίαση των Αποθηκών Δεδομένων. Η τεχνική που χρησιμοποιείται στις Αποθήκες Δεδομένων είναι το Μοντέλο Διαστάσεων (Dimensional Modeling) ή διαφορετικά το πολυδιάστατο μοντέλο.

Το Μοντέλο βασίζεται στη θεώρηση των δεδομένων μέσω ενός πολυδιάστατου μοντέλου δεδομένων το οποίο απεικονίζει τα δεδομένα σε μορφή κύβου. Αν και ο όρος κύβος περιγράφει μια γεωμετρική δομή τριών διαστάσεων, στις Αποθήκες Δεδομένων ο κύβος δεδομένων είναι ν- διαστάσεων. Δηλαδή, η χρήση ενός κύβου διαστάσεων επιτρέπει τη θεώρηση των δεδομένων σε πολλαπλές διαστάσεις. Τα βασικά στοιχεία του Μοντέλου είναι οι :

- Πίνακες Διαστάσεων (Dimension Tables) με πληροφορία για τις διαστάσεις του κύβου.
- Πίνακες Γεγονότων (Fact Tables) με μέτρα (κάποια μετρήσιμα μεγέθη) και κλειδιά προς τους σχετιζόμενους πίνακες διαστάσεων.

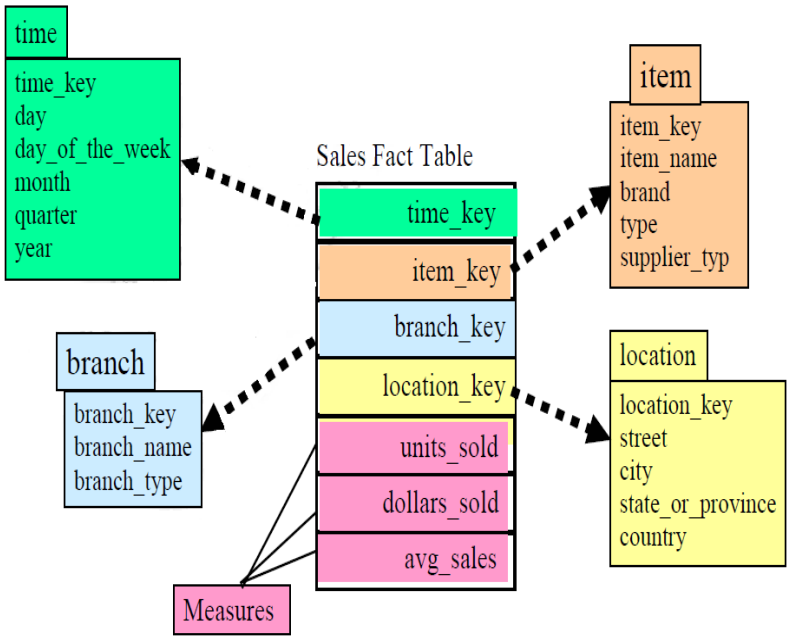
Ένα πολυδιάστατο μοντέλο δεδομένων οργανώνεται γύρω από ένα κεντρικό θέμα, όπως για παράδειγμα οι πωλήσεις, το οποίο εμφανίζεται στον πίνακα γεγονότων. Το **σχήμα 5** αποτελεί μια απεικόνιση του πολυδιάστατου μοντέλου δεδομένων.



Σχήμα 5. Απεικόνιση ενός 3- διάστατου κύβου δεδομένων με διαστάσεις χρόνου, είδους προϊόντος και τοποθεσίας. Το μέτρο συνολικές πωλήσεις προϊόντος εκφράζεται σε χιλιάδες δολάρια.

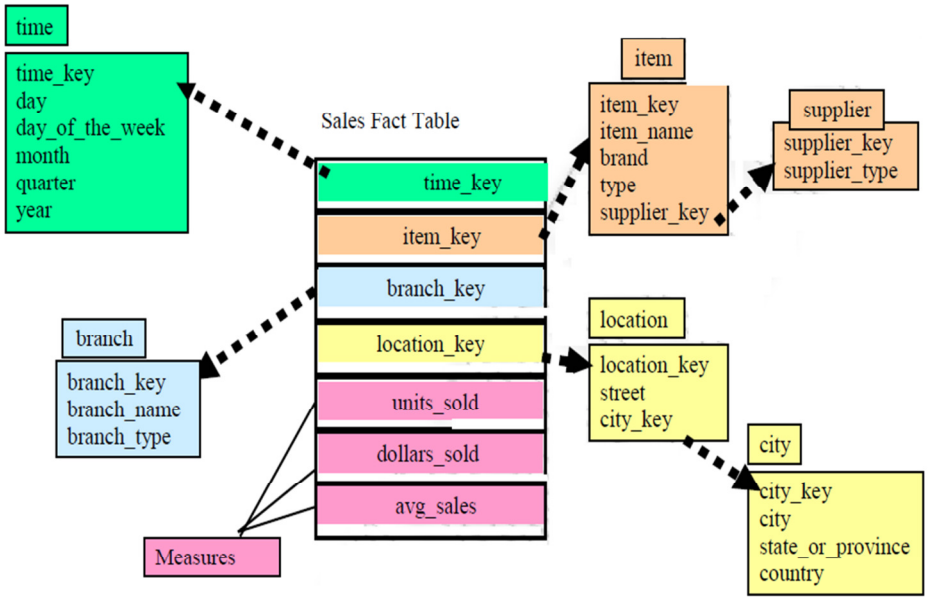
Υπάρχουν δύο βασικές κατηγορίες σχημάτων για τη σχεδίαση των βάσεων των Αποθηκών Δεδομένων, το Σχήμα Αστέρα (Star Schema) και το Σχήμα Χιονονιφάδας (Snowflake Schema) ενώ μια επιπλέον τεχνική σχεδίασης αποτελούν οι Αστερισμοί Γεγονότων (Fact Constellations) ή εναλλακτικά Σχήμα Γαλαξία καθώς πρόκειται για συλλογή σχημάτων αστέρων.

Σχήμα Αστέρα: Αποτελείται από έναν κεντρικό πίνακα γεγονότων και κάποιους από-κανονικοποιημένους πίνακες διαστάσεων. Τα μέτρα είναι τα ενδιαφέροντα μεγέθη υπό μέτρηση (π.χ. units_ sold, dollars_ sold στο πίνακα SALES). Για κάθε διάσταση του μοντέλου, εισάγεται ένας πίνακας (π.χ. Time, Branch, Location και Item), ο οποίος παρέχει όλα τα επίπεδα συνάθροισης (levels of aggregation) καθώς και τις σχετικές τους ιδιότητες. Το **σχήμα 6** παρουσιάζει ένα παράδειγμα σχήματος αστέρα.



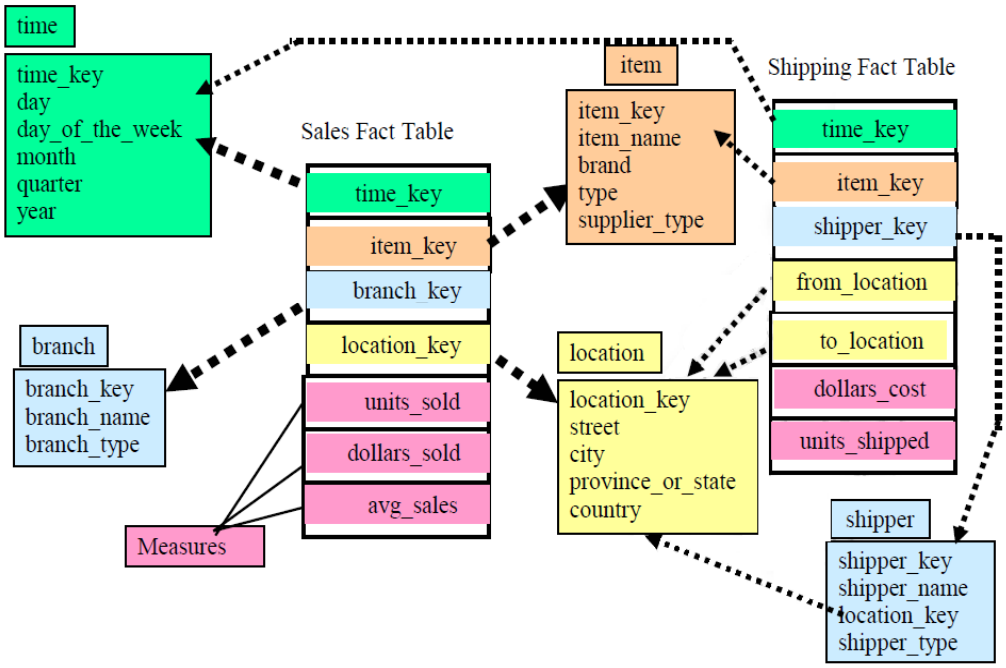
Σχήμα 6. Παράδειγμα σχήματος Αστέρα.

Σχήμα Χιονονιφάδας: Αποτελεί την κανονικοποιημένη εκδοχή του σχήματος αστέρα. Για κάθε επίπεδο της ιεραρχίας των διαστάσεων εισάγεται και ένας πίνακας. Το σχήμα αυτό εγγυάται την ακεραιότητα των δεδομένων (όπως όλα τα κανονικοποιημένα σχήματα), αλλά είναι πιο αργό στις απαντήσεις των ερωτήσεων. Στο **σχήμα 7** παρουσιάζεται ένα παράδειγμα βάσης που έχει το ίδιο περιεχόμενο με την βάση του σχήματος 6, μόνο που είναι οργανωμένη με το σχήμα χιονονιφάδας. Ουσιαστικά αποτελεί μια βελτίωση του σχήματος αστέρα, όπου η ιεραρχία των διαστάσεων αναπαριστάται με κανονικοποίηση των πινάκων διαστάσεων.



Σχήμα 7. Παράδειγμα σχήματος Χιονονιφάδας

Αστερισμός Γεγονότων: Το σχήμα αυτό χρησιμοποιείται όταν χρειάζεται να υπάρχουν πολλοί πίνακες γεγονότων, οι οποίοι να μοιράζονται τους πίνακες διαστάσεων. Είναι συχνό φαινόμενο στις Αποθήκες Δεδομένων αλλά πιο σπάνιο στη σχεδίαση Συλλογών Δεδομένων. το σχήμα 8 απεικονίζει ένα παράδειγμα Αστερισμού Γεγονότων με χρήση του ίδιου περιεχομένου βάσης, όπως και στα σχήματα 6 και 7.



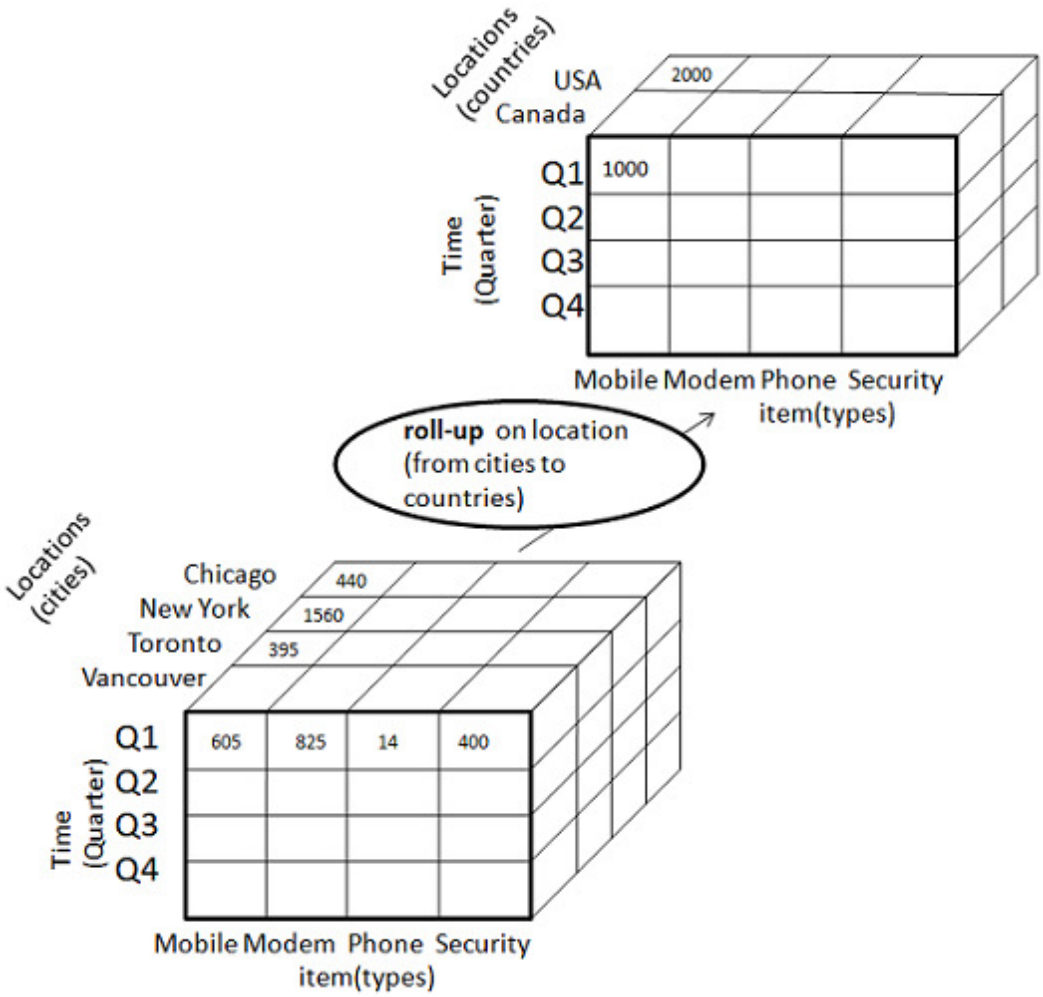
Σχήμα 8. Παράδειγμα σχήματος Αστερισμού Γεγονότων

Η αναλυτική επεξεργασία δεδομένων είναι τμήμα των εφαρμογών στήριξης αποφάσεων και των στρατηγικών πληροφοριακών συστημάτων. Η λειτουργία της αναλυτικής επεξεργασίας δεδομένων (OLAP) χαρακτηρίζεται από την δυναμική πολυδιάστατη ανάλυση των δεδομένων ενός οργανισμού με εκτέλεση ερωτήσεων πάνω στα δεδομένα. Οι ερωτήσεις έχουν συγκεκριμένη και πολύπλοκη δομή, ενώ η πληροφορία που αντλούν έχει πολυδιάστατο χαρακτήρα. Τα πολυδιάστατα μοντέλα δεδομένων περιέχουν ν – διάστατους πίνακες που συχνά αποκαλούνται υπερκύβοι (cubes ή hyper cubes). Κάθε διάσταση έχει μια ιεραρχία επιπέδων, π.χ. η διάσταση ΄΄ Γεωγραφική Τοποθεσία ΄΄ έχει τα επίπεδα πόλη, περιοχή, χώρα. Οι τιμές (μετρήσιμα μεγέθη) που περιέχουν οι υπερκύβοι αντιστοιχούν στις στήλες των σχεσιακών πινάκων. Ένα παράδειγμα αναλυτικής επεξεργασίας δεδομένων είναι μια εφαρμογή

που εκτελεί ερωτήσεις για να μπορεί να έχει συγκεντρωτικά δεδομένα για τις πωλήσεις ανά προϊόν, ανά μήνα και ανά περιοχή ενός οργανισμού. Η παρουσίαση των αποτελεσμάτων των πωλήσεων μπορεί να προκαλέσει το χρήστη στην εκτέλεση μιας πιο συγκεντρωτικής ερώτησης, ώστε να πάρει ως απάντηση τα δεδομένα που αφορούν τις ετήσιες πωλήσεις ανά προϊόν και περιοχή, ή να εκτελέσει μια πιο λεπτομερή ερώτηση παίρνοντας ως απάντηση τις μηνιαίες πωλήσεις κάθε προϊόντος ανά συγκεκριμένο πελάτη.

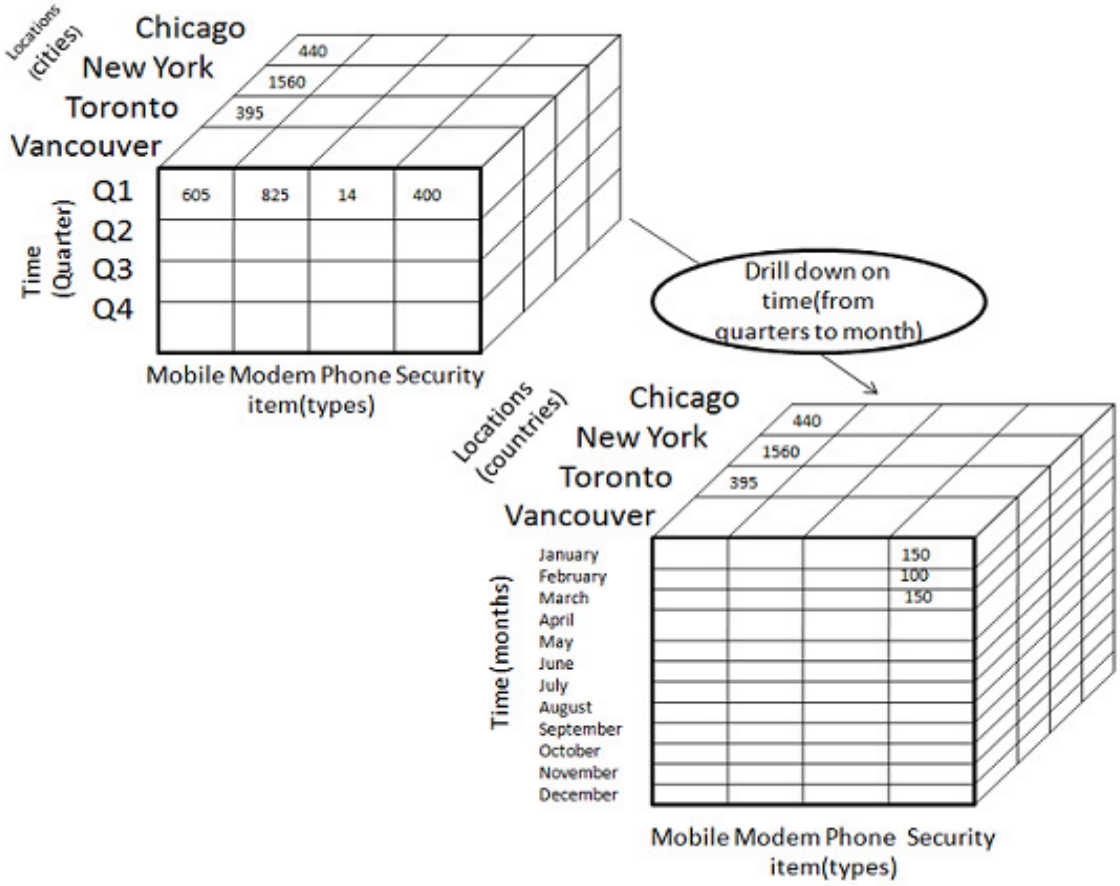
Οι κύβοι δίνουν τη δυνατότητα πλοήγησης στις ιεραρχίες των διαστάσεών τους. Η πλοήγηση είναι δυνατή από τις λειτουργίες τις οποίες παρέχουν. Οι OLAP λειτουργίες που γίνονται συνήθως στους κύβους είναι οι παρακάτω ενώ σχετικά παραδείγματα παρουσιάζονται στα παρακάτω σχήματα:

- **Roll- up:** πρόκειται για πράξη με την οποία εκτελούμε ένα βήμα ανόδου στην ιεραρχία μιας διάστασης. 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. Παράδειγμα στην διάσταση Τοποθεσία (location) ανεβαίνουμε από το επίπεδο Πόλη (city) στο επίπεδο Χώρα (country).



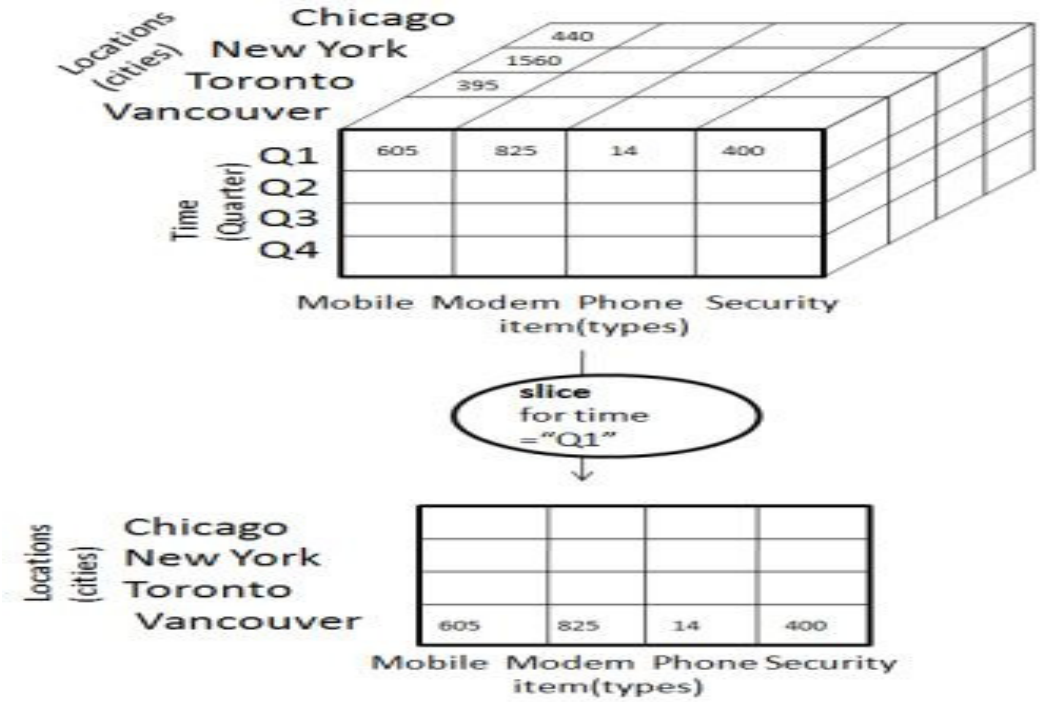
Σχήμα 10. Η πράξη Roll- up.

- **Drill- down:** Είναι η αντίστροφη πράξη του Roll- up, όπου πάμε από ένα υψηλότερο επίπεδο ιεραρχίας μιας διάστασης σε ένα χαμηλότερο. Παράδειγμα στη διάσταση Τοποθεσία (location) κατεβαίνουμε από το επίπεδο Χώρα (country) στο επίπεδο Πόλη (city).



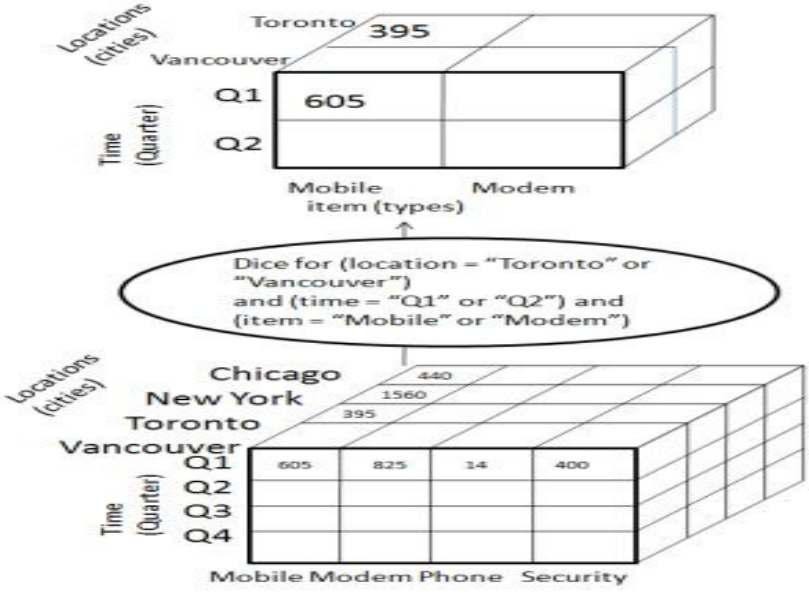
Σχήμα 11. Η πράξη Drill- down.

- **Slice:** Πρόκειται για πράξη επιλογής δεδομένων σε μια συγκεκριμένη διάσταση. Ένα επίπεδο (slice)είναι ένα υποσύνολο ενός υπερκύβου σύμφωνα με μία περιοχή τιμών ή μια συγκεκριμένη τιμή ενός επιπέδου διάστασης.



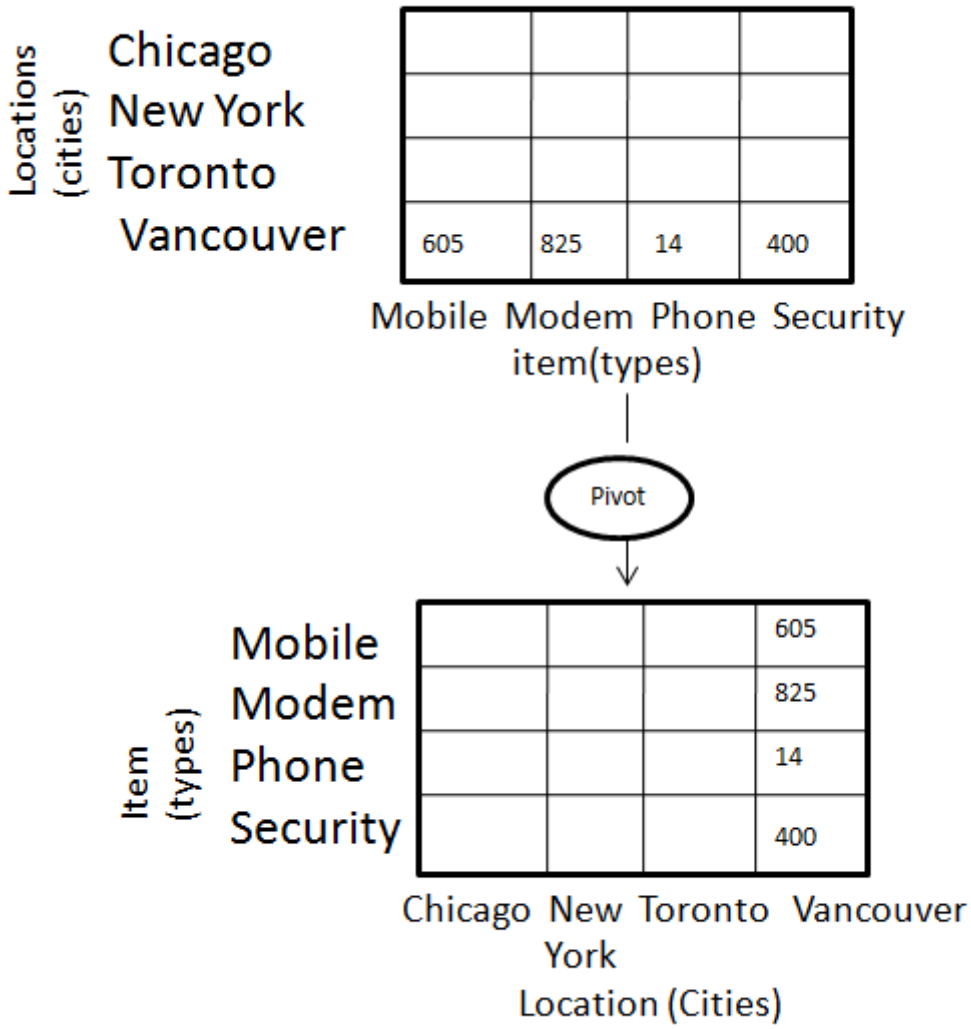
Σχήμα 12. Η πράξη Slice

- **Dice:** Πρόκειται για μια λειτουργία επιλογής δεδομένων από δύο ή και περισσότερες διαστάσεις.



Σχήμα 13. Η πράξη Dice

- **Pivot:** Πρόκειται για λειτουργία αλλαγής της διάταξης των διαστάσεων ώστε να διευκολυνθεί η ανάλυση. Κατά την περιστροφή, δεν μεταβάλλονται ούτε μειώνονται τα δεδομένα του υπερκύβου, απλά αλλάζει ο τρόπος παρουσίασής τους στην εφαρμογή ανάλυσης.



Σχήμα 14. Η πράξη Pivot.

3.2 Εξόρυξη Γνώσης από τα Δεδομένα

Τα τελευταία χρόνια, κυρίως λόγω των δυνατοτήτων που προσφέρουν οι νέες τεχνολογικές εξελίξεις, τεράστιος είναι ο όγκος των δεδομένων κάθε είδους που αποθηκεύονται σε αρχεία και βάσεις δεδομένων. εκείνοι οι οποίοι έχουν την ικανότητα να συλλέγουν πληροφορίες και δεδομένα και έπειτα να τα αναλύουν και να τα αξιοποιούν, μοιραία είναι σε θέση να πρωταγωνιστήσουν σε όποιο πεδίο δραστηριοποιούνται. Η πληροφορία και η αξιοποίησή της, καθώς και η ανάλυση διαφόρων δεδομένων τα οποία μπορούν να συλλεχθούν δίνει τη δυνατότητα σε κάθε ενδιαφερόμενο να αποκτήσει ένα ανταγωνιστικό πλεονέκτημα στο χώρο στον οποίο δραστηριοποιείται και να πάρει τελικά τις καλύτερες αποφάσεις σε θέματα που τον αφορούν. Τέτοιου είδους αναλύσεις, που λαμβάνουν χώρα σε ποιοτικά αλλά και αριθμητικά δεδομένα γίνονται με τη βοήθεια τεχνικών εξόρυξης γνώσης από δεδομένα (Data Mining), οι οποίες παρέχουν τη δυνατότητα εξαγωγής κανόνων και άρα αποφάσεων με τη βοήθεια των ηλεκτρονικών υπολογιστών.

Η σημερινή εξέλιξη στις λειτουργίες και στα προϊόντα εξόρυξης γνώσης είναι αποτέλεσμα της πολυετούς επιρροής διάφορων επιστημονικών κλάδων όπως, της Μηχανικής Μάθησης (Machine Learning), της Αναγνώρισης Κανόνων (Pattern Recognition), των Βάσεων Δεδομένων (Data bases), της Στατιστικής(Statistics), της Τεχνικής Νοημοσύνης (Artificial Intelligence – AI) και των Έμπειρων Συστημάτων (Expert Systems). Οι περισσότεροι αλγόριθμοι και οι τεχνικές προέρχονται από αυτά τα πεδία. Η βάση όλων των παραπάνω είναι η απόσπαση κανόνων που περιέχουν γνώση, μέσα στο πλήθος δεδομένων.

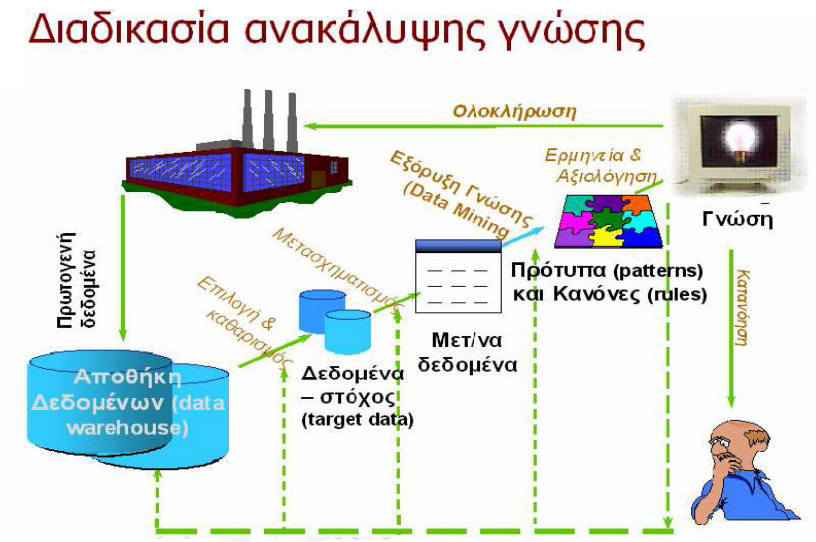
Στη συνέχεια θα γίνει αναφορά στο τι ακριβώς αντιπροσωπεύει η διαδικασία εξόρυξης γνώσης από δεδομένα και τι είδους δεδομένα χρησιμοποιεί, στα στάδια της εξόρυξης γνώσης από βάσεις δεδομένων ή πολυδιάστατα μοντέλα στις τεχνικές εξόρυξης γνώσης που χρησιμοποιούνται γενικά, καθώς και στις νέες τάσεις που επικρατούν στην αναπτυσσόμενη τεχνολογία γύρω από το επιστημονικό πεδίο της εξόρυξης γνώσης από δεδομένα.

3.2.1 Εξόρυξη γνώσης από δεδομένα και ανακάλυψη γνώσης σε βάσεις δεδομένων

Ο τεράστιος όγκος δεδομένων που αποθηκεύεται σε αρχεία, βάσεις δεδομένων και άλλα αποθηκευτικά μέσα επιβάλλει την ανάπτυξη δυναμικών μέσων ανάλυσης και ερμηνείας τέτοιων δεδομένων με σκοπό την εξαγωγή χρήσιμης γνώσης και την λήψη αποφάσεων. Η διαδικασία Data Mining, η ελληνική απόδοση της οποίας είναι ‘‘ Εξόρυξη Γνώσης από Δεδομένα ή Ανεύρεση Γνώσης από Δεδομένα’’, είναι η αναλυτική διαδικασία η οποία έχει σχεδιαστεί με σκοπό να αναλύει και να εξερευνεί δεδομένα σε μεγάλες ποσότητες και στη συνέχεια να δημιουργεί κανόνες και σχέσεις μεταξύ των μεταβλητών που ενδιαφέρουν να ερευνηθούν. Η όλη διαδικασία βασίζεται στην χρησιμοποίηση αλγορίθμων που αναζητούν κανόνες μεταξύ των μεταβλητών των δεδομένων και έπειτα βρίσκουν συσχετισμούς ή κανόνες μέσα από τεράστιες βάσεις αποθηκευμένων δεδομένων/ πληροφοριών. Επίσης η διαδικασία εξόρυξης γνώσης από δεδομένα αναφέρεται συχνά και ως Πληροφοριακή Τεχνολογία (Computerizend Technology) η οποία χρησιμοποιεί πολύπλοκους αλγορίθμους που δημιουργούν κανόνες και σχέσεις αναλύοντας τεράστιες βάσεις δεδομένων, με στόχο την λήψη στρατηγικών αποφάσεων. Η εξόρυξη γνώσης από τα δεδομένα μπορεί να οριστεί απλά ως η εύρεση πληροφορίας που είναι κρυμμένη σε μεγάλες ποσότητες δεδομένων. Υπάρχουν πολλοί άλλοι όροι που έχουν παρόμοια σημασία με την εξόρυξη γνώσης από τα δεδομένα όπως, η εξαγωγή γνώσης (knowledge extraction), η ανάλυση δεδομένων / προτύπων (data/ pattern analysis), η αρχαιολογία δεδομένων (data archaeology) και η εκβάθυνση δεδομένων (data dredging).

Οι όροι ανακάλυψης γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases – KDD) και η εξόρυξη γνώσης από δεδομένα συχνά αναφέρονται στην ίδια έννοια, δηλαδή στη διαδικασία ανακάλυψης χρήσιμων, συνήθως κρυμμένων προτύπων από τα δεδομένα. Γενικά, έχει οριστεί ότι ‘‘ Η ανακάλυψη γνώσης είναι η μη τετριμμένη διαδικασία αναγνώρισης έγκυρων, πρωτότυπων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων από τα δεδομένα’’. Επίσης, ένας δεύτερος ορισμός που δίνεται είναι ότι η διαδικασία ανακάλυψης γνώσης περιλαμβάνει ‘‘Το σχεδιασμό αποθηκών δεδομένων (data warehousing), τη συλλογή δεδομένων στόχου, τον

καθορισμό, την προεπεξεργασία, την μετατροπή και την ελαχιστοποίηση των δεδομένων, την εξόρυξη γνώσης, την επιλογή μοντέλου ή συνδυασμού μοντέλων, την αξιολόγηση και τελικά την ενοποίηση και χρησιμοποίηση της εξαγόμενης γνώσης'' όπως φαίνεται στο σχήμα 15.



Σχήμα 15. Η Εξόρυξη Γνώσης αποτελεί τον πυρήνα της διαδικασίας ανακάλυψης γνώσεις σε βάσεις δεδομένων

Γενικά, η ανακάλυψη γνώσης σε βάσεις δεδομένων είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων από τα δεδομένα ενώ η εξόρυξη γνώσης από τα δεδομένα είναι η χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και προτύπων που παράγονται με την διαδικασία KDD. Η διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων είναι μια επαναληπτική διαδικασία που εκτός από την εξόρυξη γνώσης, περιλαμβάνει μια μεθοδολογία για την εξαγωγή και την προετοιμασία της γνώσης,

καθώς επίσης και τη λήψη αποφάσεων σχετικών με τις ενέργειες που πρέπει να γίνουν όταν ολοκληρωθεί η εξόρυξη γνώσης.

Η διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων είναι μια διαλογική και επαναληπτική διαδικασία, δηλαδή μπορεί να απαιτηθεί η επιστροφή σε κάποιο προηγούμενο βήμα όπως φαίνεται στο παραπάνω σχήμα. Η διαδικασία KDD μπορεί να διαχωριστεί στα παρακάτω βήματα:

Ορισμός του προβλήματος (Defining the problem): Στο βήμα αυτό ορίζεται το πλαίσιο δράσης της διαδικασίας KDD, δηλαδή καθορίζονται οι προσδοκίες για τα αποτελέσματα από την εξόρυξη γνώσης που περιλαμβάνουν ουσιαστικά τις απαιτήσεις των αναλυτών, τις στρατηγικές marketing, τις προβλέψεις και την υποστήριξη αποφάσεων.

Συλλογή δεδομένων (Data Collection): Το βήμα αυτό περιλαμβάνει τον εντοπισμό των δεδομένων που είναι διαθέσιμα, την απόκτηση επιπρόσθετων δεδομένων που είναι αναγκαία για την ανάλυση και τελικά την ενσωμάτωση όλων αυτών σε ένα σύνολο δεδομένων το οποίο θα περιλαμβάνει τα χαρακτηριστικά (attributes) που θα ληφθούν υπόψη. Οι αλγόριθμοι εξόρυξης γνώσης εκπαιδεύονται και ανακαλύπτουν πρότυπα από τα δεδομένα που είναι κάθε φορά διαθέσιμα, και επομένως σε κάθε περίπτωση είναι απαραίτητη η μέγιστη δυνατή συλλογή χαρακτηριστικών.

Καθαρισμός των δεδομένων (Data Cleaning): Η αξιοπιστία των δεδομένων αποτελεί ένα πολύ σημαντικό σημείο στη διαδικασία KDD. Στο βήμα αυτό πραγματοποιεί καθαρισμός των δεδομένων, δηλαδή διαχείριση των ελλιπών τιμών (missing values) και απομάκρυνση δεδομένων με θόρυβο ή δεδομένων με ακραίες τιμές (outliers). Ο καθαρισμός των δεδομένων επιτυγχάνεται με τη χρησιμοποίηση σύνθετων στατικών μεθόδων ή αλγόριθμων εξόρυξης γνώσης (π.χ. Bayesian formula ή δέντρα απόφασης).

Μετασχηματισμός των δεδομένων (Data Transformation): Στο βήμα αυτό τα δεδομένα μετασχηματίζονται σε κατάλληλες μορφές για εξόρυξη γνώσης. Αυτό επιτυγχάνεται με την εφαρμογή μεθόδων εξομάλυνσης, κανονικοποίησης των τιμών των χαρακτηριστικών και διακριτοποίησης των συνεχών μεταβλητών καθώς κάποιοι αλγόριθμοι εξόρυξης γνώσης συμπεριφέρονται καλύτερα όταν χρησιμοποιούνται σαν διακριτά δεδομένα.

Επιλογή μεθόδου εξόρυξης γνώσης από τα δεδομένα (Data mining): Σε αυτό το βήμα εφαρμόζονται έξυπνες τεχνικές (κατηγοριοποίηση – classification, συσταδοποίηση – clustering, κανόνες συσχετίσεων – association rules στις οποίες θα γίνει εκτενή αναφορά σε επόμενες παραγράφους) εξαγωγής δυνητικά χρήσιμων προτύπων από τα δεδομένα. Οι δύο βασικοί στόχοι της εξόρυξης γνώσης είναι η περιγραφή και η πρόβλεψη. Εφόσον έχει οριστεί η στρατηγική που θα ακολουθηθεί, επιλέγεται και εκτελείται ο αλγόριθμος εξόρυξης γνώσης από τα δεδομένα. Η απόδοση και τα εξαγόμενα αποτελέσματα εξαρτώνται απ' τα προηγούμενα βήματα.

Αξιολόγηση προτύπου (Pattern Evaluation): Σε αυτό το βήμα γίνεται εκτίμηση και ερμηνεία των εξορυχθέντων προτύπων (κανόνες, συσχετισμοί, αξιοπιστία κλπ) σε σύγκριση με τους στόχους που είχα τεθεί κατά τον αρχικό ορισμό του προβλήματος, δηλαδή στο πρώτο βήμα της διαδικασίας. Ουσιαστικά, αξιολογείται η χρησιμότητα του μοντέλου και τεκμηριώνεται η γνώση που ανακαλύφθηκε η οποία είναι διαθέσιμη για περεταίρω χρήση.

Παρουσίαση της ανακαλυφθείσας γνώσης (Knowledge representation): Σε αυτό το τελευταίο βήμα η γνώση που ανακαλύφθηκε οπτικοποιείται και παρουσιάζεται στο χρήστη. Δηλαδή, χρησιμοποιούνται τεχνικές οπτικοποίησης που βοηθούν τους χρήστες στην κατανόηση και ερμηνεία των αποτελεσμάτων της εξόρυξης γνώσης και κατ' επέκταση την αποτελεσματικότητα της χρήσης της διαδικασίας KDD.

Τα παραπάνω βήματα συνδυάζονται και επαναλαμβάνονται κατά την διαδικασία KDD, καθώς μετά από κάθε αξιολόγηση η παρουσίαση της ανακαλυφθείσας γνώσης στο χρήστη, μπορεί να γίνει εκ νέου συλλογή δεομένων και σχηματισμός αυτών, εκτέλεση διαφορετικών αλγορίθμων κλπ, ώστε να εξαχθούν καλύτερα και πιο κατάλληλα αποτελέσματα.

3.2.2 Πληθώρα Αποθηκευμένης Πληροφορίας – Εξόρυξη Γνώσης από Διαφορετικούς Τύπους Δεδομένων

Τα σύγχρονα αποθηκευτικά μέσα παρέχουν τη δυνατότητα συλλογής μεγάλου όγκου δεδομένων, από απλές αριθμητικές μετρήσεις και έγγραφα κειμένου έως πιο σύνθετη πληροφορία όπως χωρικά δεδομένα, κανάλια πολυμέσων και έγγραφα υπέρ-κειμένου.

Στη συνέχεια παρουσιάζεται η ποικιλία της πληροφορίας που συλλέγεται σε ψηφιακή μορφή σε βάσεις δεδομένων και απλά αρχεία με τη μορφή λίστας η οποία βέβαια δεν είναι απαραίτητα αποκλειστική:

Επιχειρηματικές Συναλλαγές (Business transactions): Στην επιχειρηματική βιομηχανία κάθε συναλλαγή που καταγράφεται συνήθως σχετίζεται με το χρόνο και έχει διάφορες μορφές όπως αγορά, ανταλλαγή, τραπεζική συναλλαγή, απόθεμα στις αποθήκες, αποτέλεσμα οικονομικής χρήσης κλπ. Γενικά, χάρη στη διαδεδομένη χρήση του γραμμοκώδικα (bar code) εκατομμύρια συναλλαγών αποθηκεύονται καθημερινά και το πρόβλημα που ανακύπτει είναι η αποτελεσματική χρήση τους σε εύλογο χρονικό διάστημα για τη λήψη αποφάσεων.

Επιστημονικά Δεδομένα (Scientific Data): Διάφορα επιστημονικά ερευνητικά κέντρα συσσωρεύουν τεράστιο όγκο δεδομένων τα οποία χρήζουν ανάλυσης, ωστόσο το πρόβλημα είναι ότι ποιο γρήγορα αποθηκεύονται καινούργια δεδομένα από ότι μπορούν να αναλυθούν τα ήδη συγκεντρωμένα.

Ιατρικά και προσωπικά δεδομένα (Medical and personal data): Κυβερνητικοί φορείς, νοσηλευτικά ιδρύματα, εταιρείες και οργανισμοί αποθηκεύουν πολύ σημαντικές ποσότητες προσωπικών δεδομένων με σκοπό τη διαχείριση των ανθρώπινων πόρων, την καλύτερη κατανόηση της αγοράς ή απλώς την εξυπηρέτηση της πελατείας τους. Παρόλο που αυτό το είδος δεδομένων εμπίπτει σε θέματα ιδιωτικότητας και προστασίας των δεδομένων, η πληροφορία αυτή συλλέγεται, χρησιμοποιείται και μοιράζεται καθώς συσχετιζόμενοι με άλλα δεδομένα αναδεικνύει θέματα γύρω από την συμπεριφορά και τις προτιμήσεις των πελατών.

Βίντεο παρακολούθησης και φωτογραφίες (Surveillance video and pictures): Η πληροφορία αυτή που αποθηκεύεται στα μέσα βιντεοσκόπησης και ψηφιοποιείται για μελλοντική χρήση και ανάλυση.

Δορυφόροι (Satellite sensing): Υπάρχουν αμέτρητοι δορυφόροι (στατικοί ή σε τροχιά) γύρω από την υφήλιο, η οποίοι καθημερινά συλλέγουν δεδομένα και φωτογραφίες ώστε μελλοντικά να αναλυθούν.

Παιχνίδια (Games): Καθημερινά συλλέγονται τεράστιες ποσότητες δεδομένων για παιχνίδια, παίχτες και αθλητές. Τα δεδομένα αυτά χρησιμοποιούνται από προπονητές

και αθλητές για την βελτίωση των επιδόσεων και την καλύτερη κατανόηση των αντιπάλων καθώς και από δημοσιογράφους στις ανταποκρίσεις τους.

Ψηφιακά μέσα (Digital Media): Η εξέλιξη των ψηφιακών μέσω αποθήκευσης βοήθησε στην ψηφιοποίηση συλλογών εικόνας και ήχου που διαθέτουν τα μέσα ενημέρωσης με αποτέλεσμα να βελτιώσουν τη διαχείριση των οπτικοακουστικών αποθεμάτων τους.

Δεδομένα Σχεδιαστικών Συστημάτων και Δεδομένα Λογισμικού Εφαρμοσμένης Μηχανικής (CAD – Computer Assisted Systems And Software Engineering Data): Τα συστήματα αυτά παράγουν τεράστιες ποσότητες δεδομένων και ειδικά το Λογισμικό Εφαρμοσμένης Μηχανικής αποτελεί κύρια πηγή δεδομένου με κώδικα, βιβλιοθήκες συναρτήσεων, αντικείμενα κλπ., των οποίων η διαχείριση και η διατήρηση απαιτεί τη χρησιμοποίηση ισχυρών εργαλείων.

Εικονικοί κόσμοι (Virtual Wolds): Πολλές εφαρμογές χρησιμοποιούν τρισδιάστατους εικονικούς χώρους με αποτέλεσμα να υπάρχει διαθέσιμη αξιοσημείωτη ποσότητα πηγών πληροφορίας για αντικείμενα και χώρους εικονικής πραγματικότητας. Η διαχείριση τέτοιων πηγών βρίσκεται σε ερευνητικό στάδιο, ωστόσο το μέγεθος των συλλογών τέτοιας πληροφορίας συνεχίζει να αυξάνεται.

Εκθέσεις κειμένου, υπομνήματα και μηνύματα ηλεκτρονικού ταχυδρομείου (Text reports, memos and e- mail messages): Το μεγαλύτερο μέρος της επικοινωνίας εντός και μεταξύ εταιρειών ή ερευνητικών οργανισμών ακόμα και των απλών ανθρώπων μεταξύ τους, βασίζεται σε εκθέσεις κειμένου και υπομνήματα τα οποία ανταλλάσουν μέσω e-mails. Τα μηνύματα αυτά αποθηκεύονται σε ψηφιακή μορφή για μελλοντική χρήση και αναφορά και δημιουργούν ψηφιακές βιβλιοθήκες.

Οι πηγές πληροφορίας του παγκόσμιου ιστού (The World Wide Web Repositories): Με το ξεκίνημα του Παγκόσμιου Ιστού το 1993, έγγραφα ποικίλων μορφών, περιεχομένου και περιγραφής έχουν συλλεχθεί και διασυνδεθεί με υπέρ- συνδέσμους με αποτέλεσμα να αποτελούν τη μεγαλύτερη πηγή αποθηκευμένης πληροφορίας που έχει κατασκευαστεί έως τώρα. Ο Παγκόσμιος Ιστός παρά τη δυναμική και μη δομημένη φύση του, τα ετερογενή χαρακτηριστικά του, την ασυνέπεια και τον πλεονασμό, αποτελεί την πιο σημαντική συλλογή δεδομένων που χρησιμοποιείται για αναφορά, λόγω της ποικιλίας των θεμάτων που καλύπτει και τις άπειρες συνεισφορές

πηγών πληροφορίας και εκδοτών. Πολλοί πιστεύουν ότι ο Παγκόσμιος Ιστός θα αποτελέσει τη συλλογή της ανθρώπινης γνώσης.

Γενικά, τα είδη της αποθηκευμένης πληροφορίας είναι ουσιαστικά ανεξάντλητα και επομένως η εξόρυξη γνώσης θα πρέπει να είναι εφαρμόσιμη σε οποιοδήποτε είδος δεδομένων. ωστόσο, οι αλγόριθμοι μπορεί να διαφέρουν όταν εφαρμόζονται σε διαφορετικούς τύπους δεδομένων. Γενικά, η εξόρυξη γνώσης χρησιμοποιείται για τύπους δεδομένων που παρουσιάζουν σημαντικές διαφορές και αποθηκεύονται με ποικίλες μορφές όπως απλά αρχεία, σχεσιακές βάσεις δεδομένων, αποθήκες δεδομένων, μη δομημένες πηγές πληροφορίας (Ο Παγκόσμιος Ιστός), βάσεις χωρικών δεδομένων, βάσεις χρονολογικών δεδομένων κλπ. Στη συνέχεια δίνονται κάποια σχετικά παραδείγματα με μεγαλύτερη λεπτομέρεια.

Απλά αρχεία (Flat files): Τα αρχεία αυτά αποτελούν τις πιο συνηθισμένες πηγές δεδομένων για τους αλγορίθμους εξόρυξης γνώσης, κυρίως στο επίπεδο της έρευνας. Πρόκειται για απλά αρχεία κειμένου ή αρχεία δυαδικής μορφοποίησης και τα δεδομένα που περιέχουν είναι συνήθως στοιχεία συναλλαγών, χρονολογικά δεδομένα, επιστημονικές μετρήσεις κλπ.

Σχεσιακές Βάσεις Δεδομένων (Relational Data Bases): Οι βάσεις αυτές αποτελούνται από σύνολα πινάκων, με γραμμές και στήλες, που περιέχουν τις τιμές των χαρακτηριστικών των οντοτήτων ή τιμές των χαρακτηριστικών των συσχετίσεων των οντοτήτων. Η γλώσσα που χρησιμοποιείται στις σχεσιακές βάσεις δεδομένων είναι η SQL η οποία επιτρέπει την ανάκτηση και τον έλεγχο των αποθηκευμένων δεδομένων στους πίνακες. Οι αλγόριθμοι εξόρυξης γνώσης προσαρμόζονται με μεγαλύτερη ευκολία στις σχεσιακές βάσεις δεδομένων απ' ότι στα απλά αρχεία, κυρίως λόγω της κανονικοποιημένης δομής τους.

Αποθήκες Δεδομένων (Data Warehouses): Η Αποθήκη Δεδομένων είναι ουσιαστικά μια βάση δεδομένων στην οποία συλλέγονται δεδομένα από πολλές πηγές, συχνά ετερογενής, με σκοπό την ανάλυση τους ως σύνολο κάτω από το ίδιο ενοποιημένο σχήμα. Τα ετερογενή δεδομένα φορτώνονται, καθαρίζονται, μετασχηματίζονται και τελικά συγκεντρώνονται σε μια Αποθήκη με πολυδιάστατη δομή στα δεδομένα της οποίας μπορεί να εφαρμοστεί OLAP επεξεργασία. Οι OLAP κύβοι περιέχουν σύνθετα μέλη και διαστάσεις με ιεραρχίες, με αποτέλεσμα, σε αντίθεση με τις σχεσιακές βάσεις, να μην διατηρούν μεγάλη λεπτομέρεια για τα δεδομένα αυτό συμβαίνει κατά

τη διαδικασία δημιουργίας συναθροίσεων. Επομένως, είναι δύσκολη η ανακάλυψη κρυμμένης γνώσης ανάμεσα σε τέτοια μέλη και διαστάσεις. Το πρόβλημα μπορεί να ξεπεραστεί με τη χρησιμοποίηση μοντέλων εξόρυξης γνώσης πάνω σε OLAP πηγές. Στον πίνακα 2 παρουσιάζεται τι είναι εφικτό και τι όχι με το OLAP και τα μοντέλα εξόρυξης γνώσης (Data Mining Models).

Βάσεις Δεδομένων Συναλλαγών (Transaction Databases): Μια βάση συναλλαγών είναι ένα σύνολο εγγραφών – συναλλαγών, που η κάθε μία συμβαίνει σε κάποια χρονική στιγμή, έχει ένα αναγνωριστικό και ένα σύνολο αντικειμένων. Επειδή, η σχεσιακές βάσεις δεν επιτρέπουν τη δημιουργία εμφωλευμένων πινάκων (π.χ. ένα σύνολο δεδομένων να αποτελεί την τιμή ενός χαρακτηριστικού), οι συναλλαγές συνήθως αποθηκεύονται σε απλά αρχεία ή σε δύο κανονικοποιημένους πίνακες συναλλαγών, έναν για τις συναλλαγές και έναν για το σύνολο των αντικειμένων. Η τεχνική εξόρυξης που εφαρμόζεται σε τέτοια δεδομένα ονομάζεται ανάλυση δεδομένων από το « καλάθι της νοικοκυράς» (market- basket analysis) ή αλλιώς κανόνες συσχετίσεων (association rules) αφού ουσιαστικά γίνεται προσπάθεια μελέτης των κρυμμένων στοιχείων ανάμεσα στα αντικείμενα.

Βάσεις πολυμέσων (Multimedia Databases): Οι βάσεις αυτές περιέχουν βίντεο, φωτογραφίες και γενικά οπτικοακουστικά δεδομένα. Η εξόρυξη γνώσης από πηγές πολυμέσων απαιτεί τη χρησιμοποίηση τεχνολογιών αναγνώρισης προτύπων, αναγνώριση φωνής, γραφικά με υπολογιστές κλπ.

Βάσεις Χωρικών Δεδομένων (Spatial Databases): Εδώ αποθηκεύεται γεωγραφική πληροφορία, όπως χάρτες κλπ. Τα χωρικά δεδομένα είναι δεδομένα τα οποία έχουν μια χωρική συνιστώσα ή συνιστώσα θέσης. Είναι ακόμα κατά μια έννοια δεδομένα αντικειμένων που βρίσκονται σε ένα φυσικό χώρο ο οποίος δηλώνεται ρητά με ένα ή και περισσότερα γνωρίσματα θέσης, όπως η διεύθυνση ή το γεωγραφικό πλάτος / μήκος. Κλασικοί αλγόριθμοι εξόρυξης γνώσης βρίσκουν εφαρμογές και σε τέτοια δεδομένα, ωστόσο έχουν αναπτυχθεί νέες τεχνικές που προσαρμόζονται καλύτερα στις ιδιαιτερότητες των χωρικών δεδομένων.

Βάσεις χρονολογικών δεδομένων (Time –Series Databases): Οι βάσεις δεδομένων δεν περιέχουν συνήθως χρονολογικά δεδομένα καθώς τα δεδομένα τους αφορούν σε ένα συγκεκριμένο σημείο στο χώρο. Οι βάσεις χρονολογικών δεδομένων περιέχουν δεδομένα που σχετίζονται με το χρόνο και οι συνεχής ροή νέων δεδομένων προς αυτές

δημιουργεί την ανάγκη ανάλυσης σε πραγματικό χρόνο. Η εξόρυξη γνώσης σε τέτοια δεδομένα περιλαμβάνει τη μελέτη των τάσεων και των συσχετίσεων που παρουσιάζονται ενδιάμεσα της εξέλιξης των διαφορετικών μεταβλητών καθώς και την πρόβλεψη των τάσεων και της κινητικότητας των μεταβλητών σε βάθος χρόνου.

Ο Παγκόσμιος Ιστός (The World Wide Web): Αποτελεί τη μεγαλύτερη διαθέσιμη ετερογενή πηγή πληροφορίας. Ο Παγκόσμιος Ιστός συγκεντρώνει δεδομένα που προέρχονται από αναρίθμητες πηγές πληροφορίας, συγγραφείς και εκδότες ενώ παράλληλα παρέχει καθημερινά τη δυνατότητα σε εκατομμύρια χρήστες να έχουν πρόσβαση στα δεδομένα αυτά. Τα δεδομένα στο Παγκόσμιο Ιστό είναι οργανωμένα σε διασυνδεδεμένα έγγραφα διαφόρων μορφοποιήσεων, όπως κείμενα, ήχος, εικόνα, βίντεο, πρωτογενή δεδομένα ή ακόμα και εφαρμ0ογές. Ουσιαστικά, τον Παγκόσμιο Ιστό συνθέτουν τρείς συνιστώσες : το περιεχόμενο του Ιστού, δηλαδή τα διαθέσιμα έγγραφα, η δομή του Ιστού που περιλαμβάνει τους υπέρ- συνδέσμους και τις σχέσεις μεταξύ των εγγράφων και τέλος η χρήση του Ιστού, δηλαδή πως και πότε προσπελαύνονται οι διάφορες πηγές. Επομένως, η εξόρυξη γνώσης στο Παγκόσμιο Ιστό διακρίνεται σε εξόρυξη γνώσης από το περιεχόμενο του Παγκόσμιου Ιστού (web content mining), σε εξόρυξη γνώσης από τη δομή του Παγκόσμιου Ιστού (web structure mining) και σε εξόρυξη γνώσης από τη χρήση του Παγκόσμιου Ιστού (web usage mining).

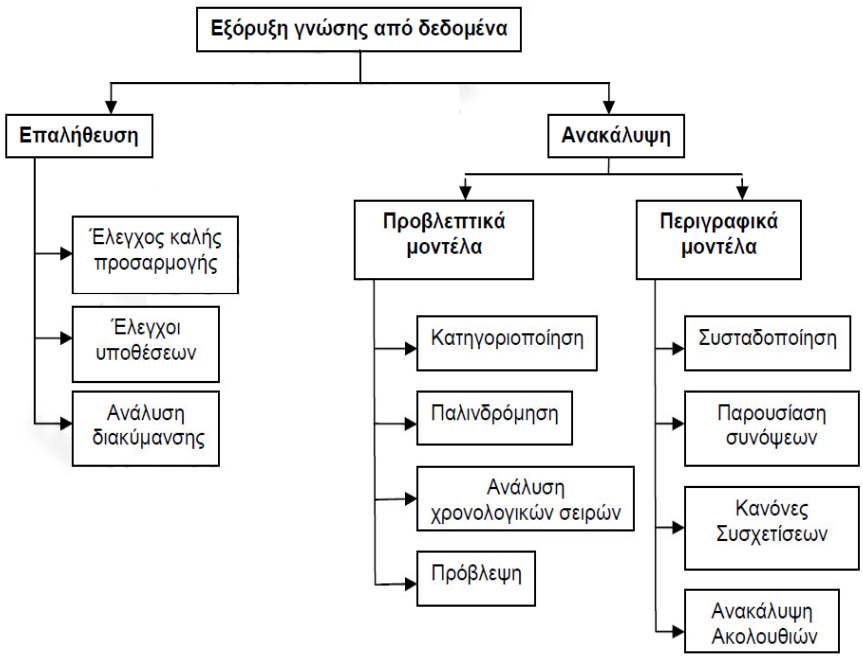
Πίνακας 2. Διαφορές OLAP και Data Mining

OLAP	**DATA MINING**
Εστιάζει σε ιστορικά δεδομένα	Εστιάζει σε μελλοντικά αποτελέσματα ή τάσεις
Συναθροίζει δεδομένα με τη χρήση προκαθορισμένης ομαδοποίησης	Προϋποθέτει λεπτομερή δεδομένα
Επαλήθευση / Αντικειμενικά αποτελέσματα	Ανακάλυψη
Ad- hoc ερωτήματα και εκθέσεις	Τεχνικές Στατιστικής και Μηχανικής Μάθησης
Περιορισμένη δυνατότητα εξαγωγής αξιόπιστων εκτιμήσεων με προβλέψεις	Μοντέλα δεδομένων διαθέσιμα για πρόβλεψη, ανακάλυψη προτύπων, εκτίμηση και παραγωγή σωστών

	αποτελεσμάτων για ανάλυση τάσεων και διενέργεια προβλέψεων
Το OLAP μπορεί να χρησιμοποιηθεί ως πηγή δεδομένων για τα μοντέλα εξόρυξης γνώσης	Τα αποτελέσματα των μοντέλων εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν σε OLAP εφαρμογές ως νέες μεταβλητές προς πρόβλεψη ή ως χαρακτηριστικά

3.2.3 Γενική Αναφορά στις Μεθόδους Εξόρυξης Γνώσης από Δεδομένα

Ο τεράστιος όγκος δεδομένων που αποθηκεύονται σε ποικίλες μορφές οδήγησε και στην ανάπτυξη πολλών μεθόδων εξόρυξης γνώσης με διαφορετικούς σκοπούς και στόχους. Στο παρακάτω σχήμα παρουσιάζεται μια γενική ταξινόμηση των μεθόδων εξόρυξης γνώσης από τα δεδομένα.



Σχήμα16. Ταξινόμηση των μεθόδων εξόρυξης γνώσης

Όπως φαίνεται στο παραπάνω σχήμα, υπάρχουν δύο βασικοί τύποι εξόρυξης γνώσης: η ***επαλήθευση*** και η ***ανακάλυψη***. Κατά την επαλήθευση το σύστημα καλείται να

επιβεβαιώσει τις υποθέσεις που έχει κάνει ο χρήστης ενώ κατά την ανακάλυψη το σύστημα καλείται να βρει νέους κανόνες και πρότυπα μέσα από αυτόνομες διαδικασίες. Οι μέθοδοι επαλήθευσης ασχολούνται κυρίως με την εκτίμηση μιας υπόθεσης που προτείνεται από μια εξωτερική πηγή. Πρόκειται ουσιαστικά για τις παραδοσιακές μεθόδους της Στατιστικής (π.χ. έλεγχος καλής προσαρμογής, έλεγχος υποθέσεων, ανάλυση διακύμανσης) οι οποίες σχετίζονται λιγότερο με την εξόρυξη από δεδομένα συγκριτικά με τις μεθόδους ανακάλυψης. Οι μέθοδοι ανακάλυψης είναι αυτές που εντοπίζουν αυτόματα πρότυπα στα δεδομένα.

Το επόμενο στάδιο της ανακάλυψης χωρίζεται στη δημιουργία δύο μοντέλων του ***προβλεπτικού*** και του ***περιγραφικού.***

Το προβλεπτικό μοντέλο (predictive model) έχει ως στόχο την πρόγνωση της τιμής μιας μεταβλητής μέσα από τις τιμές άλλων μεταβλητών που είναι γνωστές. Δηλαδή επιδιώκει να κάνει κάποια πρόβλεψη για τις τιμές των δεδομένων με χρησιμοποίηση γνωστών αποτελεσμάτων που έχει βρει από άλλα δεδομένα. Ως εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούνται για τη μοντελοποίηση μιας πρόβλεψης αναφέρονται η κατηγοριοποίηση, η παλινδρόμηση, η ανάλυση χρονολογικών σειρών και η πρόβλεψη.

Το περιγραφικό μοντέλο (descriptive model) έχει ως στόχο την περιγραφή όλου του συνόλου δεδομένων ή της διαδικασίας που παράγει τα δεδομένα αναγνωρίζοντας πρότυπα ή συσχετίσεις ανάμεσα στα δεδομένα. Το περιγραφικό μοντέλο, σε αντίθεση με το προβλεπτικό, στοχεύει στην ερμηνεία των δεδομένων ερευνώντας τις ιδιότητές τους, τις συσχετίσεις που υπάρχουν ανάμεσά τους, χωρίς να προβλέπει νέες ιδιότητες. Η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών είναι περιγραφικές εργασίες εξόρυξης γνώσης από δεδομένα.

Η εξόρυξη γνώσης μπορεί να χρησιμοποιηθεί για την επίλυση εκατοντάδων επιχειρησιακών προβλημάτων. Στη συνέχεις ακολουθεί μια συνοπτική αναφορά στις εργασίες εξόρυξης γνώσης των δύο μοντέλων ανακάλυψης βάσει του σχήματος 16.

Κατηγοριοποίηση (classification): Περιλαμβάνει την κατασκευή ενός μοντέλου που απεικονίζει ένα στοιχείο σε μια από ένα σύνολο από προκαθορισμένες κατηγορίες - - κλάσεις (classes). Αναφέρεται και ως εποπτευόμενη μάθηση, επειδή οι κατηγορίες καθορίζονται πριν εξεταστούν τα δεδομένα. Στους αλγορίθμους κατηγοριοποίησης οι κατηγορίες πρέπει να ορίζονται με βάση τις τιμές των γνωρισμάτων των δεδομένων

(δεδομένα εκπαίδευσης) και τα νέα δεδομένα να κατηγοριοποιούνται με βάση τη γνώση που παρέχουν τα δεδομένα εκπαίδευσης.

Παλινδρόμηση (regression): Ένας από τους κύριους σκοπούς της προσαρμογής καμπυλών είναι η εκτίμηση της εξαρτημένης μεταβλητής από την ανεξάρτητη μεταβλητή. Η μέθοδος ή η διαδικασία εκτίμησης ονομάζεται παλινδρόμηση. Ουσιαστικά, χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή υπό πρόβλεψη. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συναρτήσεων (π.χ. γραμμική, λογαριθμική κλπ) και καθορίζει τη συνάρτηση που μοντελοποιεί καλύτερα τα δεδομένα που έχουν δοθεί. Η κύρια διαφορά της παλινδρόμησης με την κατηγοριοποίηση είναι ότι το υπό πρόβλεψη χαρακτηριστικό παίρνει συνεχείς τιμές.

Ανάλυση χρονολογικών σειρών ή χρονοσειρών (time series analysis): Μελετά την τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο με κάποια περιοδικότητα (π.χ. ημερήσια, εβδομαδιαία, ,μηνιαία κλπ). Ως χρονολογική σειρά ορίζεται η ακολουθία των τιμών μιας μεταβλητής οι οποίες λαμβάνονται σε προκαθορισμένα χρονικά σημεία που συνήθως ισαπέχουν ή αναφέρονται σε διαδοχικές περιόδους ίδιας διάρκειας. Η γραφική απεικόνιση των χρονολογικών σειρών, οι οποίες εκφράζονται σε απόλυτα ή σε σχετικά μεγέθη, γίνεται βάσει ειδικών διαγραμμάτων, τα λεγόμενα χρονοδιαγράμματα. Υπάρχουν τρεις βασικές λειτουργίες που χρησιμοποιούνται στην ανάλυση χρονοσειρών. Η πρώτη περιλαμβάνει τη χρησιμοποίηση μονάδων μέτρησης απόστασης ώστε να καθοριστούν οι ομοιότητες ανάμεσα σε διαφορετικές χρονοσειρές. Η δεύτερη λειτουργία εξετάζει τη δομή της χρονοσειράς για να κατηγοριοποιήσει τη συμπεριφορά της. Τέλος, η Τρίτη λειτουργία χρησιμοποιεί διαγράμματα χρονοσειρών για την πρόβλεψη μελλοντικών τιμών.

Πρόβλεψη (prediction): Αναφέρεται στην πρόβλεψη ελλιπών αριθμητικών τιμών ή στην αύξηση/μείωση τάσεων που σχετίζονται με δεδομένα ως προς τα χρόνο. Η πρόβλεψη μπορεί να θεωρηθεί ως κατηγοριοποίηση καθώς και η κυρίαρχη ιδέα είναι η χρησιμοποίηση μεγάλου αριθμού δεδομένων προηγούμενων περιόδων για απόδοση πιθανών μελλοντικών τιμών στα δεδομένα. Δηλαδή οι εφαρμογές πρόβλεψης επιδιώκουν την απόδοση μιας τιμής σε μια μελλοντική κατάσταση παρά σε μια τρέχουσα.

Συσταδοποίηση (clustering): Είναι παρόμοια με την κατηγοριοποίηση, δηλαδή επιχειρεί να βρει ομάδες – συστάδες (clusters) παρατηρήσεων που είναι κοντά μεταξύ τους ως προς τα γνωρίσματα των χαρακτηριστικών που περιλαμβάνουν. Η κύρια διαφορά της με την κατηγοριοποίηση είναι ότι οι συστάδες δεν είναι προκαθορισμένες αλλά ορίζονται από τα ίδια τα δεδομένα. Αναφέρεται και ως μη εποπτευόμενη μάθηση καθώς η κλάση στην οποία ανήκουν τα δεδομένα εκπαίδευσης δεν είναι εκ των προτέρων γνωστή. Η βασική αρχή που διέπει όλες τις προσεγγίσεις συσταδοποίησης βασίζεται στη μεγιστοποίηση της ομοιότητας μεταξύ αντικειμένων που ανήκουν στην ίδια ομάδα (intra-class similarity) και στην ελαχιστοποίηση της ομοιότητας μεταξύ αντικειμένων διαφορετικών ομάδων (inter-class similarity).

Παρουσίαση συνόψεων (summarization) ή Χαρακτηρισμός (characterization): Απεικονίζει τα δεδομένα σε υποσύνολά τους με συνοδευτικές απλές περιγραφές και χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων. εξάγει αντιπροσωπευτικές πληροφορίες για τη βάση δεδομένων και παράγει τους ονομαζόμενους ***χαρακτηριστικούς κανόνες (characteristics rules).*** Καθώς ένας κύβος δεδομένων περιέχει συγκεντρωμένα δεδομένα, οι απλές λειτουργίες OLAP ταιριάζουν στο σκοπό της παρουσίασης συνόψεων.

Κανόνες συσχετίσεων (association rules): Ένα κανόνας συσχετίσεων είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ διαφορετικών χαρακτηριστικών σε ένα σύνολο δεδομένων. οι κανόνες συσχετίσεων βρίσκουν εφαρμογή στην ανάλυση του «καλαθιού αγοράς» (market basket analysis) καθώς σκοπός τους είναι η αναγνώριση προϊόντων που συνήθως αγοράζονται μαζί.

Ανακάλυψη ακολουθιών (sequence discovery): Χρησιμοποιείται για τον καθορισμό προτύπων σε σειριακά δεδομένα. Σειρές διακριτών τιμών ή καταστάσεων γνωρισμάτων δεδομένων συνθέτουν μια ακολουθία. Τα δεδομένα τόσο στην ανακάλυψη ακολουθιών όσο και στην ανάλυση χρονολογικών σειρών περιέχουν γειτονικές παρατηρήσεις που αλληλοεξαρτώνται, με τη μόνη διαφορά ότι στην πρώτη περίπτωση τα δεδομένα είναι διακριτά ενώ στη δεύτερη είναι συνεχή. Επίσης, η διαφορά της ανακάλυψης ακολουθιών με τους κανόνες συσχετίσεων έγκειται στο γεγονός ότι τα μοντέλα ακολουθίας θεωρούν ότι τα προϊόντα αγοράζονται με κάποια σειρά ενώ τα μοντέλα συσχετίσεων θεωρούν ότι κάθε προϊόν έχει την ίδια πιθανότητα να αγοραστεί και δεν εξαρτάται από τις άλλες αγορές.

ΚΕΦΑΛΑΙΟ 4: ΑΝΑΛΥΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ (OLAP ΑΝΑΛΥΣΗ)

Σκοπός της παρούσας ενότητας είναι η παρουσίαση των εξαγόμενων αποτελεσμάτων και συμπερασμάτων που προήλθαν από την αναλυτική επεξεργασία των δεδομένων της Αποθήκης Μεταναστευτικών Δεδομένων στο περιβάλλον του Business Intelligence Management Studio του Microsoft SQL Server 2005 Analysis Services. Ωστόσο, λόγω της ιδιαιτερότητας που παρουσιάζουν τα διαθέσιμα δεδομένα των δύο ασφαλιστικών οργανισμών επιλέχθηκε μια πιο ευέλικτη μορφή (data source) της αρχικής Αποθήκης Δεδομένων με κατάλληλη προπαρασκευή των δεδομένων. Επομένως, το πολυδιάστατο μοντέλο δεδομένων-κύβος που δημιουργήθηκε για την OLAP ανάλυση, με τις αντίστοιχες διαστάσεις, ιεραρχίες και μέτρα, βασίστηκε τόσο στους αρχικούς πίνακες του σχεσιακού σχήματος όσο και σε βοηθητικούς πίνακες-όψεις (views) που χρησιμοποιήθηκαν με στόχο την εξαγωγή καλύτερων αποτελεσμάτων.

4.1 Η χρήση της Microsoft Access

Η Microsoft Access είναι ένα από τα πιο δημοφιλή προγράμματα διαχείρισης βάσεων δεδομένων που κυκλοφορούν στην αγορά. Η μεγάλη διάδοσή της τα τελευταία χρόνια, οφείλεται στην απλότητα και ευκολία στη χρήση της , καθώς και στη δυναμικότητά της να δημιουργεί εφαρμογές διαχείρισης βάσεων δεδομένων σε σχετικά μικρό χρονικό διάστημα.

Η Microsoft Access έχει όλα τα χαρακτηριστικά ενός κλασσικού συστήματος διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) και αρκετά παραπάνω. Δεν είναι μόνο ένα πανίσχυρο, ευέλικτο και εύκολο στη χρήση RDBMS, αλλά και ένα πλήρες εργαλείο ανάπτυξης εφαρμογών για βάσεις δεδομένων.

Μπορούμε να χρησιμοποιήσουμε την Access για να κατασκευάσουμε και να εκτελέσουμε μια εφαρμογή φτιαγμένη για τα δικά μας μέτρα. Μπορούμε να περιορίζουμε, να επιλέγουμε και να προσθέτουμε τα δεδομένα σας με τη χρήση ερωτημάτων (Queries). Μπορούμε να δημιουργούμε φόρμες (Forms) για την εμφάνιση και την τροποποίηση των δεδομένων μας.

Μπορούμε επίσης να δημιουργήσουμε απλές ή πολύπλοκες αναφορές (Reports). Οι φόρμες και οι αναφορές αναφέρονται σε κάποιον πίνακα ή σε κάποιο ερώτημα και έτσι «κληρονομούν» τις ιδιότητες του πίνακα ή του ερωτήματος αντίστοιχα. Αυτό σημαίνει ότι οι μορφές (Formats) και οι κανόνες εγκυρότητας (Validation Rules) χρειάζεται να οριστούν μόνο μία φορά κατά τον σχεδιασμό ενός πίνακα.

Ανάμεσα στα ισχυρότερα χαρακτηριστικά της Access είναι και οι Οδηγοί (Wizards), τους οποίους μπορούμε να χρησιμοποιούμε για την κατασκευή πινάκων και ερωτημάτων και για τον ορισμό μιας μεγάλης ποικιλίας φορμών και αναφορών. Οι Οδηγοί αυτοί είναι έτοιμες σχεδιάσεις που έχει ενσωματωμένες η Access για να μας βοηθάει στη σχεδίαση της βάσης δεδομένων.

Η Access περιλαμβάνει ακόμη την περιεκτική γλώσσα προγραμματισμού Microsoft Visual Basic for Applications (VBA), που μπορούμε να χρησιμοποιήσουμε για να κατασκευάσουμε πολύ ισχυρές εφαρμογές.

4.1.1 Σαν Προσωπικό RDBMS

Η Access είναι ένα θαυμάσιο εργαλείο για τη διαχείριση προσωπικών στοιχείων στον δικό μας προσωπικό υπολογιστή. Θα μπορούσαμε να δημιουργήσουμε έναν κατάλογο με τις διευθύνσεις, τις ημερομηνίες γενεθλίων και τις επετείους των φίλων μας. Αν σας αρέσει το μαγείρεμα, θα σας ήταν χρήσιμη μια βάση δεδομένων για συνταγές. Ίσως ακόμα να θέλετε να παρακολουθείτε τις συλλογές των video ταινιών ή των βιβλίων σας.

Φανταστείτε στη βάση δεδομένων που κάνετε για τους φίλους σας, να μπορείτε να κρατάτε και από μια φωτογραφία για τον καθένα καθώς και από ένα αγαπημένο τους

μουσικό κομμάτι. Ή σε μια βάση δεδομένων για τα CD's που έχετε, να μπορείτε να καταχωρείτε για κάθε CD την εικόνα του και από ένα ακουστικό δείγμα του. Οι δυνατότητες της Access είναι απεριόριστες.

4.1.2 Χαρακτηριστικά των Windows

Η Access χρησιμοποιεί όλα τα γνωστά μας εύχρηστα χαρακτηριστικά των Windows, όπως τα πολλά παράθυρα, τα μενού, τις γραμμές εργαλείων και τους πτυσσόμενους καταλόγους. Μπορεί ακόμα να επικοινωνεί και να ανταλλάσσει δεδομένα (κείμενα, λογιστικά φύλλα, γραφήματα, σκίτσα, εικόνες και ήχους) με τ' άλλα προγράμματα των Windows.

Η Access χρησιμοποιεί τη Διασύνδεση Πολλών Εγγράφων (MDI) των Windows 95 για να μας επιτρέπει την ταυτόχρονη εργασία με πολλά διαφορετικά αντικείμενα. Δηλαδή θα μπορούμε να δουλεύουμε με πολλούς πίνακες, φόρμες, αναφορές, μακροεντολές ή υπομονάδες την ίδια στιγμή.

4.1.3. Η Αρχιτεκτονική της Microsoft Access

Η Access θεωρεί οτιδήποτε μπορεί να έχει όνομα σαν αντικείμενο (object). Τα βασικά αντικείμενα μιας βάσης δεδομένων της Access είναι οι πίνακες (tables), τα ερωτήματα (queries), οι φόρμες (forms), οι αναφορές (reports), οι μακροεντολές (macros) και οι υπομονάδες (modules).

Σε παλιότερα προγράμματα διαχείρισης βάσεων δεδομένων (όπως ήταν η dBase III+, η dBase IV, κ.α.), με τον όρο βάση δεδομένων εννοούσαμε μόνο τα αρχεία στα οποία

αποθηκεύαμε δεδομένα και η σύνδεση των αρχείων μεταξύ τους ήταν πολύ δύσκολη ή και αδύνατη.

Στην Access, ο όρος βάση δεδομένων περιλαμβάνει και όλα τα βασικά αντικείμενα που συσχετίζονται με τα αποθηκευμένα δεδομένα, καθώς και τα αντικείμενα που ορίζουμε για την αυτοματοποίηση της χρήσης των δεδομένων μας.

Ακολουθεί μια σύντομη, αλλά και περιεκτική περιγραφή των βασικών αντικειμένων μιας βάσης δεδομένων της Access.

Πίνακας (Table)

Είναι ένα αντικείμενο που ορίζουμε και το χρησιμοποιούμε για την αποθήκευση των δεδομένων μας. Κάθε πίνακας περιέχει πληροφορίες για ένα συγκεκριμένο θέμα, όπως είναι οι πελάτες, οι παραγγελίες τους, οι μαθητές κ.ά.

Οι πίνακες περιέχουν πεδία (fields) ή στήλες (columns), όπου αποθηκεύονται τα διαφορετικά είδη πληροφοριών, όπως είναι το όνομα ενός πελάτη ή ο βαθμός ενός μαθητή και εγγραφές (records) ή γραμμές (rows) που περιέχουν όλες τις πληροφορίες για μια συγκεκριμένη περίπτωση του πίνακα, όπως π.χ. όλες οι πληροφορίες για έναν μαθητή που ονομάζεται Αντωνιάδης.

Σε κάθε πίνακα μπορούμε να ορίσουμε ένα βασικό ή πρωτεύον κλειδί (primary key), που είναι ένα ή περισσότερα πεδία που χαρακτηρίζουν μοναδικά την εγγραφή μέσα στον πίνακα και ένα ή περισσότερα ευρετήρια (indexes) για να μπορούμε να αυξήσουμε την ταχύτητα πρόσβασης στα δεδομένα μας. Το πρωτεύον κλειδί μπορεί να είναι ο κωδικός ενός πελάτη, το ΑΦΜ ενός φορολογούμενου, ο αριθμός μητρώου ενός μαθητή, η πινακίδα ενός αυτοκινήτου κ.ά.

Σ' έναν πίνακα μπορούμε να έχουμε ένα μόνο πρωτεύον κλειδί και, αν θέλουμε, ένα ή περισσότερα ευρετήρια. Για παράδειγμα, στον πίνακα με τα στοιχεία των πελατών, πρωτεύον κλειδί μπορεί να είναι ο κωδικός του πελάτη, αλλά μόνο αυτός, και σαν ευρετήρια μπορούμε να ορίσουμε όποια πεδία θέλουμε. Τα ευρετήρια είναι χρήσιμα μόνο για γρήγορη αναζήτηση όταν ο πίνακάς μας έχει πολλές και μεγάλες εγγραφές.

Ερώτημα (Query)

Είναι ένα αντικείμενο που «απομονώνει» ότι στοιχεία θέλουμε και μας δίνει μια συγκεκριμένη άποψη των δεδομένων μας, η οποία άποψη μπορεί να προέρχεται από

έναν ή περισσότερους πίνακες. Μπορούμε να ορίσουμε ερωτήματα για να δημιουργήσουμε νέους πίνακες από τα δεδομένα ενός ή περισσοτέρων ήδη υπαρχόντων πινάκων.

Για παράδειγμα, μπορούμε να ορίσουμε ένα ερώτημα που θα παίρνει δεδομένα από τους πίνακες πελατών και παραγγελιών και θα δημιουργεί έναν νέο πίνακα, που θα αναφέρεται όμως σαν ερώτημα, όπου θα περιέχονται τα στοιχεία επώνυμο, όνομα και ποσότητα παραγγελίας από τους πελάτες που έκαναν παραγγελίες ενός συγκεκριμένου προϊόντος τον περασμένο μή να. Τα στοιχεία που δημιουργεί αυτό το ερώτημα μπορούμε μετά να τα επεξεργαστούμε σαν έναν νέο πίνακα.

Φόρμα (Form)

Είναι ένα αντικείμενο που χρησιμεύει κατά κύριο λόγο για την εισαγωγή και την εμφάνιση των δεδομένων μας ή για τον έλεγχο της εκτέλεσης της εφαρμογής. Μπορούμε να χρησιμοποιούμε φόρμες για να έχουμε μια ωραία παρουσίαση των δεδομένων που προέρχονται από ερωτήματα ή/και πίνακες. Οι φόρμες μπορούν ακόμα να εκτελούν μακροεντολές ή διαδικασίες της γλώσσας VBA σαν απόκριση σε κάποια συμβάντα, για παράδειγμα να υπολογίζεται η αξία του ΦΠΑ σε μια παραγγελία.

Αναφορά (Report)

Είναι ένα αντικείμενο σχεδιασμένο για τη μορφοποίηση, την εκτέλεση υπολογισμών, την εκτύπωση και τη σύνοψη κάποιων επιλεγμένων δεδομένων. Πριν τυπώσουμε μια αναφορά, μπορούμε να τη δούμε στην οθόνη μας (preview ή προεπισκόπιση).

Μακροεντολή (Macro)

Είναι ένα αντικείμενο που αποτελεί το δομημένο ορισμό μιας ή περισσοτέρων ενεργειών που θέλουμε να εκτελέσει η Access σαν απόκριση σ' ένα ορισμένο συμβάν. Για παράδειγμα, μπορούμε να σχεδιάσουμε μια μακροεντολή που θα ανοίγει μια δεύτερη φόρμα, σαν απόκριση στην επιλογή ενός στοιχείου της κύριας φόρμας.

Μπορούμε επίσης να έχουμε μια μακροεντολή που θα ελέγχει την εγκυρότητα των στοιχείων ενός πεδίου όταν θα κάνουμε αλλαγές σ' αυτά. Μπορούμε ακόμα να περιλάβουμε συνθήκες στις μακροεντολές για να ορίζουμε πότε πρέπει να εκτελεστούν κάποιες ενέργειες των μακροεντολών και πότε όχι.

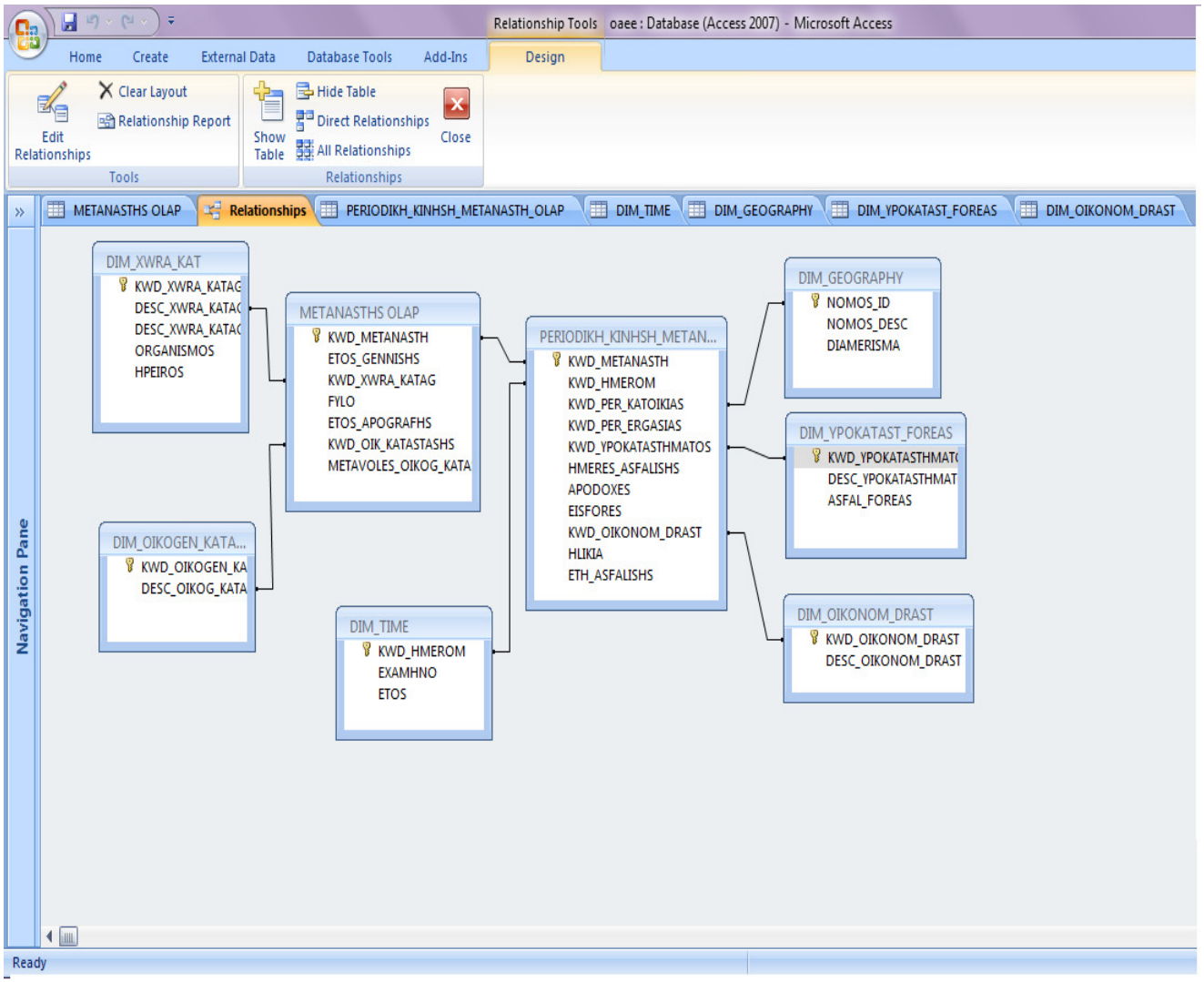
Μπορούμε να χρησιμοποιήσουμε μακροεντολές για το άνοιγμα και την εκτέλεση ερωτημάτων, για το άνοιγμα πινάκων ή για την εκτύπωση ή την εμφάνιση αναφορών. Ακόμα, μπορούμε μέσα από μια μακροεντολή, να εκτελούμε άλλες μακροεντολές ή διαδικασίες της VBA.

Υπομονάδα (Module)

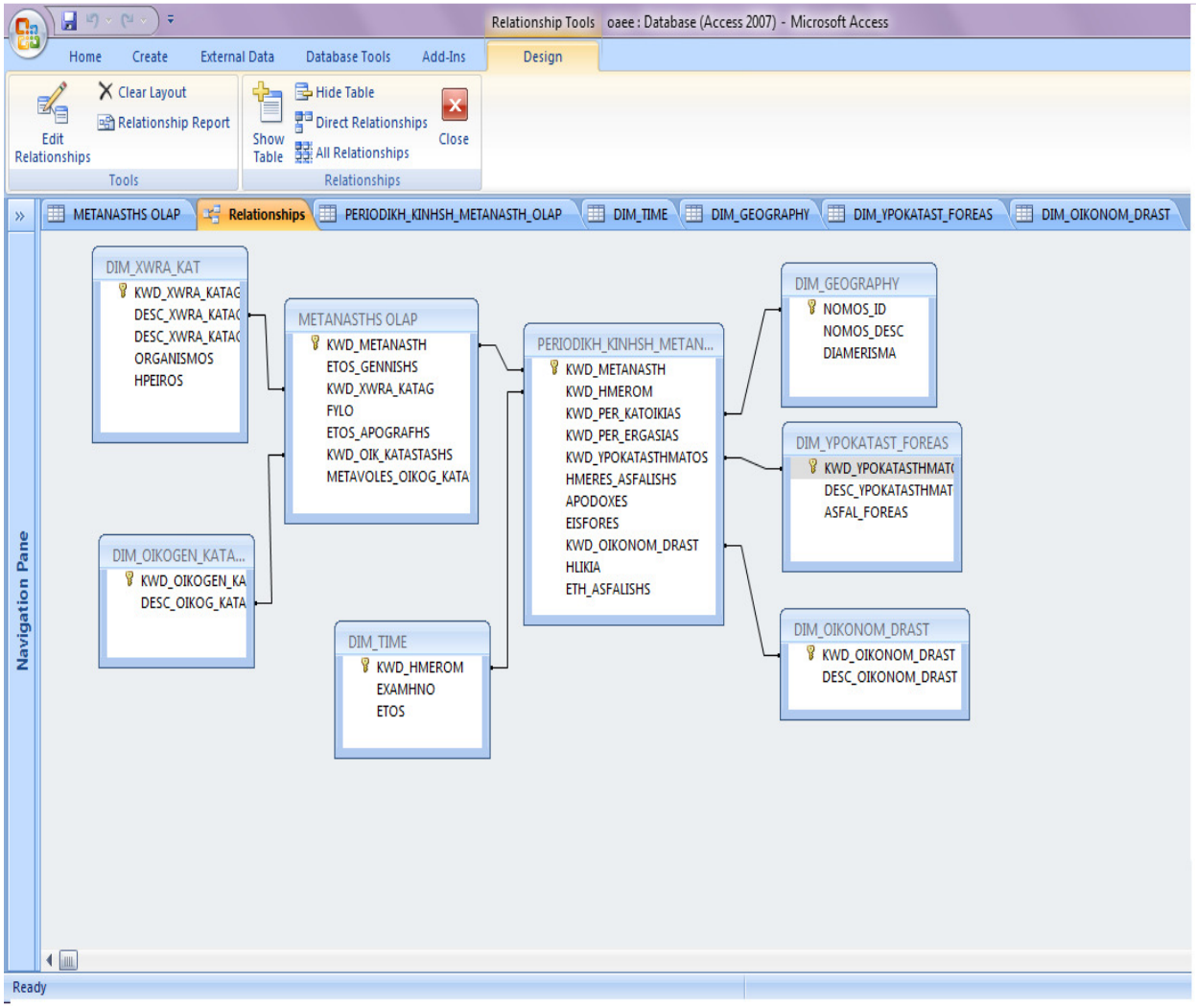
Είναι ένα αντικείμενο που περιέχει διαδικασίες (procedures ή functions) που τις ορίσαμε εμείς με τη χρήση της VBA. Οι υπομονάδες παρέχουν μια πιο διακριτική ροή των ενεργειών και μας επιτρέπουν να παγιδεύουμε τα λάθη, κάτι που δεν μπορούμε να κάνουμε με τις μακροεντολές. Μπορεί να είναι αυτόνομα αντικείμενα, με συναρτήσεις που μπορούν να κληθούν από οποιοδήποτε σημείο της εφαρμογής μας, ή μπορεί να συσχετίζονται απευθείας με τις φόρμες ή τις αναφορές για να αποκρίνονται μόνο στα συμβάντα των φορμών και των αναφορών.

4.2 Προπαρασκευή της Αποθήκης Δεδομένων για την OLAP Ανάλυση

Για την πραγματοποίηση OLAP Ανάλυσης σε κύβο με διαστάσεις, ιεραρχίες και μέτρα, δημιουργήθηκε μια πιο ευέλικτη μορφή πηγής δεδομένων (data sources) από αυτή της αρχικής Αποθήκης Μεταναστευτικών Δεδομένων, της οποίας η σχεσιακή αναπαράσταση η οποία δίνεται στο παρακάτω σχήμα 17 και στο σχήμα 18 δίνεται ο κύβος που δημιουργήθηκε πάνω στο data source της Αποθήκης του σχήματος 17.



Σχήμα 17. Η Αποθήκη δεδομένων IKA_OAEE_DW.ds, που θα χρησιμοποιηθεί ως data source στον κύβο IKA_OAEE_DW.cube



Σχήμα 18. Ο κύβος ΙΚΑ_ΟΑΕΕ_DW.cube

Το σχήμα 17 της Αποθήκης δεδομένων διατηρεί την ίδια δομή με το σχεσιακό σχήμα της αποθήκης δεδομένων είναι δηλαδή Σχήμα Χιονονιφάδας (Snowflake Schema). Η κύρια διαφορά του σχήματος 17 με το αρχικό σχεσιακό σχήμα είναι οι βοηθητικοί πίνακες PERIODIKH_KINHSH_METANASTH_OLAP και METANASTHS_OLAP που αντικατέστησαν τους πίνακες FACT_PERIODIKH_KINHSH και

DIM_METANASTHS αντίστοιχα. Η αντικατάσταση πραγματοποιήθηκε για τη δημιουργία επιπλέον πεδίων, χρήσιμων για τους σκοπούς της OLAP ανάλυσης.

Ο πίνακας γεγονότων PERIODIKH_KINHSH_METANASTH_OLAP αποτελεί μια όψη (view) που διατηρεί όλα τα πεδία του πίνακα FACT_PERIODIKH_KINHSH, ενώ παράλληλα διαθέτει δύο επιπλέον πεδία – μέτρα, τα [ETH_ASFALISHS]. Στα πεδία αυτά, τόσο η ηλικία όσο και τα έτη ασφάλισης, δηλαδή ο χρόνος ασφάλισης κάθε ασφαλισμένου στον αντίστοιχο ασφαλιστικό φορέα, υπολογίζονται για τις χρονικές περιόδους που καλύπτει η Αποθήκη Δεδομένων, δηλαδή για την περίοδο Ιούνιος 2004 έως Ιούνιος 2008 για τους ασφαλισμένους του ΙΚΑ_ΕΤΑΜ και για την περίοδο Ιούνιος 2007 έως Δεκέμβριος 2009 για τους ασφαλισμένους του ΟΑΕΕ. Το SQL Script δημιουργίας του view PERIODIKH_KINHSH_METANASTH_OLAP δίνεται παρακάτω.

Ο πίνακας METANASTHS-OLAP, που στον κύβο χρησιμοποιείται ως πίνακας διάστασης αλλά και ως δευτερεύον πίνακας γεγονότων, αποτελεί επίσης μια όψη (view) που διατηρεί όλα τα πεδία του πίνακα DIM_METANASTHS,ενώ διαθέτει ένα επιπλέον πεδίο το [METAVOLES_OIK_DRAST]. Το πεδίο αυτό δημιουργήθηκε από την ανάγκη για καταγραφή πληροφορίας που σχετίζεται με τη μεταβολή της οικονομικής δραστηριότητας των ασφαλισμένων ως προς τις χρονικές περιόδους που εξετάζονται. Εξ' ορισμού, αλλά και λόγω της έλλειψης διαθέσιμων στοιχείων, για τους ασφαλισμένους του ΟΑΕΕ θεωρήθηκε ότι διατηρούν σταθερή την οικονομική τους δραστηριότητα, καθώς απασχολούνται ως ελεύθεροι επαγγελματίες. Όσον αφορά στους ασφαλισμένους του ΙΚΑ-ΕΤΑΜ, για τους οποίους παράχθηκαν στοιχεία οικονομικής δραστηριότητας για την περίοδο Ιούνιος 2004 έως Ιούνιος 2008 σύμφωνα με τα επίσημα δημοσιευμένα εξαμηνιαία στατιστικά δελτία του ΙΚΑ, το συγκεκριμένο πεδίο καταγράφει πληροφορία σχετικά με το αν κάποιοι ασφαλισμένοι διατηρούν την οικονομική τους δραστηριότητα για ορισμένες χρονικές περιόδους ή αλλάζουν συνεχώς. Με αυτό τον τρόπο κατέστη δυνατή η παρακολούθηση της κινητικότητας των ασφαλισμένων του ΙΚΑ ανάμεσα στις διάφορες κατηγορίες οικονομικής δραστηριότητας, καθώς απασχολούνται με εξαρτημένη εργασία και κατά τη διάρκεια της ασφαλιστικής τους ιστορίας μπορούν να εναλλάσσουν οικονομικές δραστηριότητες και κατ' επέκταση εργοδότες. Στην περίπτωση που υπήρχαν διαθέσιμα στοιχεία μητρώου εργοδοτών θα ήταν πιο εύκολη η εξαγωγή

συμπερασμάτων σχετικά με την εργασιακή κινητικότητα των ασφαλισμένων του ΙΚΑ. Το SQL Script δημιουργίας του view METANASTHS_OLAP δίνεται παρακάτω:

SQL Scripts δημιουργίας views

```
/*Create views for OLAP analysis*/

CREATE VIEW [dbo].[PERIODIKH_KINHSH_METANASTH_OLAP]

AS

SELECT dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_KATOIKIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_ERGASIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_YPOKATASTHMATOS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM,
       dbo.FACT_PERIODIKH_KINHSH.HMERES_ASFALISHS,
       dbo.FACT_PERIODIKH_KINHSH.APODOXES,
       dbo.FACT_PERIODIKH_KINHSH.EISFORES,
       dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST,

       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_GENNHSHS AS HLIKIA,
       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_APOGRAFHS AS
           ETH_ASFALISHS

FROM   dbo.FACT_PERIODIKH_KINHSH INNER JOIN
       dbo.DIM_METANASTHS ON
       dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH

       dbo.DIM_METANASTHS.KWD_METANASTH

       INNER JOIN dbo.DIM_TIME ON

       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM = dbo.DIM_TIME.KWD_HMEROM
```

```
CREATE VIEW [dbo].[METANASTHS_OLAP]

AS
SELECT KWD_METANASTH, ETOS_GENNHSHS, KWD_XWRA_KATAG, FYLO,

       ETOS_APOGRAFHS, KWD_OIKOG_KATAST,
       CASE CHANGE WHEN 0 THEN 'STABLE OAEE'
       WHEN 1 THEN 'STABLE IKA' ELSE 'IKA CHANGED' END

       AS METAVOLES_OIK_DRAST
FROM

      (SELECT dbo.DIM_METANASTHS.KWD_METANASTH,
       dbo.DIM_METANASTHS.ETOS_GENNHSHS,
       dbo.DIM_METANASTHS.KWD_XWRA_KATAG,

       dbo.DIM_METANASTHS.FYLO, dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
       dbo.DIM_METANASTHS.KWD_OIKOG_KATAST,

       COUNT(DISTINCT dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST)
       AS CHANGE

  FROM dbo.DIM_METANASTHS INNER JOIN
       dbo.FACT_PERIODIKH_KINHSH ON
       dbo.DIM_METANASTHS.KWD_METANASTH =

       dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH
  GROUP BY dbo.DIM_METANASTHS.KWD_METANASTH,

                dbo.DIM_METANASTHS.ETOS_GENNHSHS,
                dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
                dbo.DIM_METANASTHS.FYLO,

                dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
                                dbo.DIM_METANASTHS.KWD_OIKOG_KATAST) AS A


/*Create views and named queries for Data mining models*/
```

```
CREATE VIEW [dbo].[PERIODIKH_KINHSH_METANASTH_DM]

AS
SELECT dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_KATOIKIAS,

       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_ERGASIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_YPOKATASTHMATOS,

       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM,
       dbo.FACT_PERIODIKH_KINHSH.HMERES_ASFALISHS,
       dbo.FACT_PERIODIKH_KINHSH.APODOXES,

       dbo.FACT_PERIODIKH_KINHSH.EISFORES,
  ISNULL(dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST, 0) AS

       KWD_OIKONOM_DRAST,
            dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_GENNHSHS AS HLIKIA,
       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_APOGRAFHS AS

           ETH_ASFALISHS
FROM   dbo.FACT_PERIODIKH_KINHSH INNER JOIN

            dbo.DIM_METANASTHS ON dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH =
       dbo.DIM_METANASTHS.KWD_METANASTH
       INNER JOIN dbo.DIM_TIME ON

       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM = dbo.DIM_TIME.KWD_HMEROM
```

```
CREATE VIEW [dbo].[METANASTHS_DM]

AS
SELECT A.KWD_METANASTH, A.ETOS_GENNHSHS, B.DESC_XWRA_KATAG,
       B.DESC_XWRA_KATAG_ENG, B.HPEIROS, A.FYLO, A.ETOS_APOGRAFHS,
   ISNULL (C.DESC_OIKOG_KATAST, 'ΑΓΝΣΤΟ') AS OIKOG_KATAST,

   CASE CHANGE WHEN 0 THEN 'STABLE OAEE' WHEN 1 THEN 'STABLE IKA' ELSE
   'IKA CHANGED' END

   AS METAVOLES_OIK_DRAST FROM

    (SELECT dbo.DIM_METANASTHS.KWD_METANASTH,
            dbo.DIM_METANASTHS.ETOS_GENNHSHS,
            dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
            dbo.DIM_METANASTHS.FYLO,
            dbo.DIM_METANASTHS.ETOS_APOGRAFHS,

        ISNULL(dbo.DIM_METANASTHS.KWD_OIKOG_KATAST, 0) AS
            KWD_OIKOG_KATAST,

        COUNT(DISTINCT dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST) AS
              CHANGE

        FROM  dbo.DIM_METANASTHS INNER JOIN
              dbo.FACT_PERIODIKH_KINHSH ON
              dbo.DIM_METANASTHS.KWD_METANASTH =
              dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH

        GROUP BY dbo.DIM_METANASTHS.KWD_METANASTH,
                 dbo.DIM_METANASTHS.ETOS_GENNHSHS,
                 dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
                 dbo.DIM_METANASTHS.FYLO,
                 dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
                 dbo.DIM_METANASTHS.KWD_OIKOG_KATAST)
      AS A INNER JOIN

      dbo.DIM_XWRA_KAT AS B ON A.KWD_XWRA_KATAG = B.KWD_XWRA_KATAG FULL
      OUTER JOIN
      dbo.DIM_OIKOGEN_KATAST AS C ON A.KWD_ OIKOG_KATAST =
          C.KWD_OIKOG_KATAST
```

```
CREATE VIEW [dbo].[OIKONOM_DRAST_DM]

AS

SELECT [KWD_OIKONOM_DRAST]
      ,[DESC_OIKONOM_DRAST]

  FROM [IKA_OAEE_DW].[dbo].[DIM_OIKONOM_DRAST]
UNION ALL
SELECT 0 AS Expr1,'ΑΓΝΣΤΗ' AS Expr2

CREATE VIEW [dbo].[PLITHOS_ANA_OIKOG_KATAST] AS
SELECT * FROM (

SELECT A.OIKOG_KATAST,(PLITHOS*100/OL_PLITHOS)* PLITHOS_AGN/100 AS OLA FROM
(SELECT OIKOG_KATAST,COUNT(*) AS PLITHOS FROM METANASTHS_DM

WHERE OIKOG_KATAST < >'ΑΓΝΣΤΟ' GROUP BY OIKOG_KATAST ) AS
A ,

(SELECT COUNT(*) AS OL_PLITHOS FROM METANASTHS_DM WHERE OIKOG_KATAST <
>'ΑΓΝΣΤΟ') AS B ,

(SELECT COUNT(*) AS PLITHOS_AGN FROM METANASTHS_DM WHERE
OIKOG_KATAST='ΑΓΝΣΤΟ') AS C ) AS D




CREATE NAMED QUERY [dbo].[METANASTHS_DM_ADVANCED]

AS

SELECT            [KWD_METANASTH]
      ,[ETOS_GENNHSHS]
```

```
      ,[DESC_XWRA_KATAG]
      ,[DESC_XWRA_KATAG_ENG]
      ,[HPEIROS]
      ,[FYLO] ,[ETOS_APOGRAFHS]
      ,[OIKOG_KATAST]
      ,[METAVOLES_OIK_DRAST],


CASE WHEN ROW_NUMBER() OVER (ORDER BY CASE WHEN OIKOG_KATAST='ΑΓΝΣΤΟ' THEN
0 ELSE 1 END,KWD_METANASTH)<34706 THEN 'ΕΓΓΑΜΟΣ'
WHEN ROW_NUMBER() OVER (ORDER BY CASE WHEN OIKOG_KATAST='ΑΓΝΣΤΟ'

THEN 0 ELSE 1 END,KWD_METANASTH) BETWEEN 34706 AND 62864 THEN 'ΑΓΑΜΟΣ'
WHEN ROW_NUMBER() OVER (ORDER BY CASE WHEN OIKOG_KATAST='ΑΓΝΣΤΟ' THEN 0
ELSE 1 END,KWD_METANASTH) BETWEEN 62864 AND 64173 THEN 'ΔΙΑΖΕΥΓΜΕΝΟΣ'

ELSE OIKOG_KATAST END AS OIKOG_KATAST_2
FROM [IKA_TEBE].[dbo].[METANASTHS_DM]
```

Το σχήμα 18 απεικονίζει τον κύβο που δημιουργήθηκε πάνω στο data sources του σχήματος 17, ο οποίος αποτελείται από δύο πίνακες γεγονότων (PERIODIKH_KINHSH_METANASTH_OLAP και METANASTHS_OLAP) με τα αντίστοιχα μέτρα και τις διαστάσεις τους. Στον κύβο έχουν διατηρηθεί και κάποια πεδία της αρχικής Αποθήκης Δεδομένων με κενές τιμές που ωστόσο δεν επηρεάζουν σημαντικά την OLAP ανάλυση. Στη συνέχεια παραθέτονται τρεις πίνακες οι οποίοι συνοψίζουν τις διαστάσεις, ιεραρχίες και τα μέτρα του κύβου IKA_OAEE_DW.cube. Ο πίνακας (3) παρουσιάζει όλες τις διαστάσεις του κύβου με τις αντίστοιχες ιεραρχίες τους. Στον πίνακα (4) περιγράφονται οι ομάδες μέτρων (measure groups), δηλαδή τα μετρήσιμα μεγέθη του κύβου, που δημιουργούνται και επιλέγονται με τον Cube Wizard. Τέλος, στον πίνακα (5) παρουσιάζονται τα επιπλέον μέτρα που ορίστηκαν ως υπολογιζόμενα μέλη (Calculated Members) πάνω στον συγκεκριμένο κύβο, ορισμένα από τα οποία υπολογίστηκαν με βάση τα

μέτρα του πίνακα (4). Τα Calculated Members υπολογίστηκαν με MDX Scripts τα οποία δίνονται παρακάτω:

MDX Scripts δημιουργίας μέτρων (Calculated Members) στον κύβο IKA_OAEE_DW.cube

[ASFALISMENOI OAEE]

```
sum([DIM YPOKATAST FOREAS].[ASFAL  FOREAS].&[OAEE],
                    [Measures].[UNIQUE METANASTHS])
```

[ASFALISMENOI IKA]

```
sum([DIM YPOKATAST FOREAS].[ASFAL FOREAS].&[IKA ETAM],
                        [Measures].[UNIQUE METANASTHS])
```

[MO HLIKIAS] // Υπολογισμός μέσου όρου ηλικίας όλων των ασφαλισμένων

```
[Measures].[SUM HLIKIA]/[Measures].[PERIODIKH KINHSH METANASTH OLAP
                                    Count]
```

[MO EISFORWN]// Υπολογισμός μέσου όρου εισφορών ανά έτος για τους ασφαλισμένους του ΟΑΕΕ

```
SUM([DIM TIME].[ETOS].CURRENTMEMBER,[Measures].[EISFORES])/SUM([DIM
    TIME].[ETOS].CURRENTMEMBER,[Measures].[ASFALISMENOI OAEE])
```

[POSOSTO METANASTWN ANA OIKON DRAST]

```
([Measures].[ASFALISMENOI IKA],[DIM OIKONOM DRAST].[DESC OIKONOM
DRAST].CURRENTMEMBER)/([Measures].[ASFALISMENOI IKA],[DIM OIKONOM
DRAST].[DESC OIKONOM DRAST].[ALL])
```

[POSOSTO METANASTWN ANA ETHNIKOTHTA]

```
([Measures].[METANASTHS OLAP Count],[DIM XWRA KAT].[DESC XWRA
KATAG].CurrentMember)/([Measures].[METANASTHS OLAP Count],[DIM XWRA
KAT].[DESC XWRA KATAG].[ALL])
```

Πίνακας 3. Σύνοψη των διαστάσεων και ιεραρχιών του κύβου ΙΚΑ_ΟΑΕΕ_DW.cube

Διάσταση	**Ιεραρχία**	**Περιγραφή**
DIM_TIME	ETOS → EXAMHNO	Διάσταση Χρόνου
DIM_OIKONOM_DRAST	Δεν έχει οριστεί Ιεραρχία	Διάσταση Οικονομικής Δραστηριότητας των ασφαλισμένων
DIM_YPOKATAST_FOREAS	AFALISTIKO TAMEIO → ASFAL FOREAS	Διάσταη Ασφαλιστικού Φορέα (ΙΚΑ-ΕΤΑΜ, ΟΑΕΕ)
DIM_YPOKATAST_FOREAS	YPOKASTHMA → DESC YPOKATASTHMATOS	Διάσταση Περιγραφής Υποκαταστημάτων Ασφαλιστικού Φορέα(ΙΚΑ-ΕΤΑΜ,ΟΑΕΕ)
METANASTHS_OLAP	Δεν έχει οριστέι Ιεραρχία	Διάσταση με δημογραφικά στοιχεία του μετανάστη ασφαλισμένου
DIM_XWRA_KAT	PROELEYSH METANASTH → ORGANISMOS → DESC_XWRA_KATG DESC_XWRA_KATAG_ENG	Διάσταση χώρας καταγωγής των μεταναστών ασφαλισμένων
DIM_OIKOGEN_KATAST	MARITAL STATUS DESC_OIKOG_KATAST	Διάσταση οικογενειακής κατάστασης
DIM_PER_KATOIKIAS	TOPOTHESIA → DIAMERISMA → NOMOS DESC	Διάσταση γεωγραφικής περιοχής κατοικίας του ασφαλισμένου
DIM_PER_ERGASIAS	TOPOTHESIA → DIAMERISMA → NOMOS DESC	Διάσταση γεωγραφικής περιοχής εργασίας του ασφαλισμενου
HLIKIAKES OMADES	AGE_GROUPS	Διάσταση ηλικιακών

	HLIKIA (διακρίνεται στα διαστήματα) (17-24), (25-29), (30-32), (33-36), (37-40), (41-47) και (48-70)	ομάδων.Τα αντίστοιχα διαστήματα προήλθαν με επιλογή από τα Properties (Discretization Method – Equal Areas)
PALAIOTHTA	XRONOS_ASFALISHS ⟶ ETH_ASFALISHS (διακρίνονται στα διαστήματα) (0-3, (4-7), (8-12), (13-16), (17-21), (22-26) και (27-46)	Διάσταση παλαιότητας, δηλαδή χρόνου ασφάλισης κάθε ασφαλισμένου στον αντίστοιχο ασφαλιστικό φορέα. Τα αντίστοιχα διαστήματα προήλθαν με επιλογή από τα Properties (Discretization Method-Clusters)
FYLO	GENDER ⟶ FYLO (Α,Γ)	Διάσταση φύλο ασφαλισμένου
METAVOLH ERGASIAS	Δεν έχει οριστεί Ιεραρχία	Διάσταση μεταβολή εργασίας

Πίνακας 4. Ομάδες μέτρων (measure groups) του κύβου ΙΚΑ_ΟΑΕΕ_DW.cube

Measure Groups	Μετρήσιμα Μεγέθη-Πεδία που δημιουργήθηκαν με τον Cube Wizard	Περιγραφη
PERIODIKH_KINHSH_ METANASTH_OLAP	PERIODIKH KINHSH METANASTH OLAP Count	Πλήθος Μεταναστών του συγκεκριμένου πίνακα.
	APODOXES	Ποσά αποδοχών ασφαλισμένων (κενό πεδίο).
	EISFORES	Ποσά εισφορών ασφαλισμένων του ΟΑΕΕ.
	HLIKIA	Η ηλικία του ασφαλισμένου για κάθε χρονική περίοδο που εξετάζεται.
	ETH ASFALISHS	Τα έτη ασφάλισης κάθε ασφαλισμένου για κάθε χρονική περίοδο που εξετάζεται.
	SUM HLIKIA	Συνολική ηλικία ασφαλισμένων, ως βοηθητικό μέτρο υπολογισμού του μέσου όρου ηλικίας.
METANASTHS_OLAP	ETOS GENNHSHS	Έτος γέννησης ασφαλισμένων.
	ETOS APOGRAFHS	Έτος απγραφής ασφαλισμένων.
	METANASTHS OLAP Count	Πλήθος Μεταναστών του συγκεκριμένου πίνακα.
PLITHOS_METANASTWN	UNIQUE METANASTHS	Εύρεση του διακριτού πήθους των μεταναστών (με count distinct). Βοηθητικό μέτρο.

Πινακας 5. Calculated Members πάνω στον κύβο ΙΚΑ_ΟΑΕΕ_DW.cube

Ορισμός μετρήσιμων μεγεθών σε όλο τον κύβο ως Calculated Members	Περιγραφή
ASFALISMENOI OAEE	Πλήθος ασφαλισμένων μεταναστών του ΟΑΕΕ. Βοηθητικό Μέτρο.
ASFALISMENOI IKA	Πλήθος ασφαλισμένων του ΙΚΑ-ΕΤΑΜ. Βοηθητικό Μέτρο.
MO HLIKIAS	Μέσος όρος ηλικίας του συνόλου των ασφαλισμένων.
MO EISFORWN	Μέσος όρος εισφορών ανά χρονική περίοδο των ασφαλισμένων του ΟΑΕΕ.
POSOSTO METANASTWN ANA OIKON DRAST	Ποσοστό ασφαλισμένων μεταναστών του ΙΚΑ-ΕΤΑΜ ανά οικονομική δραστηριότητα.
POSOSTO METANASTWN ANA ETHNIKOTHTA	Ποσοστό του συνόλου των ασφαλισμένων μεταναστών ανά εθνικότητα.

4.3 Παραδείγματα OLAP Ανάλυσης στον Κύβο

Στην προηγούμενη ενότητα ορίστηκε ο κύβος δεδομένων με τις διαστάσεις, τις ιεραρχίες και τα μετρήσιμα μεγέθη του. Στη συνέχεια, δλινονται με τη μορφή screenshot κάποια ενδεικτικά παραδείγματα επεξεργασίας των δεδομένων του κύβου στο Analysis Servicees του Microsoft SQL Server 2005 με εφαρμογή διαφόρων λειρουργιών OLAP για την εξαγωγήχρήσιμων συμπερασμάτων. Δηλαδή, σε κάθε ενδεικτικό παράδειγμα θα εκτελούνται κάποιες από τις επόμενες και θ ασχολιάζονται τα σχετικά αποτελέσματα.

Συσσώρευση (Roll-up): Πρόκειται για μια λειτουργία με την οποία εκτελείται ένα βήμα ανόδου στην ιεραρχία μιας διάστασης (π.χ. από ημέρα σε μήνα). Ο κύβος που προκύπτει από τη λειτουργία της συνάθροισης της πληροφορίας περιέχει πιο ομαδοποιημένα δεδομένα, με βάση τη διάσταση στην οποία έγινε η ομαδοποίηση. Η ανάβαση στην ιεραρχία μπορεί συνεχιστεί με όμοιο τρόπο. Γενικα, η λειτουργία αυτή παρέχει συγκεντρωτικά αποτελέσματα τα οποία μπορούν να χρησιμοποιηθούν για την εξαγωγή στατιστικών στοιχείων για τα δεδομένα που αποθηκεύονται στην Αποθήκη.

Εμβάθυνση (Drill-down): Είναι η αντίστροφη πράξη του roll-up, όπου εκτελείται ένα βήμα καθόδου από ένα υψηλότερο επίπεδο της ιεραρχίας μιας διάστασης σε ένα χαμηλότερο. Με την εφαρμογή της συγκεκριμένης λειτουργίας, ο κύβος επιστρέφει αποτελέσματα με μεγάλο βαθμό λεπτομέρειας. Επίσης, η λειτουργία αυτή παρέχει τη δυνατότητα στον αναλυτή να διατρέξει ακόμη και ολόκληρη την ιεραρχία μιας κλάσης δεδομένων και να φτάσει στο χαμηλότερο επίπεδο λεπτομέρειας.

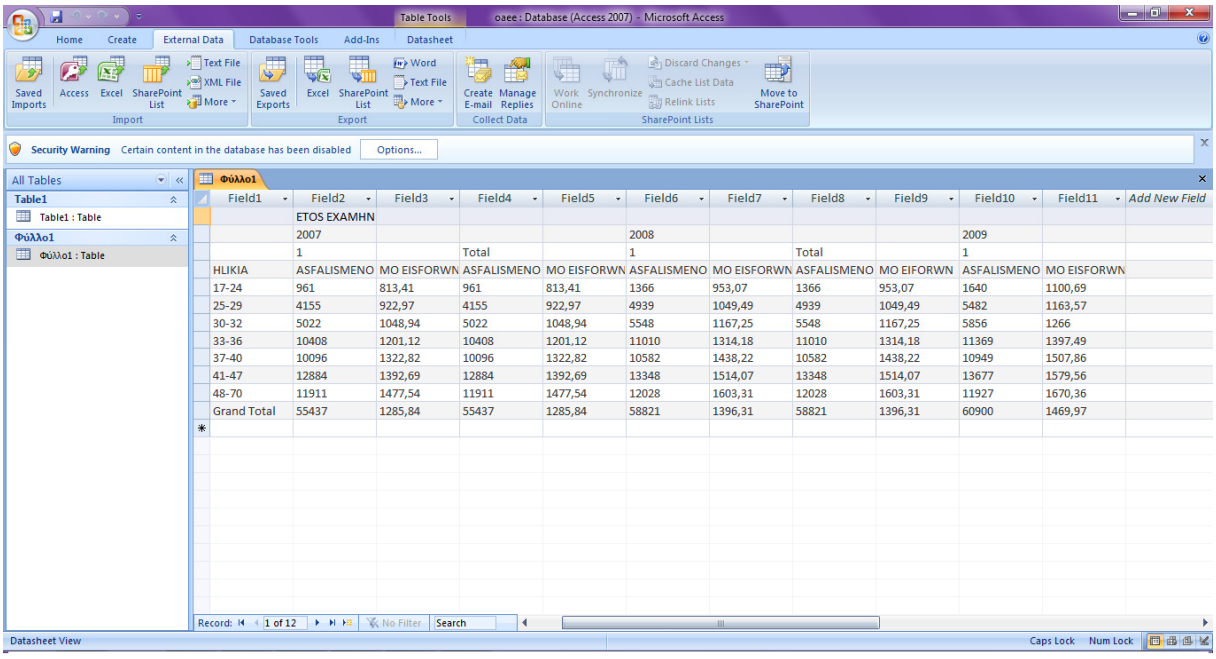
Τεμαχισμός (Slice): Πρόκειται για λειτουργία επιλογής δεδομένων σε μια συγκεκριμένη διάσταση. Ένα επίπεδο (slice) είναι ένα υποσύνολο ενός υπερκύβου σύμφωνα με μία περιοχή τιμών ή μια συγκεκριμένη τιμή ενός επιπέδου διάστασης (οριζόντιος τεμαχισμός). Ουσιαστικά, η συγκεκριμένη λειτουργία φιλτράρει τα αποτελέσματα που δίνει ο κύβος ως προς μία διάστασή του.

Κομμάτιασμα (Dice): Πρόκειται για λειτουργία επιλογής δεδομένων από δύο ή και περισσότερες διαστάσεις (κάθετος τεμαχισμός). Ουσιαστικά, η συγκεκριμένη λειτουργία φιλτράρει περισσότερες από μια διαστάσεις του κύβου για την εξαγωγή των επιθυμητών αποτελεσμάτων.

Περιστροφή (Pivot): Πρόκειται για λειτουργία αλλαγής της διάταξης των διαστάσεων ώστε να διευκολυνθεί η ανάλυση. Κατά την περιστροφή, δεν μεταβάλλονται ούτε μειώνονται τα δεδομένα του υπερκύβου, απλά αλλάζει ο τρόπος παρουσίασής τους στην εφαρμογή ανάλυσης.

Παράδειγμα 1:

Μας ενδιαφέρει να μελετήσουμε την εξέλιξη μέσου όρου των εξαμηνιαίων εισφορών των ασφαλισμένων του ΟΑΕΕ ανά ηλικιακή ομάδα και για την χρονική περίοδο τριών εξαμήνων, Ιούνιος 2007 έως Ιούνιους 2009.



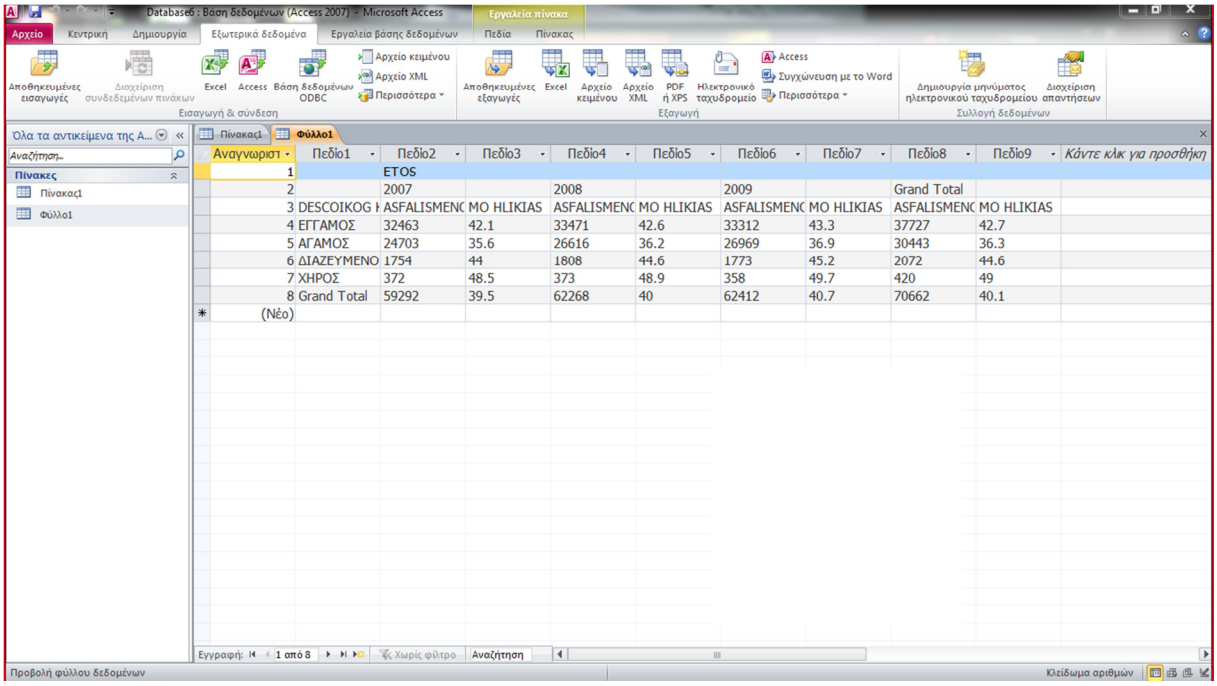
Field1	Field2	Field3	Field4	Field5	Field6	Field7	Field8	Field9	Field10	Field11
	ETOS EXAMHN									
	2007				2008				2009	
	1		Total		1		Total		1	
HLIKIA	ASFALISMENO	MO EISFORWN	ASFALISMENO	MO EISFORWN	ASFALISMENO	MO EISFORWN	ASFALISMENO	MO EIFORWN	ASFALISMENO	MO EISFORWN
17-24	961	813,41	961	813,41	1366	953,07	1366	953,07	1640	1100,69
25-29	4155	922,97	4155	922,97	4939	1049,49	4939	1049,49	5482	1163,57
30-32	5022	1048,94	5022	1048,94	5548	1167,25	5548	1167,25	5856	1266
33-36	10408	1201,12	10408	1201,12	11010	1314,18	11010	1314,18	11369	1397,49
37-40	10096	1322,82	10096	1322,82	10582	1438,22	10582	1438,22	10949	1507,86
41-47	12884	1392,69	12884	1392,69	13348	1514,07	13348	1514,07	13677	1579,56
48-70	11911	1477,54	11911	1477,54	12028	1603,31	12028	1603,31	11927	1670,36
Grand Total	55437	1285,84	55437	1285,84	58821	1396,31	58821	1396,31	60900	1469,97

Εικόνα 1. Εξαγόμενα αποτελέσματα 1ου παραδείγματος

Όπως ήταν αναμενόμενο, στις μικρές ηλικίες οι καταβληθείσες εξαμηνιαίες εισφορές είναι μικρότερες συγκριτικά με αυτές στις μεγαλύτερες ηλικιακές κλάσεις. Επίσης, οι εισφορές ακολουθούν αυξητική πορεία ως προς το χρόνο. Τέλος, παρατηρούμε ότι δε κάθε εξάμηνο το σύνολο των ασφαλισμένων μεταβάλλεται χωρίς ποτέ να φτάνει τον πραγματικό αριθμό ασφαλισμένων του ΟΑΕΕ που εξετάζουμε, δηλαδή τους 70662 ασφαλισμένους. Αυτό συμβαίνει, επειδή τα διαθέσιμα οικονομικά στοιχεία για την τριετία 2007-2009 αντιπροσοπεύουν τις αξαμηνιαίες εισφορές των τακτικών ασφαλισμένων, δηλαδή αυτών που δεν χρωστού κάποια δόση.

Παράδειγμα 2:

Μας ενδιαφέρει να μελετήσουμε τους ασφαλισμένους του ΟΑΕΕ ως πρός την οικογενειακή τους κατάσταση και το μέσο όρο ηλικίας τους για την τριετία 2007 εως 2009, κατά την οποία υπάρχουν σχετικά διαθέσιμα στοιχεία :



Αναγνωριστ	Πεδίο1	Πεδίο2	Πεδίο3	Πεδίο4	Πεδίο5	Πεδίο6	Πεδίο7	Πεδίο8	Πεδίο9
1		ΕΤΟΣ							
2		2007		2008		2009		Grand Total	
3	DESCOIKOG Η	ASFALISMENC	MO HLIKIAS	ASFALISMENC	MO HLIKIAS	ASFALISMENC	MO HLIKIAS	ASFALISMENC	MO HLIKIAS
4	ΕΓΓΑΜΟΣ	32463	42.1	33471	42.6	33312	43.3	37727	42.7
5	ΑΓΑΜΟΣ	24703	35.6	26616	36.2	26969	36.9	30443	36.3
6	ΔΙΑΖΕΥΜΕΝΟ	1754	44	1808	44.6	1773	45.2	2072	44.6
7	ΧΗΡΟΣ	372	48.5	373	48.9	358	49.7	420	49
8	Grand Total	59292	39.5	62268	40	62412	40.7	70662	40.1
(Νέο)									

Εικόνα 2. Εξαγόμενα αποτελέσματα 2^{ου} παραδείγματος.

Όπως παρατηρούμε, κατά την τριετία 2007 έως 2009 οι περισσότεροι ασφαλισμένοι του ΟΑΕΕ, 37.727 ασφαλισμένοι επί του συνόλου των 70.662 είναι έγγαμοι με μέσο όρο ηλικίας τα 42,7 έτη.

Συμπέρασμα

Τα τελευταία χρόνια, η εξέλιξη των υπολογιστικών συστημάτων και των τεχνολογιών αποθήκευσης δεδομένων σε συνδιασμό με την διάδοση των συστημάτων διαχείρισης βάσεων δεδομένων οδήγησαν στη συγκέντρωση και αποθήκευση τεράστιου όγκου πληροφορίας κάθε είδους. Η σύγχρονη κοινωνία, αναφέρεται συχνά ως κοινωνία της πληροφορίας καθώς για κάθε είδος πληροφορίας που παράγεται υπάρχει η δυνατότητα καταγραφής του και άρα αποθήκευσής του, με αποτέλεσμα να διογκώνεται το μέγεθος των αντίστοιχων βάσεων δεδομένων. Ωστόσο, η συνολική διαχείριση και αξιοποίηση του τεράστιου όγκου της διαθέσιμης πληροφορίας αποτελεί ένα σημαντικό και δύσκολο έργο το οποίο μπορεί να αντιμετωπιστεί αποτελεσματικά μόνο με την χρήση σύγχρονων τεχνολογιών πληροφορικής για την ανάπτυξη προηγμένων συστημάτων διαχείρισης βάσεων δεδομένων και υποστήριξης λήψης αποφάσεων.

Η λήψη αποφάσεων είναι μια από τις πιο σημαντικές αρμοδιότητες ενός σύγχρονου διοικητικού στελέχους. Τα επιτεύγματα της επιστήμης της πληροφορικής, αποτελούν σήμερα ένα σύγχρονο τεχνολογικό υπόβαθρο για την ανάπτυξη εξελιγμένων συστημάτων υποστήριξης λήψης αποφάσεων. Τα αρχιτεκτονήματα δεδομένων (Data Warehousing) και η Online ανάλυση δεδομένων (OLAP), είναι μέρος των εφαρμογών στήριξης αποφάσεων και αποτελούν μία σύγχρονη προσέγγιση στο πρόβλημα της υποστήριξης αποφάσεων. Η εφαρμογή ενός τέτοιου μοντέλου, παρέχει την δυνατότητα άμεσης ανάλυσης των δεδομένων, υποστηρίζει τις διαδικασίες λήψης απόφασης και όλα αυτά μέσα από ενέργειες απαλλαγμένες από τεχνικά θέματα.

Βιβλιογραφία

Elnasri R., Navathes S. (1996), Θεμελιώδεις Αρχές Συστημάτων Βάσεων Δεδομένων, Δίαυλος, Αθήνα.

Viescas J (1997)., Ο Οδηγός της Microsoft για την Access, Κλειδάριθμος, Αθήνα.

Dunham M. H., "Data Mining", Εκδόσεις Νέων Τεχνολογιών, Αθήνα 2004.

DATE C. J., "Εισαγωγή στα Συστήματα Βάσεων Δεδεομένων", Εκδόσεις Κλειδάριθμος, 6η Έκδοση τόμος Α.

DATE C. J., "Εισαγωγή στα Συστήματα Βάσεων Δεδεομένων", Εκδόσεις Κλειδάριθμος, 6η Έκδοση τόμος Β.

Κάππος Θ. Γιάννης (2003). Δουλέψτε με τις βάσεις δεδομένων και την *access*, Θεσσαλονίκη: Εκδόσεις Τζιόλας.