ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Οι τεχνολογίες OLAP και Data warehousing

Του φοιτητή:
Δαραβίγκα Δημήτριου
Αρ. Μητρώου: 05/2933

Επιβλέπων Καθηγητής
κ. Δέρβος Δημήτριος

Θεσσαλονίκη 2012

Πρόλογος

Η πτυχιακή αυτή εργασία με θέμα «Οι τεχνολογίες OLAP και Data Warehousing» εκπονήθηκε στο πλαίσιο των σπουδών μου στο Τμήμα Πληροφορικής του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος της Θεσσαλονίκης.

Στην πτυχιακή αυτή υπάρχει μια επισκόπηση των τεχνολογιών OLAP και Data Warehousing. Πρόκειται για τεχνολογίες που αναπτύχθηκαν την τελευταία εικοσαετία, εξαιτίας των διαρκώς αυξανόμενων αναγκών των επιχειρήσεων για τεχνολογίες διαχείρισης βάσεων δεδομένων μεγάλου όγκου. Οι νέες τεχνολογίες θα χειρίζονταν τις βάσεις δεδομένων, με τρόπο τέτοιο, ώστε να προκύπτουν χρήσιμα συμπεράσματα για την υποστήριξη αποφάσεων στις επιχειρήσεις.

Περίληψη

Η πτυχιακή παρουσιάζει τις τεχνολογίες OLAP και Data Warehouse, καθώς και κάποια παραδείγματα στο εμπορικό πρόγραμμα IBM Infosphere αλλά και στο open source πρόγραμμα saiku.

Η αποθήκη δεδομένων είναι μια συγκεντρωτική αποθήκη, όπου διατηρούνται όλες οι πληροφορίες για την ανάλυση, σε έναν οργανισμό. Τα δεδομένα προέρχονται από διάφορες πηγές με σκοπό την αναλυτική επεξεργασία και τη δημιουργία αναφορών.

Η OLAP είναι μία μεθοδολογία, η οποία υποστηρίζει την ανάλυση δεδομένων σε πολυδιάστατο περιβάλλον. Τα δεδομένα που προέρχονται από μία αποθήκη δεδομένων, οπτικοποιούνται σε κύβους και μέσω των πράξεων OLAP, προβάλλονται τα κομμάτια του κύβου που χρειάζεται ο χρήστης.

Abstract

This thesis presents OLAP and Data Warehousing technologies, as well as examples of these technologies with the commercial IBM Infosphere program and the open source program saiku.

The data warehouse is a centralized repository where all information, which an organization uses, is retained for analysis. The data is gathered from various sources to analytical processing and reporting.

OLAP is a methodology that supports data analysis in multidimensional environment. The data gathered from a data warehouse is visualized in cubes and with OLAP operations, the needed part of the cube could be viewed.

Ευχαριστίες

Πρωτίστως, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Δημήτριο Δέρβο, που μου έδωσε τη δυνατότητα να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο.
Στη συνέχεια, θα ήθελα να ευχαριστήσω την οικογένεια μου και τους φίλους μου για τη στήριξη και τη βοήθεια τους όχι μόνο κατά τη διάρκεια εκπόνησης της πτυχιακής, αλλά και σε όλη τη διάρκεια της φοιτητικής μου ζωής.

Πίνακας περιεχομένων

Πρόλογος 0
Περίληψη 2
Abstract 3
Ευχαριστίες 4
Ευρετήριο εικόνων 7
Ευρετήριο πινάκων 7
Εισαγωγή 8
Κεφάλαιο 1: Αποθήκες Δεδομένων (Data Warehouse) 10
Εισαγωγή 10
1.1 Κίνητρα ανάπτυξης αποθηκών δεδομένων 11
1.2 Ορισμός 14
1.3 Επιχειρησιακή Ευφυΐα κι Αποθήκες Δεδομένων 15
1.4 Οι στόχοι μιας αποθήκης δεδομένων 16
1.5 Εισαγωγή δεδομένων στην αποθήκη 18
1.6 Κατασκευή της αποθήκης δεδομένων 21
1.7 Σχήμα αστέρα 22
1.8 Πρόσθετα σχεσιακά σχήματα 25
1.9 Υποστήριξη αποφάσεων: Ανάλυση των δεδομένων αποθήκης 27
1.10 Διαφορές OLTP – Αποθήκης Δεδομένων 28
Επίλογος 31
Κεφάλαιο 2: On-Line Analytical Processing (OLAP) 32
Εισαγωγή 32
2.1 Γενικά 33
2.2 FASMI test 34
2.3 Ο κύβος OLAP 35
2.4 Πράξεις OLAP 37
2.5 Τύποι Συστημάτων OLAP 40
2.6 Η γλώσσα ερωτημάτων MDX (MultiDimensional eXpressions) 44
Επίλογος 46
Κεφάλαιο 3: Παραδείγματα κι εφαρμογές 47
Εισαγωγή 47
3.1 Παράδειγμα στο IBM InfoSphere Warehouse 48
3.1.1 IBM InfoSphere Warehouse 48
3.1.1.1 Η βάση 48
3.1.1.2 Design Studio 48

3.1.2 Παράδειγμα Data Warehousing 49
3.1.3 Παράδειγμα OLAP 51
3.2 Παραδείγματα στο Saiku 54
3.2.1 Saiku Analytics 54
3.2.1.1 Η βάση 54
3.2.2 Παράδειγμα.....54
3.2.3 Πράξεις OLAP.....56
Βιβλιογραφία.....58
Άλλες πηγές.....60
Οδηγός χρήσης λογισμικού.....61

Ευρετήριο εικόνων

Εικόνα 1: Βασικές λειτουργίες αποθηκών δεδομένων 13
Εικόνα 2: Πυραμίδα Business Intelligence 15
Εικόνα 3: Μοντέλο διαδικασίας αποθήκης δεδομένων 18
Εικόνα 4: Σχήμα αστέρα 22
Εικόνα 5: Σχήμα αστέρα από τη ΒΔ πιστωτικών καρτών της υποθετικής εταιρίας Acme 23
Εικόνα 6: Βάση Δεδομένων πιστωτικών καρτών της υποθετικής εταιρίας Acme 23
Εικόνα 7: Σχήμα Χιονοστιβάδας 25
Εικόνα 8: Σχήμα αστερισμού 26
Εικόνα 9: Σχέση OLAP-OLTP 29
Εικόνα 10: Ο κύβος OLAP 36
Εικόνα 11: Ιεραρχία "Τοποθεσία" 36
Εικόνα 12: Η πράξη της σύμπτυξης (roll-up) 37
Εικόνα 13: Η πράξη της ανάπτυξης (drill-down) 38
Εικόνα 14: Η πράξη του κοψίματος σε φέτες (slice) 38
Εικόνα 15: Η πράξη του τεμαχισμού σε κύβους (dice) 39
Εικόνα 16: Η πράξη της περιστροφής (pivot) 39
Εικόνα 17: Σύστημα MOLAP 40
Εικόνα 18: Σύστημα ROLAP 41
Εικόνα 19: Σύστημα HOLAP 43
Εικόνα 20: Πίνακες διαστάσεων σχήματος Marts 49
Εικόνα 21: Σχήμα αστέρα παραδείγματος 50
Εικόνα 22: Ροή δεδομένων για BRANCH_LOCATION 50
Εικόνα 23: Σχήμα αστέρα για το παράδειγμα OLAP 51
Εικόνα 24: Δομή OLAP project 52
Εικόνα 25: Saiku 54

Ευρετήριο πινάκων

Πίνακας 1: Διαφορές Σχεσιακών ΒΔ - Αποθηκών Δεδομένων 30

Εισαγωγή

Με την παγκοσμιοποίηση και την αύξηση του ανταγωνισμού στις σημερινές αγορές, υπάρχει μια τεράστια αύξηση στη συλλογή δεδομένων. Οργανισμοί συλλέγουν περισσότερες πληροφορίες και τις διατηρούν για μεγαλύτερο χρονικό διάστημα έτσι ώστε να μπορέσουν να τις αναλύσουν και να τις χρησιμοποιήσουν για να διατηρήσουν την εστίαση στα βασικά κριτήρια των επιχειρήσεων, όπως τα ακόλουθα:

- Παρακολούθηση της απόδοσης
- Ικανοποίηση των πελατών
- Οικονομική ενημέρωση
- Ρυθμιστικές απαιτήσεις που βασίζονται στα διοικητικά όργανα του κλάδου.

Για να είναι επιτυχής και για να αναπτυχθεί μια επιχείρηση, σήμερα, σημαίνει υπέρβαση των ορίων της τρέχουσας υποδομής και κίνηση προς την κατεύθυνση ενός πιο ευέλικτου και ισχυρού περιβάλλοντος Data Warehousing. Οι επιχειρήσεις θα πρέπει να ενισχύσουν το Data Warehousing περιβάλλον τους και να το καταστήσουν ικανό να υποστηρίξει μια ποικιλία πηγών δεδομένων με δομημένα και μη δομημένα δεδομένα και να μπορέσει να εκθέτουν το σύνολο των εν λόγω πληροφοριών στους χρήστες μέσα από ένα ενοποιημένο περιβάλλον εργασίας.
Η παρούσα πτυχιακή εργασία έχει ως στόχο μια περιγραφή των τεχνολογιών Data Warehousing και OLAP καθώς και την πρακτική εφαρμογή τους, μέσω παραδειγμάτων στο περιβάλλον IBM DB2 Data Warehouse και στο περιβάλλον Saiku.
Στο πρώτο κεφάλαιο θα παρουσιαστεί η τεχνολογία Data Warehousing. Αρχικά, παραθέτοντας τα κίνητρα για την ανάπτυξη των αποθηκών δεδομένων και τον κύριο ορισμό τους, κατόπιν, με αναφορά στους στόχους μιας αποθήκης δεδομένων, στον τρόπο εισαγωγής των δεδομένων, στην κατασκευή της αποθήκης και στα σχεσιακά σχήματα που χρησιμοποιούνται για την κατασκευή. Τέλος, θα γίνει σύγκριση των σχεσιακών Βάσεων Δεδομένων με τις Αποθήκες Δεδομένων.

Στο δεύτερο κεφάλαιο θα παρουσιαστούν κάποια γενικά στοιχεία για την OLAP, το FASMI test, ο κύβος κι οι πράξεις OLAP όπως κι οι υποκατηγορίες OLAP. Τέλος, υπάρχει μια σύντομη περιγραφή της γλώσσα MDX.

Στο τρίτο κεφάλαιο θα παρουσιαστούν παραδείγματα των τεχνολογιών Data Warehouse κι OLAP στα προγράμματα IBM Infosphere Warehouse και saiku.

Κεφάλαιο 1: Αποθήκες Δεδομένων (Data Warehouse)

Εισαγωγή

Η λήψη αποφάσεων σε επίπεδο οργανισμών απαιτεί μια ολοκληρωμένη άποψη για όλους τους τομείς μιας επιχείρησης, και ως εκ τούτου πολλοί οργανισμοί έχουν δημιουργήσει συγκεντρωτικές αποθήκες δεδομένων, οι οποίες περιέχουν δεδομένα που προέρχονται από διάφορες βάσεις δεδομένων, οι οποίες συντηρούνται από διαφορετικές μονάδες της επιχείρησης, μαζί με συνοπτικές ιστορικές πληροφορίες.

Σε αυτό το κεφάλαιο παρουσιάζονται τα κίνητρα για την ανάπτυξη αποθηκών δεδομένων, ο ορισμός της αποθήκης, οι στόχοι της, ο τρόπος με τον οποίο γίνεται η εισαγωγή των δεδομένων κι η κατασκευή της αποθήκης. Γίνεται αναφορά στο σχήμα αστέρα και τα υπόλοιπα σχεσιακά σχήματα. Τέλος, γίνεται παρουσίαση των διαφορών μεταξύ των σχεσιακών βάσεων δεδομένων και της αποθήκης δεδομένων.

1.1 Κίνητρα ανάπτυξης αποθηκών δεδομένων

Τα συστήματα διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) από τη δημιουργία τους έδωσαν λύσεις σε πολλά προβλήματα των επιχειρήσεων. Ο όρος On-Line Transaction Processing (OLTP) αποδίδει τον κύριο τρόπο λειτουργίας των σχεσιακών ΣΔΒΔ. Παραδείγματα διεργασιών OLTP είναι η καταγραφή αεροπορικών κρατήσεων μέσω του διαδικτύου, η καταγραφή αναλήψεων και καταθέσεων στα ATM κλπ. Συνοπτικά, τα σχεσιακά ΣΔΒΔ διαχειρίζονται καθημερινές λειτουργίες μιας επιχείρησης με σκοπό την διεκπεραίωση συναλλαγών και για το λόγο αυτό ονομάζονται και Επιχειρησιακές Βάσεις Δεδομένων (Operational Database).

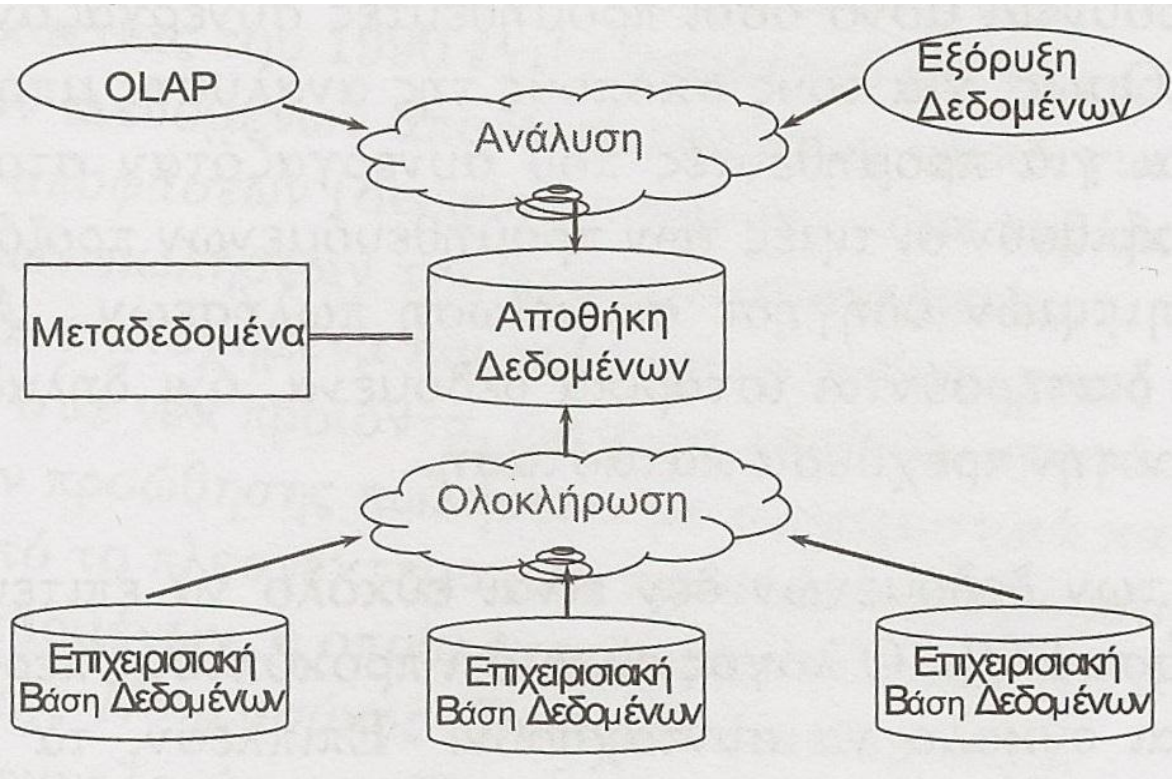
Με το πέρασμα των χρόνων, όμως, οι Βάσεις Δεδομένων των επιχειρήσεων κατέληξαν να περιέχουν τεράστιους όγκους δεδομένους, οι οποίοι, όμως, δεν μπορούσαν να αξιοποιηθούν. Στα μέσα της δεκαετίας του 1990 έγινε αντιληπτό ότι το περιεχόμενο των Επιχειρησιακών Βάσεων Δεδομένων μπορεί να αποτελέσει χρήσιμο υλικό για επεξεργασία με στόχο την υποστήριξη αποφάσεων. Για παράδειγμα, η ανάλυση των πωλήσεων της τελευταίας τριετίας σε μια επιχείρηση, μπορεί να αναδείξει χρήσιμα συμπεράσματα για το ποια προϊόντα πωλούνται συχνότερα, ποιες ομάδες καταναλωτών επηρεάζονται περισσότερο από προσφορές κι εκπτώσεις κι άλλα παρόμοια. Αμέσως γίνεται κατανοητό τα πλεονεκτήματα που προκύπτουν από την αξιοποίηση των Επιχειρησιακών Βάσεων Δεδομένων, που έχει ως αποτέλεσμα την ανάπτυξη της λεγόμενης Επιχειρησιακής Ευφυΐας (Business Intelligence).

Η τεχνολογία και τα εργαλεία των σχεσιακών ΣΔΒΔ, δεν μπορούν, όμως, να χρησιμοποιηθούν για τους σκοπούς της ανάλυσης των περιεχομένων των Επιχειρησιακών Βάσεων Δεδομένων. Αυτό συμβαίνει για τους παρακάτω λόγους:

Πρώτον, τα δεδομένα των Επιχειρησιακών Βάσεων Δεδομένων στις περισσότερες περιπτώσεις δεν έχουν καλή ποιότητα, δηλαδή υπάρχουν ελλιπή στοιχεία κι ασυνέπειες με αποτέλεσμα τα δεδομένα να μην προσφέρονται για άμεση αξιοποίηση. Δεύτερον, στα πλαίσια μιας επιχείρησης οι Επιχειρησιακές Βάσεις Δεδομένων είναι τις περισσότερες φορές ανεξάρτητες. Αυτό σημαίνει, ότι διαφορετικά τμήματα μιας επιχείρησης έχουν, συνήθως, διαφορετικές Βάσεις Δεδομένων, με αποτέλεσμα να μην αλληλοενημερώνονται και τα δεδομένα να είναι ετερογενή. Για το λόγο αυτό, πριν αναλυθούν τα δεδομένα θα πρέπει να

ομογενοποιηθούν, κι επιπλέον, θα πρέπει να επιλεγούν μόνο όσα από τα δεδομένα είναι χρήσιμα για τους σκοπούς της ανάλυσης. Τρίτον, στις Επιχειρησιακές Βάσεις Δεδομένων, διατηρούνται μόνο δεδομένα για την τρέχουσα κατάσταση. Όμως για τους σκοπούς της ανάλυσης χρειάζονται και προγενέστερα δεδομένα. Για το λόγο αυτό, θα πρέπει να διατηρούνται και ιστορικά δεδομένα. Τέταρτον, η μέχρι τότε τεχνολογία και τα εργαλεία όπως η γλώσσα SQL, δεν προσφέρονταν για περίπλοκα ερωτήματα ανάλυσης δεδομένων. Επιπροσθέτως, και τα σχεσιακά ΣΔΒΔ στο φυσικό επίπεδο δεν είναι σχεδιασμένα έτσι ώστε να μπορούν να ανταποκρίνονται στις απαιτήσεις περίπλοκων ερωτημάτων. Επομένως, η ανάγκη για νέες τεχνολογίες και νέα εργαλεία ήταν εμφανής. Τέλος, στα Σχεσιακά ΣΔΒΔ, τα δεδομένα οργανώνονται βάση μεθοδολογίες όπως το διάγραμμα Οντοτήτων - Συσχετίσεων και η κανονικοποίηση, οι οποίες ικανοποιούν τις απαιτήσεις της αποδοτικής OLTP, αλλά παράγουν βάσεις δεδομένων που στο εννοιολογικό επίπεδο είναι περίπλοκες. Για την ανάλυσης των δεδομένων, απαιτούνται διαφορετικές τεχνικές σχεδιασμού και οργάνωσης των δεδομένων στο εννοιολογικό επίπεδο.

Προκειμένου να ικανοποιηθούν οι προαναφερθείσες ανάγκες και τα κενά που προέκυπταν από τα σχεσιακά ΣΔΒΔ, αναπτύχθηκαν οι αποθήκες δεδομένων. Η τεχνολογία των αποθηκών δεδομένων παρέχει ολοκλήρωση ετερογενών πηγών δεδομένων και μια πλατφόρμα για αποδοτική ανάλυση ιστορικών δεδομένων. Επί της ουσίας, οι αποθήκες δεδομένων αποτελούν συλλογές δεδομένων που επιλέγονται από τις Επιχειρησιακές Βάσεις Δεδομένων, ολοκληρώνονται και στη συνέχεια τα δεδομένα, που υπάρχουν σε αυτές, αναλύονται με διαδικασίες όπως η OLAP ή η εξόρυξη δεδομένων. Εξαιτίας των διαφορετικών απαιτήσεων, οι αποθήκες δεδομένων διατηρούνται σε διαφορετικά υπολογιστικά συστήματα από τις επιχειρησιακές βάσεις, από τις οποίες αντιγράφονται και ολοκληρώνονται τα δεδομένα. Ανάλογα με το πόσο συχνά ενημερώνονται οι αποθήκες ως προς τις μεταβολές των δεδομένων των επιχειρησιακών βάσεων, οι αποθήκες μπορεί να μην περιέχουν τα πλέον ενημερωμένα δεδομένα. Στις περισσότερες, όμως, εφαρμογές αυτό δεν αποτελεί πρόβλημα, καθώς η ανάλυση των δεδομένων δεν επηρεάζεται από λίγες, πρόσφατες μεταβολές. Οι βασικές λειτουργίες των αποθηκών δεδομένων συνοψίζονται στην παρακάτω εικόνα:



Εικόνα 1: Βασικές λειτουργίες αποθηκών δεδομένων

1.2 Ορισμός

Ο πιο δημοφιλής ορισμός για το τι είναι μία αποθήκη δεδομένων είναι αυτός που δόθηκε από τον Αμερικανό επιστήμονα των υπολογιστών, William H. Inmon. Σύμφωνα με τον Inmon, αποθήκη δεδομένων είναι μία θεματική (subject-oriented), ολοκληρωμένη (integrated), μεταβαλλόμενη ως προς το χρόνο (time-variant), και μη ευμετάβλητη συλλογή δεδομένων για την υποστήριξη λήψης αποφάσεων.

Ο όρος “θεματική” σημαίνει ότι μία αποθήκη δεδομένων μπορεί να χρησιμοποιηθεί για να αναλύσει μια συγκεκριμένη θεματική περιοχή. Για παράδειγμα, οι Πωλήσεις μπορεί να είναι μια συγκεκριμένη θεματική περιοχή.

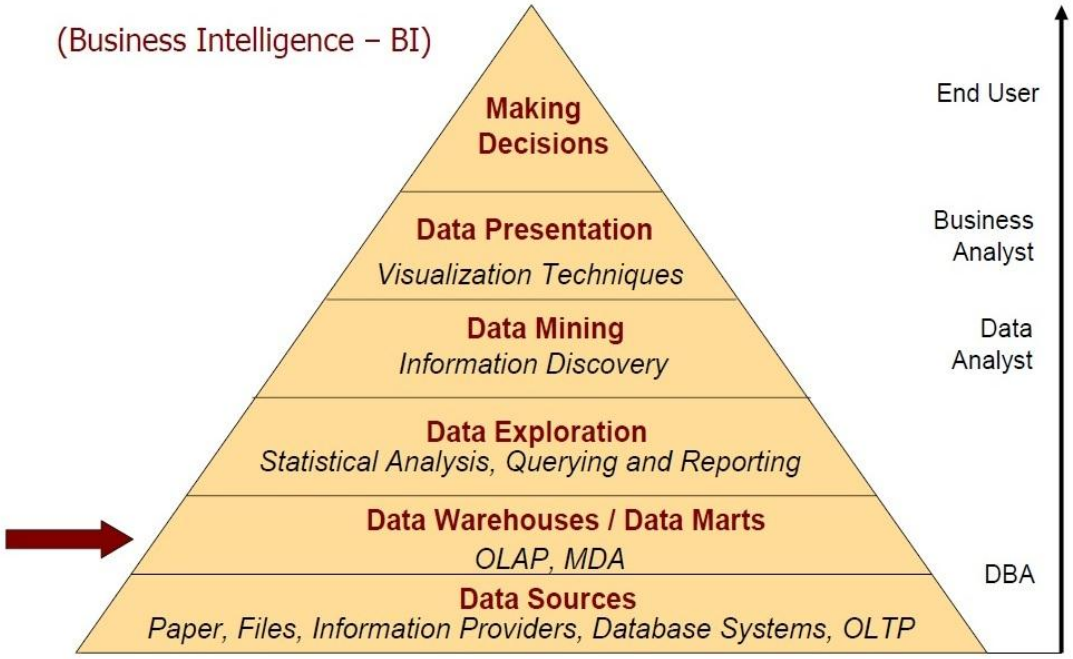
Ο όρος “ολοκληρωμένη” σημαίνει ότι μία αποθήκη δεδομένων μπορεί να ενσωματώνει δεδομένα από πολλαπλές πηγές δεδομένων, αλλά τα δεδομένα που αφορούν στην ίδια έννοια (θέμα), ορίζονται με ίδιο τρόπο. Για παράδειγμα, μια πηγή Α και μια πηγή Β μπορεί να έχουν διαφορετικούς τρόπους αναγνώρισης ενός προϊόντος, αλλά σε μια αποθήκη δεδομένων, θα υπάρχει μόνο ένας τρόπος προσδιορισμού του προϊόντος αυτού. Η ολοκλήρωση εφαρμόζεται μέσω μετασχηματισμών κατά την φόρτωση των δεδομένων.

Ο όρος "μεταβαλλόμενη ως προς το χρόνο " σημαίνει ότι όταν με το πέρας του χρόνου, αλλάζουν τα δεδομένα που σχετίζονται με μία έννοια, ιστορικά δεδομένα διατηρούνται στην αποθήκη δεδομένων. Αυτό σημαίνει ότι κάθε δεδομένο συνοδεύεται από στοιχεία που αφορούν στο χρόνο. Για παράδειγμα, αν ένας πελάτης αλλάξει διεύθυνση, διατηρείτε η παλαιά του διεύθυνση και το χρονικό διάστημα για το οποίο ίσχυε. Σε μία αποθήκη δεδομένων, κάποιος μπορεί να ανακτήσει δεδομένα τριών, έξι, δώδεκα μηνών, ή προγενέστερα. Αυτό έρχεται σε αντίθεση με ένα σύστημα συναλλαγών, όπου συχνά μόνο τα πιο πρόσφατα στοιχεία διατηρούνται. Για παράδειγμα, ένα σύστημα συναλλαγών μπορεί να κρατήσει την πιο πρόσφατη διεύθυνση του πελάτη, ενώ μια αποθήκη δεδομένων μπορεί να κρατήσει όλες τις διευθύνσεις που σχετίζονται με έναν πελάτη.

Ο όρος “μη ευμετάβλητη” σημαίνει ότι άπαξ τα δεδομένα εισαχθούν στην αποθήκη δεδομένων, δεν διαγράφονται. Σε περίπτωση που αλλάξουν αποθηκεύονται, όπως αναφέρθηκε, ως ιστορικά δεδομένα.

1.3 Επιχειρησιακή Ευφυΐα κι Αποθήκες Δεδομένων

Η επιχειρησιακή ευφυΐα (Business Intelligence, BI) περιλαμβάνει τεχνολογίες και εφαρμογές για συλλογή, αποθήκευση, ανάλυση και επεξεργασία επιχειρησιακών δεδομένων με στόχο την υποστήριξη αποφάσεων. Παράγει μεγάλες ποσότητες πληροφοριών, οι οποίες μπορούν να βοηθήσουν στην ανάπτυξη νέων δυνατοτήτων. Ο προσδιορισμός αυτών των δυνατοτήτων κι η εφαρμογή μιας αποτελεσματικής στρατηγικής μπορούν να προσφέρουν ένα ανταγωνιστικό πλεονέκτημα στην αγορά και μακροπρόθεσμη σταθερότητα. Η Business Intelligence περιλαμβάνει μεταξύ άλλων τις αποθήκες δεδομένων και την αναλυτική επεξεργασία άμεσης επικοινωνίας (OLAP).



Εικόνα 2: Πυραμίδα Business Intelligence

1.4 Οι στόχοι μιας αποθήκης δεδομένων

Οι θεμελιώδεις στόχοι μιας αποθήκης δεδομένων μπορούν να αναπτυχθούν αν κάποιος έρθει σε επαφή με το τμήμα της διοίκηση επιχειρήσεων ενός οργανισμού. Εκεί ακούγονται εκφράσεις όπως οι παρακάτω:

- “Υπάρχουν βουνά δεδομένων στην εταιρεία, αλλά δεν μπορούμε να έχουμε πρόσβαση”
- “Χρειαζόμαστε να προβάλουμε σωστά τα δεδομένα με κάθε τρόπο”
- “Θέλουμε να είναι εύκολο για τους επιχειρηματίες να παίρνουν τα δεδομένα άμεσα”
- “Απλά δείξτε μου τι είναι σημαντικό”
- “Θέλουμε οι άνθρωποι να χρησιμοποιούν τις πληροφορίες για να υποστηρίξουν περισσότερες λήψεις αποφάσεων βασισμένες στα γεγονότα”

Οι ανησυχίες αυτές είναι τόσο καθολικές ώστε να οδηγούν σε θεμελιώδεις απαιτήσεις για αποθήκες δεδομένων. Οι “απαιτήσεις” αυτές της διοίκησης επιχειρήσεων θα πρέπει να μετατραπούν σε απαιτήσεις της αποθήκης δεδομένων. Η αποθήκη δεδομένων πρέπει να καταστήσει τις πληροφορίες μιας επιχείρησης εύκολα προσβάσιμες. Τα περιεχόμενα μιας αποθήκης δεδομένων πρέπει να είναι κατανοητά. Τα δεδομένα πρέπει να είναι προφανή για τον επαγγελματία χρήστη κι όχι μόνο για τον προγραμματιστή. Η κατανόηση συνεπάγεται αναγνωσιμότητα, τα περιεχόμενα μιας αποθήκης δεδομένων χρειάζεται να επισημαίνονται με σημασία. Οι επαγγελματίες χρήστες θέλουν να διαχωρίζουν και να συνδυάζουν τα δεδομένα στην αποθήκη σε ατελείωτους συνδυασμούς, μια διαδικασία που συνήθως αναφέρεται ως *slicing and dicing*. Τα εργαλεία που έχουν πρόσβαση στην αποθήκη δεδομένων πρέπει να είναι απλά και εύκολα στη χρήση. Επίσης, πρέπει να επιστρέφουν τα αποτελέσματα του ερωτήματος στον χρήστη με τον ελάχιστο χρόνο αναμονής.

Η αποθήκη δεδομένων πρέπει να παρουσιάζει τις πληροφορίες του οργανισμού με συνέπεια. Τα δεδομένα στην αποθήκη πρέπει να είναι αξιόπιστα και πρέπει να συλλέγονται προσεκτικά από ποικίλες πηγές γύρω από τον οργανισμό, να καθαρίζονται, να εξασφαλίζεται η ποιότητα, και να απελευθερώνονται μόνο όταν είναι κατάλληλα για επεξεργασία από τον χρήστη. Οι πληροφορίες από μία

επιχειρηματική διαδικασία θα πρέπει να ταιριάζουν με τις πληροφορίες από μία άλλη επιχειρηματική διαδικασία. Εάν δύο μέτρα απόδοσης έχουν το ίδιο όνομα, τότε θα πρέπει να σημαίνουν το ίδιο πράγμα. Αντίθετα, αν τα δύο μέτρα απόδοσης δεν σημαίνουν το ίδιο πράγμα, τότε θα πρέπει να επισημαίνονται με διαφορετικό τρόπο. Αξιόπιστες πληροφορίες σημαίνουν πληροφορίες υψηλής ποιότητας. Αυτό σημαίνει ότι όλα τα δεδομένα είναι αντιπροσωπευτικά και ολοκληρωμένα. Η συνέπεια σημαίνει επίσης, ότι κοινοί ορισμοί για τα περιεχόμενα της αποθήκης δεδομένων είναι διαθέσιμοι για τους χρήστες.

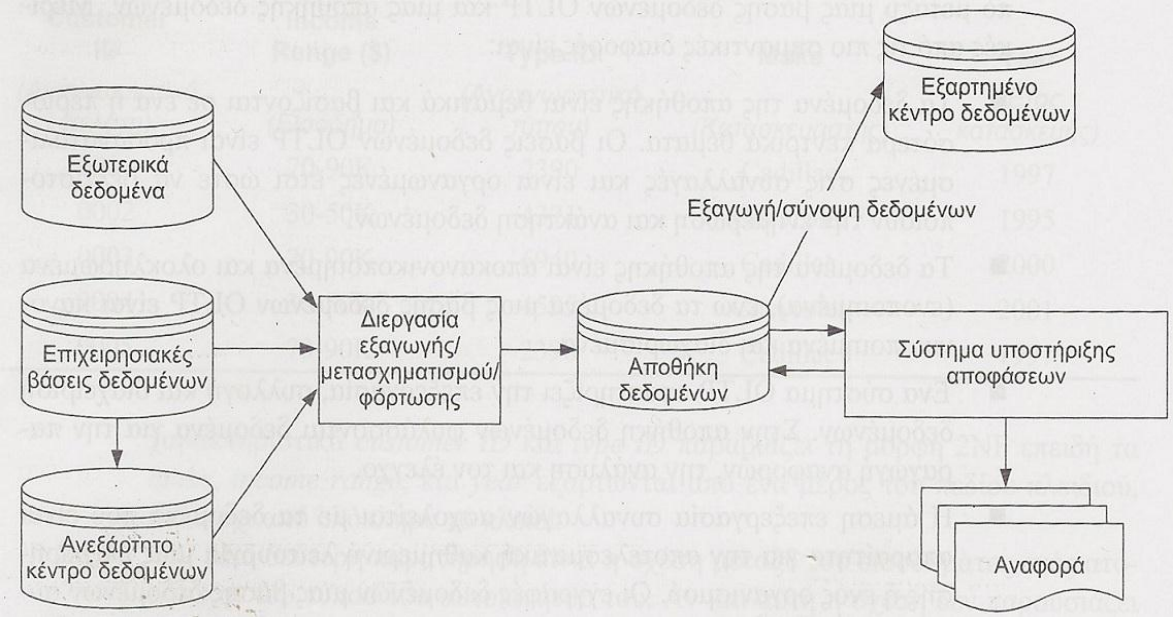
Η αποθήκη δεδομένων πρέπει να είναι προσαρμοστική και ελαστική στο να αλλάξει. Απλά είναι σχεδόν αδύνατο να αποφευχθεί η αλλαγή. Οι ανάγκες των χρηστών, οι επιχειρηματικές συνθήκες, τα δεδομένα, και η τεχνολογία αλλάζουν με το πέρασμα του χρόνου. Η αποθήκη δεδομένων πρέπει να είναι σχεδιασμένη με τέτοιο τρόπο ώστε να μπορεί να χειριστεί αυτές τις αναπόφευκτες αλλαγές. Οι αλλαγές στην αποθήκη δεδομένων πρέπει να είναι κομψές, που σημαίνει ότι δεν ακυρώνουν τα υπάρχοντα δεδομένα ή τις υπάρχουσες εφαρμογές. Τα υπάρχοντα δεδομένα και οι εφαρμογές δεν θα πρέπει να αλλάζουν ή να διακόπτονται όταν η επιχειρηματική κοινότητα θέτει νέα ερωτήματα ή όταν νέα δεδομένα προστίθενται στην αποθήκη. Αν τα περιγραφικά δεδομένα στην αποθήκη έχουν τροποποιηθεί, τότε πρέπει να αντιπροσωπεύουν κατάλληλα τις αλλαγές.

Η αποθήκη δεδομένων πρέπει να είναι ένα ασφαλές οχυρό, το οποίο προστατεύει τα “περιουσιακά στοιχεία”. Οι πολυτιμότερες πληροφορίες μιας οργάνωσης αποθηκεύονται στην αποθήκη δεδομένων. Η αποθήκη πολύ πιθανό να περιέχει πληροφορίες σχετικά με το τι πουλάει η επιχείρηση, σε ποια τιμή - δυνητικά επιβλαβή στοιχεία στα χέρια των λάθος ανθρώπων. Η αποθήκη δεδομένων πρέπει να ελέγχει αποτελεσματικά την πρόσβαση στις εμπιστευτικές πληροφορίες του οργανισμού.

Η αποθήκη δεδομένων θα πρέπει να χρησιμεύει ως θεμέλιο για τη βελτίωση της λήψης αποφάσεων. Γι αυτό και πρέπει να περιέχει τα σωστά δεδομένα για να υποστηρίξει τη λήψη αποφάσεων. Υπάρχει μόνο ένα πραγματικό παράγωγο από την αποθήκη δεδομένων: η απόφαση που παίρνεται αφού η αποθήκη δεδομένων έχει παρουσιάσει τα στοιχεία της. Οι αποφάσεις αυτές παρέχουν τον αντίκτυπο των επιχειρήσεων και την αξία που αναλογεί στην αποθήκη.

1.5 Εισαγωγή δεδομένων στην αποθήκη

Τα δεδομένα μπορούν να εισαχθούν στην αποθήκη από τρεις κύριες πηγές, όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 3: Μοντέλο διαδικασίας αποθήκης δεδομένων

Στα εξωτερικά δεδομένα συγκαταλέγονται στοιχεία όπως οικονομικοί δείκτες, πληροφορίες για τον καιρό και άλλα παρόμοια που δεν σχετίζονται με το εσωτερικό της εταιρείας. Ανεξάρτητο κέντρο δεδομένων (indipendent data mart) ονομάζεται ο χώρος αποθήκης δεδομένων, ο οποίος είναι παρόμοιος με μια αποθήκη, αλλά επικεντρώνεται σε ένα μόνο θέμα. Το ανεξάρτητο κέντρο δεδομένων δομείται με βάση επιχειρησιακά δεδομένα, αλλά και δεδομένα από εξωτερικές πηγές. Τα δεδομένα που είναι αποθηκευμένα σε ένα κέντρο μπορούν να φορτωθούν στην κεντρική αποθήκη δεδομένων για χρήση από άλλα τμήματα της εταιρείας. Ανεξάρτητα από την εξωτερική πηγή δεδομένων, για τη διαδικασία της μετακίνησης δεδομένων στην αποθήκη πιθανότητα θα χρειαστεί προγραμματισμός σε διαδικαστική γλώσσα χαμηλού επιπέδου.

Τα δεδομένα πριν εισαχθούν στην αποθήκη επεξεργάζονται με μία διεργασία εξαγωγής, μετασχηματισμού, και φόρτωσης (Extract-Transform-Load process-ETL). Στις βασικές λειτουργίες της διεργασίας αυτής περιλαμβάνονται: η εξαγωγή δεδομένων από μία ή περισσότερες από τις πηγές εισόδου, ο καθαρισμός και ο κατάλληλος μετασχηματισμός των εξαγόμενων δεδομένων, και η φόρτωση των δεδομένων στην αποθήκη. Οι μετασχηματισμοί δεδομένων χρησιμοποιούνται

συχνά για την επίλυση προβλημάτων βαθμού λεπτομέρειας των δεδομένων, τη διόρθωση ασυνεπειών στα δεδομένα μεταξύ πολλών επιχειρησιακών βάσεων δεδομένων, και τη χρονοσήμανση κάθε εγγραφής δεδομένων. Μετά το μετασχηματισμό και τον καθαρισμό τους, τα δεδομένα εισάγονται στην αποθήκη όπου αποθηκεύονται σε σχεσιακή ή πολυδιάστατη μορφή.
Κατά κανόνα, μετά την εισαγωγή τους στην αποθήκη, τα δεδομένα δεν υπόκεινται σε αλλαγές. Μια προφανής εξαίρεση του κανόνα είναι η περίπτωση κατά την οποία εντοπίζονται σφάλματα στα δεδομένα. Υπάρχουν, όμως, και ειδικές καταστάσεις, όπως όταν ένα άτομο αλλάζει τη διεύθυνσή του ή την οικογενειακή του κατάσταση.

Για παράδειγμα, έχουμε μια αποθήκη δεδομένων στην οποία φυλάσσονται συναλλαγές πελατών για δαπάνες με πιστωτικές κάρτες κι ένας πελάτης αλλάζει την οικογενειακή του κατάσταση από ανύπαντρος σε έγγαμος. Η πρώτη σκέψη είναι να αλλάξουμε απλώς την οικογενειακή κατάσταση σε όλες τις εγγραφές της αποθήκης που αναφέρονται στον πελάτη αυτό. Το πρόβλημα με σε αυτή την περίπτωση είναι ότι οποιαδήποτε ανάλυση χρησιμοποιεί τις πληροφορίες οικογενειακής κατάστασης, επεξεργάζεται αλλοιωμένα δεδομένα, καθώς οι δαπάνες που έγιναν από το άτομο όταν ήταν ανύπαντρο θεωρούνται ως δαπάνες που έγιναν από έναν παντρεμένο πελάτη. Επομένως, η καταγραφή μιας τέτοιας αλλαγής στις πληροφορίες πρέπει να γίνει με διαφορετικό τρόπο. Έχουν προταθεί πολλές λύσεις. Σε μία από αυτές προτείνεται η δημιουργία πεδίων εγγραφών για τη διατήρηση τόσο των προηγούμενων όσο και των τρεχουσών τιμών για κάθε χαρακτηριστικό Μια δεύτερη μέθοδος περιλαμβάνει τη δημιουργία μιας νέας εγγραφής κάθε φορά που αλλάζει η τιμή ενός χαρακτηριστικού σε μια υπάρχουσα εγγραφή. Κατόπιν, η υπάρχουσα και η καινούρια εγγραφή συνδέονται με ένα κλειδί. Ωστόσο η εύρεση μιας γενικής λύσης για την απεικόνιση των αλλαγών κατάστασης των εγγραφών μέσα στην αποθήκη παραμένει άλυτο πρόβλημα.
Στην αποθήκη φυλάσσονται επίσης, μεταδεδομένα (metadata). Τα μεταδεδομένα, με τεχνικούς όρους, είναι δεδομένα που αναφέρονται σε δεδομένα. Ο σκοπός τους είναι να επιτρέψουν την καλύτερη κατανόηση της φύσης των δεδομένων που περιέχονται σε μία αποθήκη. Έχουν οριστεί δύο γενικοί τύποι μεταδεδομένων: τα δομικά και τα επιχειρησιακά. Τα δομικά μεταδεδομένα δίνουν έμφαση στις περιγραφές δεδομένων, τους τύπους δεδομένων, τους κανόνες αναπαράστασης, και τις σχέσεις μεταξύ των δεδομένων. Τα επιχειρησιακά μεταδεδομένα αφορούν

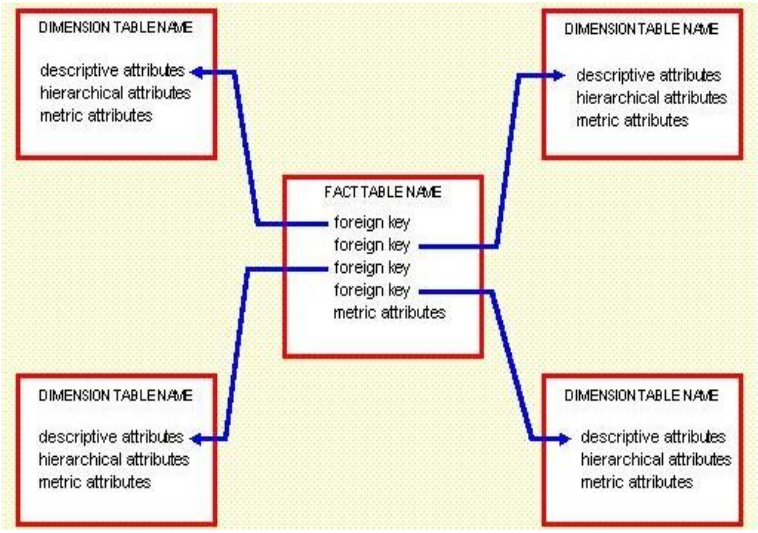
κυρίως την περιγραφή της ποιότητας και της χρήσης των δεδομένων. Μια κύρια διαφορά μεταξύ δομικών και επιχειρησιακών δεδομένων είναι ότι τα πρώτα είναι στατικά, ενώ τα δεύτερα βρίσκονται σε μια διαρκή κατάσταση αλλαγής.

1.6 Κατασκευή της αποθήκης δεδομένων

Γενικά, υπάρχουν δύο τεχνικές για την υλοποίηση μιας αποθήκης δεδομένων. Μία μέθοδος είναι η δημιουργία του μοντέλου αποθήκης ως πολυδιάστατου πίνακα. Σε αυτή την περίπτωση τα δεδομένα αποθηκεύονται σε μορφή παρόμοια με εκείνη που χρησιμοποιείται για την παρουσίαση στο χρήστη. Συνήθως, χρησιμοποιείται η δεύτερη μέθοδος, στην οποία τα δεδομένα αποθήκης αποθηκεύονται με βάση το σχεσιακό μοντέλο, ενώ χρησιμοποιείται μια μηχανή σχεσιακής βάσης δεδομένων για την παρουσίαση τους στο χρήστη σε πολυδιάστατη μορφή. Μια από τις πιο δημοφιλείς τεχνικές σχεσιακής μοντελοποίησης είναι το σχήμα αστέρα (star schema). Άλλες τεχνικές σχεσιακής μοντελοποίησης είναι το σχήμα χιονοστιβάδας και το σχήμα αστερισμού.

1.7 Σχήμα αστέρα

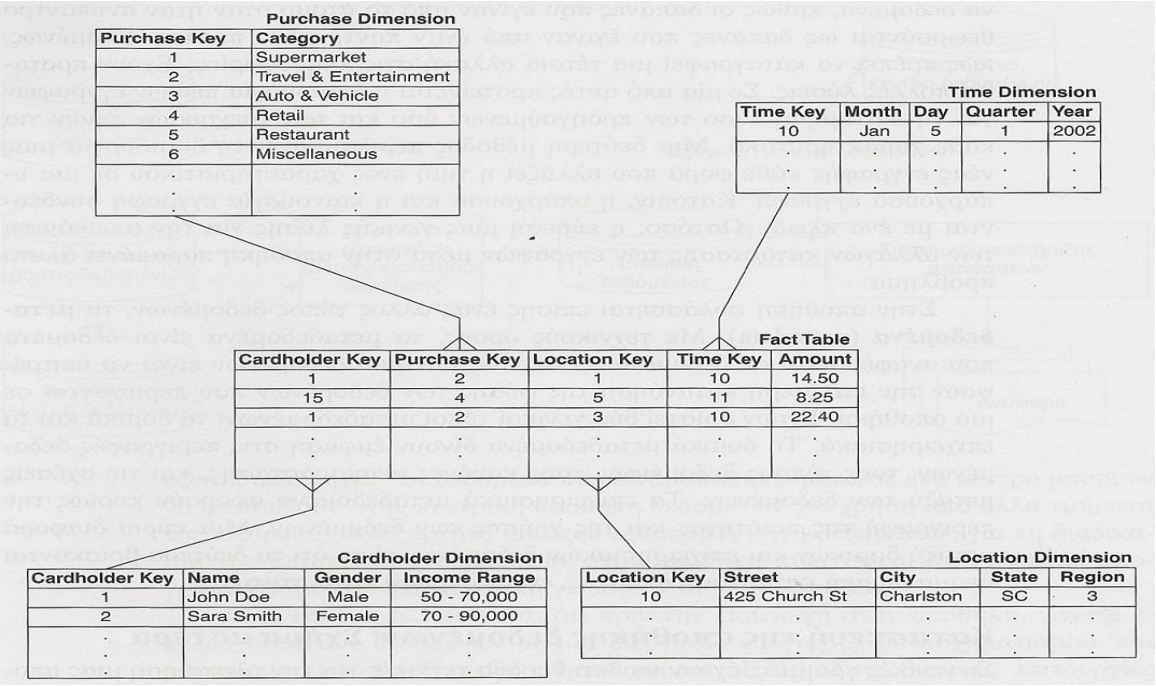
Το σχήμα αστέρα είναι ο απλούστερος τρόπος απεικόνισης ενός σχήματος αποθήκης δεδομένων. Ονομάζεται σχήμα αστέρα, επειδή το διάγραμμα μοιάζει με αστέρι. Στο κέντρο του αστέρα υπάρχει ο πίνακας γεγονότων (fact table) και στα άκρα του υπάρχουν οι πίνακες διαστάσεων, οι οποίοι είναι αποκανονικοποιημένοι. Ο πίνακας γεγονότων έχει συνήθως δύο τύπους πληροφοριών, τα ξένα κλειδιά, για σύνδεση με τους πίνακες διαστάσεων (dimension tables), και τα μέτρα (measure), που περιέχουν αριθμητικά δεδομένα. Τα κλειδιά είναι τιμές που δημιουργούνται από το σύστημα και, στο σύνολο τους, καθορίζουν με μονοσήμαντο τρόπο κάθε εγγραφή του πίνακα γεγονότων. Τα κλειδιά ορίζουν τις συντεταγμένες της πολυδιάστατης δομής που αντιπροσωπεύεται από το σχήμα αστέρα. Οι πίνακες διαστάσεων περιέχουν δεδομένα σχετικά με την κάθε διάσταση. Η σχέση μεταξύ ενός πίνακα διαστάσεων και του πίνακα γεγονότων είναι ένα-προς-πολλά. Ως επακόλουθο, οι πίνακες διαστάσεων θα είναι σημαντικά μικρότεροι από τον κεντρικό πίνακα γεγονότων. Παρόλο που ο πίνακας γεγονότων είναι μορφής 3NF, οι πίνακες διαστάσεων, όπως αναφέρθηκε, δεν είναι κανονικοποιημένοι. Η επιλογή των χαρακτηριστικών που συνθέτουν έναν πίνακα διαστάσεων καθορίζεται σε μεγάλο βαθμό από τη φύση των αναλυτικών ερωτήσεων που πρέπει να απαντηθούν από το σχήμα αστέρα. Τέλος, οι διαστάσεις του σχήματος αστέρα αναφέρονται συχνά ως αργά μεταβαλλόμενες διαστάσεις (slowly changing dimensions). Αυτό οφείλεται στο γεγονός ότι οι πληροφορίες των πινάκων διαστάσεων είναι αυτές που δεν υπόκεινται σε συχνές αλλαγές. Στη συνέχεια, υπάρχει ένα παράδειγμα για το σχήμα αστέρα από το



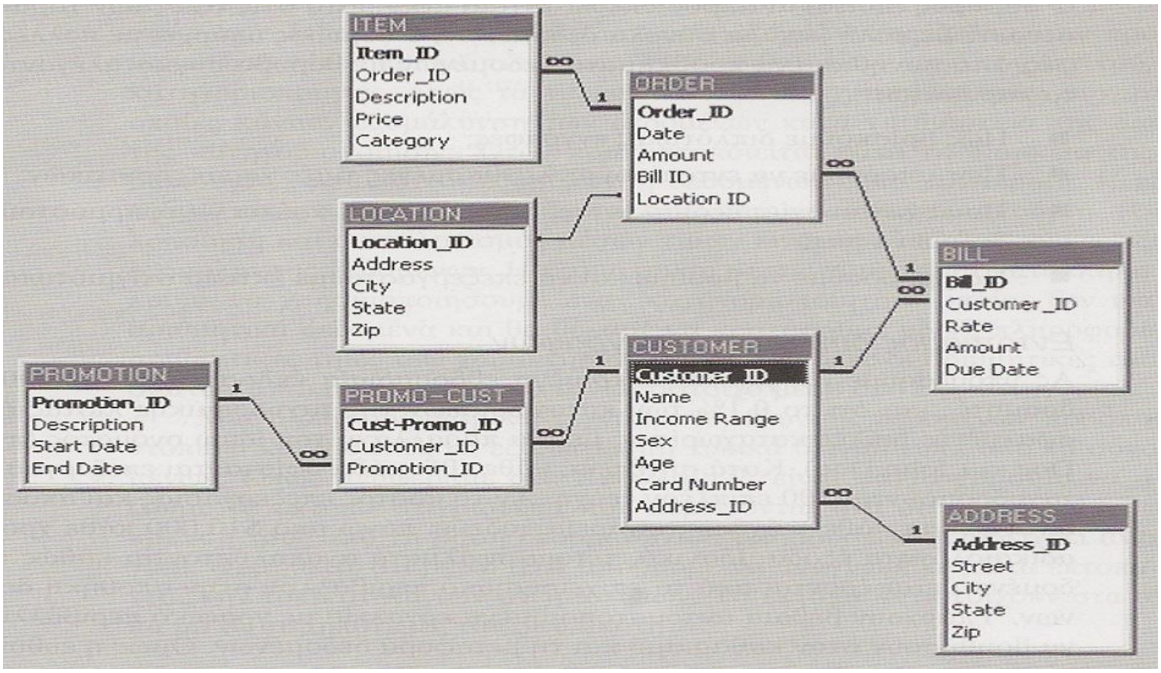
Εικόνα 4: Σχήμα αστέρα

βιβλίο «Εξόρυξη Πληροφορίας, Ένας εισαγωγικός οδηγός με παραδείγματα» των Richard J. Roiger, Michael W. Geatz.

Η Εικόνα 5 δείχνει τη διάρθρωση ενός σχήματος αστέρα που δημιουργήθηκε από τη βάση δεδομένων που φαίνεται στην Εικόνα 6 και απεικονίζει τη βάση δεδομένων πιστωτικών καρτών της υποθετικής εταιρίας Acme. Το θέμα του σχήματος αστέρα είναι οι αγορές με πιστωτικές κάρτες. Ο πίνακας γεγονότων έχει τέσσερις διαστάσεις - cardholder, purchase, location, και time.



Εικόνα 5: Σχήμα αστέρα από τη ΒΔ πιστωτικών καρτών της υποθετικής εταιρίας Acme



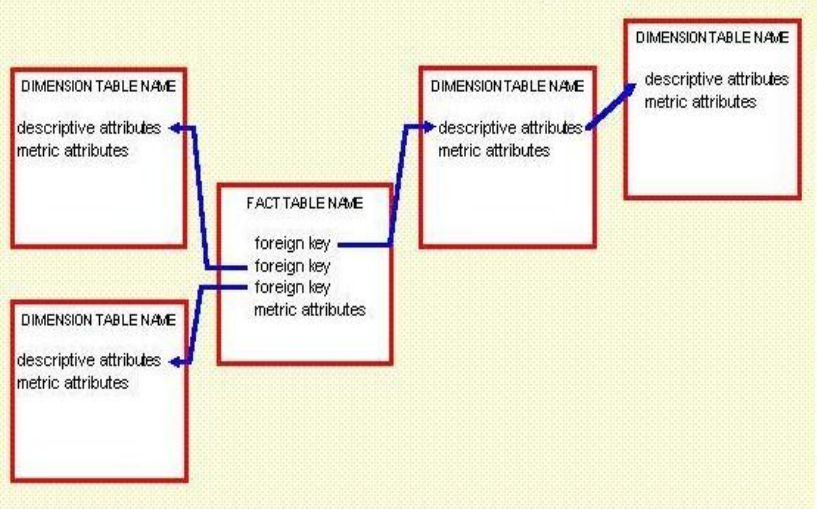
Εικόνα 6: Βάση Δεδομένων πιστωτικών καρτών της υποθετικής εταιρίας Acme

Το σχήμα αστέρα της Εικόνας 5 δείχνει τέσσερις πίνακες διαστάσεων, που σχετίζονται με το σχήμα αστέρα. Ο πίνακας διαστάσεων cardholder συνδέεται με τη διάσταση cardholder και περιέχει το όνομα, το γένος, και την τάξη εισοδήματος κάθε πελάτη που περιέχεται στη βάση δεδομένων. Ο πίνακας διαστάσεων που συνδέεται με τη διάσταση purchase περιέχει τις πιθανές κατηγορίες για κάθε αγορά με πιστωτική κάρτα. Ο πίνακας διαστάσεων που σχετίζεται με τη διάσταση location περιέχει πληροφορίες για την τοποθεσία ενός είδους που αγοράστηκε. Ο πίνακας location διατηρεί μια τιμή για κάθε πολιτεία και για κάθε περιοχή. Το μοντέλο θεωρεί ότι υπάρχουν τέσσερις περιοχές, με κάθε πολιτεία να είναι μέρος μίας, και μόνο μίας, περιοχής. Τέλος, ο πίνακας διαστάσεων για το time δείχνει το χρόνο σε μορφή ημερών, μηνών, τριμήνων, ή ετών. Εκτός από τα πεδία διαστάσεων , ο πίνακας γεγονότων μπορεί να συσχετίζει ένα ή περισσότερα γεγονότα με κάθε εγγραφή. Ο πίνακας γεγονότων περιέχει ένα γεγονός που αντιπροσωπεύει το ποσό δαπάνης για κάθε συναλλαγή πιστωτικής κάρτας. Η πρώτη καταχώρηση του πίνακα γεγονότων δείχνει την αγορά με πιστωτική κάρτα από τον John Doe, έναν άνδρα με ετήσιο εισόδημα μεταξύ $ 50000 και $ 70000. Η δαπάνη ήταν ένα είδος ταξιδιού και αναψυχής και έγινε στις 5 Ιανουαρίου του 2002. Το ποσό της δαπάνης ήταν $ 14,50.

1.8 Πρόσθετα σχεσιακά σχήματα

Σχήμα χιονοστιβάδας (snowflake schema)

Το σχήμα χιονοστιβάδας είναι μια παραλλαγή του σχήματος αστέρα. Η κύρια διαφορά μεταξύ των δύο σχημάτων είναι στον ορισμό των πινάκων διαστάσεων. Στο σχήμα χιονοστιβάδας γίνεται περαιτέρω διαίρεση σε μερικούς από τους πίνακες διαστάσεων που συνδέονται απευθείας στον πίνακα γεγονότων. Αυτό επιτρέπει στους πίνακες διαστάσεων να κανονικοποιηθούν, το οποίο με τη σειρά του σημαίνει λιγότερο συνολικό χώρο αποθήκευσης. Αν το σχήμα χιονοστιβάδας είναι πλήρες, φαίνεται ως ένα πλήρες διάγραμμα τρίτης κανονικής μορφής. Το κύριο πλεονέκτημα είναι η αύξηση στην αποτελεσματικότητα αποθήκευσης επειδή υπάρχει λιγότερος πλεονασμός.



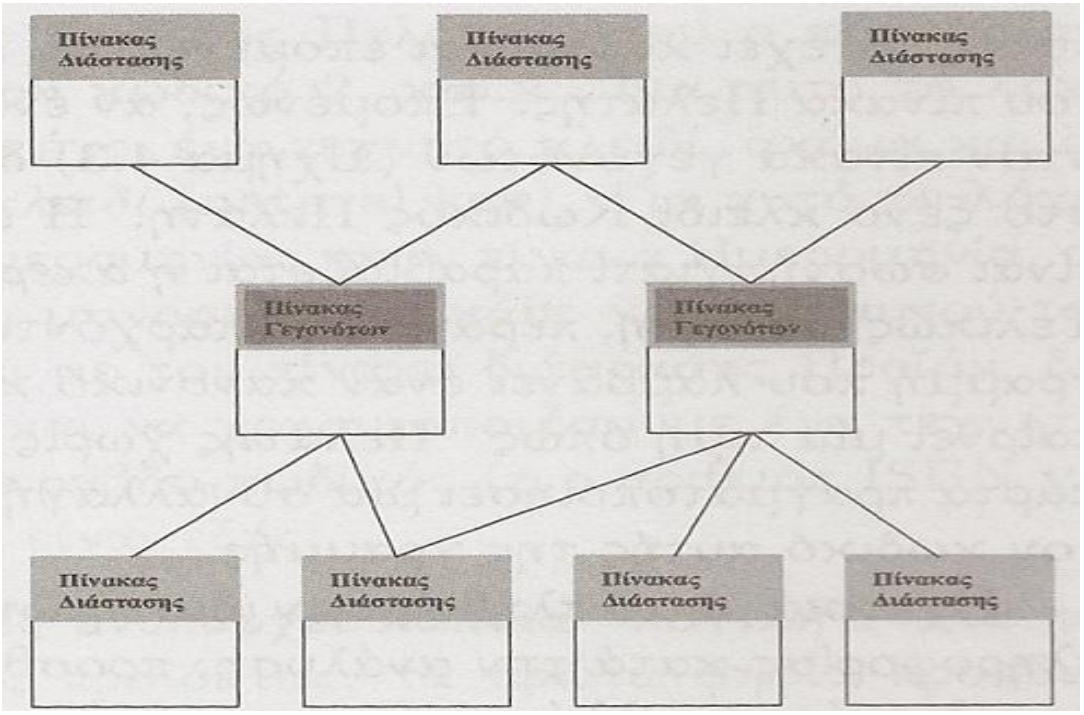
Εικόνα 7: Σχήμα Χιονοστιβάδας

Επιπλέον, επειδή οι σχεσιακοί πίνακες είναι μικρότεροι, οι πράξεις σύζευξης έχουν βελτιωμένη απόδοση. Όμως η αύξηση του συνολικού αριθμού των πινάκων και των μεταξύ τους συνδέσεων μεγαλώνει την πολυπλοκότητα των ερωτημάτων που απαιτούνται για την εξαγωγή των ίδιων πληροφοριών σε σχέση με την περίπτωση που οι πίνακες δεν έχουν κανονικοποιηθεί.

Σχήμα αστερισμού (constellation schema).

Το σχήμα αστερισμού είναι μια ακόμα παραλλαγή του σχήματος αστέρα η οποία χρησιμοποιείται, συνήθως, σε πιο σύνθετες εφαρμογές. Το σχήμα αυτό λαμβάνει υπόψη μοντέλα που απαιτούν περισσότερους από έναν κεντρικούς πίνακες γεγονότων. Οι πολυθεματικές αποθήκες δεδομένων είναι εκείνες που επωφελούνται περισσότερο από το σχήμα αστερισμού. Το κύριο μειονέκτημα του σχήματος αστερισμού είναι η αρκετά περίπλοκη σχεδίαση διότι διάφορες

παραλλαγές για συγκεκριμένα είδη συσσωμάτωσης πρέπει να εξεταστούν και να επιλεγούν. Επιπλέον, οι πίνακες διαστάσεων εξακολουθούν να είναι μεγάλοι.



Εικόνα 8: Σχήμα αστερισμού

1.9 Υποστήριξη αποφάσεων: Ανάλυση των δεδομένων αποθήκης

Η θεμελιώδης λειτουργία μιας αποθήκης δεδομένων είναι η στέγαση δεδομένων για την υποστήριξη αποφάσεων. Τα δεδομένα αντιγράφονται από την αποθήκη δεδομένων ώστε να αναλυθούν από το σύστημα υποστήριξης αποφάσεων και παράλληλα εισάγονται δεδομένα στην αποθήκη δεδομένων από το περιβάλλον αποφάσεων (Εικόνα 3). Όσα κι όποια δεδομένα εισάγονται στην αποθήκη από το σύστημα υποστήριξης αποφάσεων θα έχουν τη μορφή μεταδεδομένων που προκύπτουν από μία ή περισσότερες διαδικασίες υποστήριξης αποφάσεων. Υπάρχουν τρεις κατηγορίες υποστήριξης αποφάσεων:

1. Αναφορά δεδομένων

Η αναφορά θεωρείται το χαμηλότερο επίπεδο υποστήριξης αποφάσεων. Όμως ένα υποσύστημα αναφορών με δυνατότητα δημιουργίας ενημερωτικών αναφορών για πλειάδα διαφορετικών πελατών έχει ύψιστη σημασία για την επιτυχή λειτουργία οποιασδήποτε επιχείρησης.

2. Ανάλυση δεδομένων

Η ανάλυση δεδομένων επιτυγχάνεται με κάποιο εργαλείο ανάλυσης πολυδιάστατων δεδομένων.

3. Ανακάλυψη γνώσης

Η ανακάλυψη γνώσης συνήθως γίνεται μέσω της εξόρυξης πληροφορίας. Όμως, μερικές φορές, και χειροκίνητες τεχνικές εξόρυξης πληροφορίας που συνεπάγονται την επαναλαμβανόμενη υποβολή ερωτημάτων και ανάλυση δεδομένων μπορεί να αποκαλύψουν ενδιαφέρουσες τάσεις στα δεδομένα.

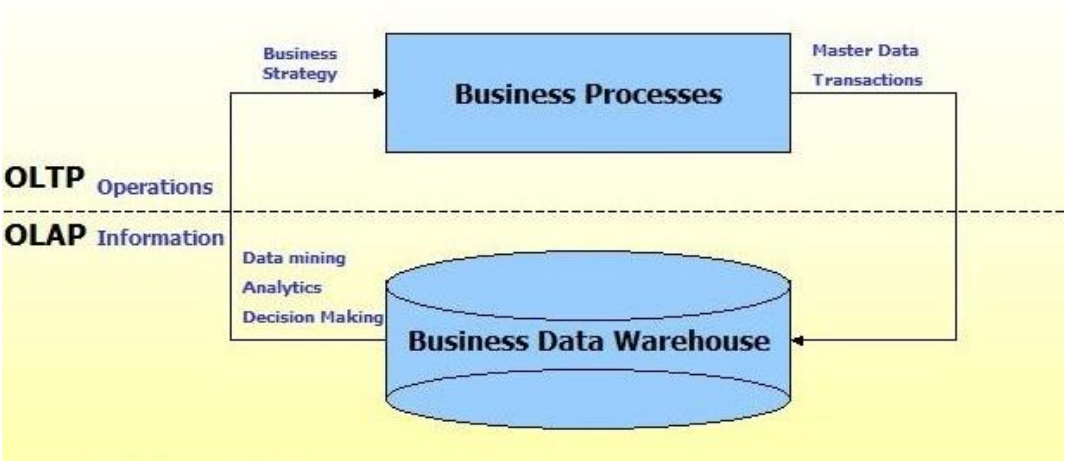
Εκτός από την αποθήκευση δεδομένων για την υποστήριξη αποφάσεων, η αποθήκη είναι ένα μέσο για τη δημιουργία μικρότερων τμηματικών αποθηκών, οι οποίες ονομάζονται εξαρτημένα κέντρα δεδομένων (dependent data marts). Συνήθως, ένα εξαρτημένο κέντρο δεδομένων αναφέρεται σε ένα θέμα μόνο και είναι σχεδιασμένο για συγκεκριμένο σκοπό. Επιπλέον, το κέντρο αυτό είναι πιθανό να περιέχει συνοπτικές πληροφορίες με χαμηλότερο βαθμό λεπτομέρειας από αυτόν της αποθήκης δεδομένων.

1.10 Διαφορές OLTP – Αποθήκης Δεδομένων

Τα πληροφοριακά συστήματα Βάσεων Δεδομένων μπορούν να χωριστούν σε συστήματα συναλλαγών (OLTP, On-line Transaction Processing) και συστήματα ανάλυσης (OLAP). Γενικά, μπορεί να θεωρηθεί ότι τα συστήματα OLTP παρέχουν δεδομένα στις αποθήκες δεδομένων, ενώ τα συστήματα OLAP βοηθούν στην ανάλυση των δεδομένων..
Υπάρχουν σημαντικές διαφορές στη δομή και το σκοπό μιας βάσης δεδομένων OLTP από μία αποθήκης δεδομένων. Οι πιο σημαντικές από τις διαφορές περιγράφονται παρακάτω:

- Τα δεδομένα σε μία αποθήκη είναι θεματικά και βασίζονται σε ένα ή περισσότερα κεντρικά θέματα. Οι βάσεις δεδομένων OLTP προσανατολίζονται στις συναλλαγές και είναι οργανωμένες με τέτοιο τρόπο ώστε να βελτιστοποιούν την ενημέρωση και την ανάκτηση δεδομένων.
- Τα δεδομένα της αποθήκης είναι αποκανονικοποιημένα και ολοκληρωμένα (ενοποιημένα), ενώ τα δεδομένα μιας βάσης δεδομένων OLTP είναι κανονικοποιημένα και διαχωρισμένα.
- Στην αποθήκη δεδομένων, τα δεδομένα φυλάσσονται για την παραγωγή αναφορών, την ανάλυση και τον έλεγχο. Σε ένα σύστημα OLTP υποστηρίζεται η επεξεργασία, η συλλογή και η διαχείριση δεδομένων.
- Η άμεση επεξεργασία συναλλαγών ασχολείται με τα δεδομένα που είναι απαραίτητα για την αποτελεσματική καθημερινή λειτουργία μιας επιχείρησης ή ενός οργανισμού. Οι εγγραφές δεδομένων μιας βάσης δεδομένων συναλλαγών είναι διαθέσιμες για πολλαπλή πρόσβαση και συνεχή ενημέρωση. Αντίθετα, η ύπαρξη των δεδομένων μιας αποθήκης οφείλεται κατά ένα μέρος στο ότι δεν είναι πλέον χρήσιμα στο περιβάλλον του OLTP. Τα περισσότερα από τα δεδομένα μιας αποθήκης είναι ιστορικά, χρονοσημασμένα, και δεν υπόκειται σε αλλαγές (είναι μόνο για ανάγνωση).
- Ο βαθμός λεπτομέρειας είναι ο όρος που χρησιμοποιείται για την περιγραφή του επιπέδου λεπτομέρειας των αποθηκευμένων πληροφοριών. Τα επιχειρησιακά δεδομένα αντιπροσωπεύουν το χαμηλότερο επίπεδο λεπτομέρειας καθώς κάθε στοιχείο δεδομένων έχει πληροφορίες για μία και μόνη συναλλαγή. Το επίπεδο λεπτομέρειας για τα δεδομένα που βρίσκονται

σε μια αποθήκη είναι θέμα σχεδίασης που εξαρτάται από τις επιθυμίες του χρήστη, αλλά και από το πλήθος των δεδομένων που συγκεντρώνονται.



Εικόνα 9: Σχέση OLAP-OLTP

Οι διαφορές συνοψίζονται στον παρακάτω πίνακα:

Χαρακτηριστικό	**Συστήματα OLTP Σχεσιακές Βάσεις Δεδομένων**	**Συστήματα OLAP Αποθήκες Δεδομένων**
Στόχος	Έλεγχος και λειτουργία των θεμελιωδών εργασιών των επιχειρήσεων.	Ο προγραμματισμός κι η επίλυση προβλημάτων, καθώς και η υποστήριξη λήψης αποφάσεων
Όγκος εργασίας	Προκαθορισμένες συναλλαγές	Συγκεκριμένα ερωτήματα ανάλυσης
Ερωτήματα	Σχετικά τυποποιημένα και απλά ερωτήματα. Επιστρέφονται λίγες εγγραφές.	Πολύπλοκα ερωτήματα που αφορούν συναθροίσεις.
Χρήστες	Απλοί εργαζόμενοι (π.χ., ταμίες),	Υψηλόβαθμα στελέχη, αναλυτές
Αριθμός Χρηστών	Χιλιάδες	Εκατοντάδες
Προσπέλαση	Για εκατοντάδες εγγραφές, ανάγνωση κι εγγραφή	Για εκατομμύρια εγγραφές, κυρίως ανάγνωση
Δεδομένα	Λεπτομερή, τόσο αριθμητικά	Συνοπτικά,

	όσο και αλφαριθμητικά	κυρίως αριθμητικά
Πηγή δεδομένων	Σχεσιακά δεδομένα, OI βάσεις δεδομένων OLTP, είναι η αρχική πηγή δεδομένων.	Ενοποιημένα δεδομένα. Τα δεδομένα OLAP προέρχονται από διάφορες βάσεις δεδομένων OLTP.
Χρονική Κάλυψη	Μόνο τρέχοντα δεδομένα	Τρέχοντα και ιστορικά δεδομένα
Ενοποίηση Δεδομένων	Με βάση την εφαρμογή	Με βάση το θέμα
Ενημέρωση	Συνεχής	Περιοδική
Μοντέλο	Κανονικοποιημένο, με πολλούς πίνακες.	Αποκανονικοποιημένο, πολυδιάστατο. Χρήση σχήματος αστέρα και/ή χιονοστιβάδας.
Απαιτούμενος χώρος	Σχετικά μικρός (100 MB – 100 GB)	Μεγαλύτερος λόγω της ύπαρξης των δομών συσσωμάτωσης και των ιστορικών δεδομένων. Απαιτούνται περισσότερα ευρετήρια από ότι στην OLTP (100GB – TB).
Ταχύτητα επεξεργασίας	Συνήθως πολύ γρήγορα (msec ή sec)	Εξαρτάται από τον όγκο των δεδομένων. Μεγάλος όγκος δεδομένων και πολύπλοκα ερωτήματα μπορεί να χρειαστούν ώρες. Η ταχύτητα των ερωτημάτων μπορεί να βελτιωθεί με τη δημιουργία ευρετηρίων.
Κατάλογοι	B – δένδρα	Κατάλογοι bitmap

Πίνακας 1: Διαφορές Σχεσιακών ΒΔ - Αποθηκών Δεδομένων

Επίλογος

Στο κεφάλαιο αυτό είδαμε πως ορίζεται η έννοια της αποθήκης δεδομένων, πως μπορεί να κατασκευαστεί μία αποθήκη και ποιες οι διαφορές της από τις παραδοσιακές βάσεις δεδομένων. Στο επόμενο κεφάλαιο, θα δούμε την τεχνολογία OLAP, η οποία αναπτύχθηκε για την ανάκτηση πληροφορίας από αποθήκες δεδομένων.

Κεφάλαιο 2: On-Line Analytical Processing (OLAP)

Εισαγωγή

Η αναλυτική επεξεργασία άμεσης επικοινωνίας είναι μια μεθοδολογία που χρησιμοποιείται για τη δυναμική ανάλυση κι επεξεργασία μεγάλου όγκου δεδομένων σε πολυδιάστατο περιβάλλον.

Στο κεφάλαιο αυτό παρουσιάζονται κάποια γενικά στοιχεία για την OLAP, το πώς ορίζονται οι εφαρμογές OLAP σύμφωνα με το FASMI test, ο κύβος OLAP και οι πράξεις που αφορούν τον κύβο, καθώς κι οι υποκατηγορίες OLAP. Τέλος, γίνεται μια σύντομη αναφορά στη γλώσσα ερωτημάτων MDX.

2.1 Γενικά

Η OLAP θεωρείται ως ένας από τους καλύτερους τρόπους για να αξιοποιηθούν οι πληροφορίες μιας αποθήκης δεδομένων. Δίνει την δυνατότητα στους τελικούς χρήστες, των οποίων η ανάλυση των αναγκών δεν είναι εύκολο να καθοριστεί εκ των προτέρων, να αναλύσουν τα δεδομένα και να τα εξερευνήσουν διαδραστικά με βάση το πολυδιάστατο μοντέλο. Αποτελεί ένα χρήσιμο εργαλείο για την επιβεβαίωση ή τη διάψευση υποθέσεων που διατυπώνονται από ανθρώπους και για την εκτέλεση μη αυτόματης εξόρυξης πληροφορίας. Μία διαδικασία OLAP αποτελείται από μία διαδρομή πλοήγησης, που αντιστοιχεί σε μια διαδικασία ανάλυσης γεγονότων από διαφορετικές οπτικές γωνίες και σε διαφορετικά επίπεδα λεπτομέρειας. Για την επεξεργασία των πληροφοριών βάσης δεδομένων με τη χρήση της OLAP, απαιτείται ένας OLAP server για να οργανώσει και να συγκρίνει τις πληροφορίες. Οι clients μπορούν να αναλύσουν διαφορετικά σύνολα δεδομένων χρησιμοποιώντας λειτουργίες ενσωματωμένες στον OLAP server. Εξαιτίας των ισχυρών δυνατοτήτων ανάλυσης δεδομένων, η διαδικασία OLAP χρησιμοποιείται συχνά για εξόρυξη δεδομένων, η οποία έχει ως στόχο να ανακαλυφθούν νέες σχέσεις μεταξύ διαφορετικών σετ δεδομένων.

Ο όρος OLAP δημιουργήθηκε από μία μικρή τροποποίηση του παραδοσιακού όρου των βάσεων δεδομένων OLTP (On-Line Transaction Processing). Εφαρμογές OLAP στις επιχειρήσεις αποτελούν διάφορες αναφορές σχετικά με πωλήσεις, marketing, προϋπολογισμούς κ.ο.κ.

Ο όρος χρησιμοποιήθηκε για πρώτη φορά από το Βρετανό επιστήμονα Edgar F. Code, ο οποίος το 1993 έγραψε τους «12 νόμους της αναλυτικής επεξεργασίας άμεσης επικοινωνίας».

Παλαιότερα, ένα ερώτημα σε ένα σύστημα διαχείρισης μιας σχεσιακής βάσης δεδομένων έδινε στον χρήστη, που ήταν υπεύθυνος για τη λήψη αποφάσεων, την απάντηση για το «τι» είχε συμβεί, αλλά όχι το «γιατί». Γι' αυτό δόθηκε έμφαση στο να ανακαλυφθεί το «γιατί». Το «τι» δεν ικανοποιούσε τις απαιτήσεις γιατί, έδινε την ευκαιρία για λύσεις μόνο σε επιφανειακά προβλήματα του σήμερα, ενώ το «γιατί» βοηθάει στη λήψη αποφάσεων στη ρίζα των προβλημάτων και στην πρόληψη μελλοντικών περιστατικών.

2.2 FASMI test

Το FASMI τεστ χρησιμοποιείται για να ορίσει και να προσδιορίσει τα χαρακτηριστικά που θα πρέπει να έχει μια εφαρμογή OLAP.
Ο ορισμός συνοψίζεται σε πέντε λέξεις-κλειδιά: Fast Analysis of Shared Multidimensional Information.

FAST. Το σύστημα στοχεύει να δίνει απαντήσεις στους χρήστες, όσο το δυνατόν ταχύτερα, σε αρκετές περιπτώσεις, μέσα σε δευτερόλεπτα.

ANALYSIS. Το σύστημα μπορεί να αντιμετωπίσει οποιαδήποτε επιχειρηματική λογική και στατιστική ανάλυση που έχει σημασία για την εφαρμογή και τον χρήστη.

SHARED. Το σύστημα υλοποιεί όλες τις απαιτήσεις ασφαλείας για την εμπιστευτικότητα και, αν η πολλαπλή πρόσβαση εγγραφής είναι απαραίτητη, κλειδώνει τις ενημερώσεις στο κατάλληλο επίπεδο. Το σύστημα θα πρέπει να είναι σε θέση να χειριστεί πολλαπλές ενημερώσεις έγκαιρα και με ασφαλή τρόπο.

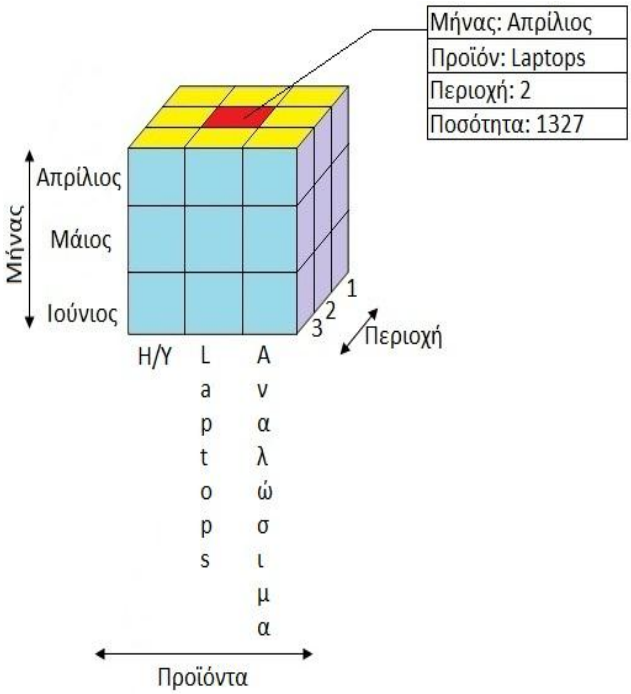
MULTIDIMENSIONAL. Η έννοια-κλειδί. Αν θα έπρεπε να δοθεί ένας μονολεκτικός ορισμός στην OLAP, θα ήταν αυτός. Το σύστημα πρέπει να παρέχει μια πολυδιάστατη εννοιολογική άποψη των δεδομένων, συμπεριλαμβανομένης της πλήρους υποστήριξης για ιεραρχίες και πολλαπλές ιεραρχίες, καθώς αυτός είναι ο πιο λογικός τρόπος για να αναλυθούν επιχειρήσεις και οργανισμοί.

INFORMATION. Όλα τα δεδομένα και οι συμπληρωματικές πληροφορίες που χρειάζονται, όπου κι αν βρίσκονται κι όσο σχετικές κι αν είναι για την εφαρμογή. Η χωρητικότητα μιας εφαρμογής μετριέται σε σχέση με την ποσότητα των δεδομένων που μπορεί να χειριστεί κι όχι με το πόσα gigabyte μπορεί να αποθηκεύσει.

2.3 Ο κύβος OLAP

Ο πυρήνας ενός συστήματος OLAP είναι ο κύβος OLAP. Πρόκειται για μια δομή δεδομένων που επιτρέπει τη γρήγορη ανάλυση των δεδομένων. Μπορεί επίσης να οριστεί ως η δυνατότητα του χειρισμού και της ανάλυσης δεδομένων από διαφορετικές οπτικές γωνίες. Η διάταξη των δεδομένων μέσα στον κύβο ξεπερνά ορισμένους περιορισμούς των σχεσιακών βάσεων δεδομένων
Μιας και ο κάθε κύβος OLAP είναι σχεδιασμένος για ένα συγκεκριμένο σκοπό, δεν είναι ασυνήθιστο να έχουμε πολλούς κύβους δομημένους από τα δεδομένα μιας μόνο αποθήκης. Στη σχεδίαση ενός κύβου δεδομένων συμπεριλαμβάνονται αποφάσεις σχετικά με το ποια χαρακτηριστικά θα περιληφθούν στον κύβο, καθώς επίσης και για το βαθμό λεπτομέρειας του κάθε χαρακτηριστικού. Ένας καλοσχεδιασμένος κύβος δομείται έτσι ώστε να αποφεύγονται περιπτώσεις στις οποίες τα κελιά δεδομένων δεν περιέχουν χρήσιμες πληροφορίες. Για παράδειγμα, ένας κύβος με δύο διαστάσεις χρόνου, μία για το μήνα και μία ακόμα για το οικονομικό τρίμηνο (Q1, Q2, Q3, Q4), αποτελεί κακή επιλογή επειδή οι συνδυασμοί (Ιανουάριος, Q4) ή (Δεκέμβριος, Q1), θα είναι πάντα κενοί. Η αναπαράσταση με κύβο δεδομένων διευκολύνει πολύ την ανάλυση, επειδή η χρήση ενός κύβου μοιάζει με τη χρήση ενός φύλλου δεδομένων.
Η οπτικοποίηση ενός κύβου τριών διαστάσεων είναι εύκολη. Υπάρχουν, όμως, και κύβοι με περισσότερες διαστάσεις, οι οποίοι χαρακτηρίζονται ως υπερκύβοι. Για να οπτικοποιήσουμε έναν κύβο με τέσσερις, για παράδειγμα, διαστάσεις, θεωρούμε *n* τρισδιάστατους κύβους, όπου *n* είναι το σύνολο των πιθανών τιμών που παίρνει η τέταρτη διάσταση.
Θα δούμε ένα παράδειγμα κύβου, τριών διαστάσεων. Οι τρεις διαστάσεις του κύβου του παραδείγματος, είναι Προϊόν, Περιοχή και Χρόνος. Δεδομένου ενός προϊόντος, μιας τοποθεσίας, και μιας τιμής χρόνου, προκύπτει το πολύ μία τιμή πωλήσεων.

Το τετράγωνο με κόκκινο χρώμα δείχνει σε έναν κύβο που περιέχει το συνολικό αριθμό laptops που πωλήθηκαν στην περιοχή 2 το μήνα Απρίλιο.



Εικόνα 10: Ο κύβος OLAP

Ο συνολικός αυτός αριθμός είναι μία αριθμητική ποσότητα, η οποία ονομάζεται αριθμητική μέτρηση (measure). Μία μέτρηση (measure) είναι μία αριθμητική τιμή που αναπαριστά τη συνάθροιση των γραμμών δεδομένων στον πίνακα γεγονότων (fact table). Για παράδειγμα, το sum (για πωλήσεις) και το count (για ποσότητες) είναι μετρήσεις. Η σχέση που σχετίζει τις διαστάσεις με την μέτρηση που μας ενδιαφέρει ονομάζεται πίνακας γεγονότων.

Κάθε διάσταση ενός κύβου OLAP μπορεί να έχει μία ή περισσότερες σχετιζόμενες ιεραρχίες εννοιών. Κάθε ιεραρχία εννοιών ορίζει μια αντίστοιχη η οποία επιτρέπει την προβολή της αντίστοιχης διάστασης από διάφορα επίπεδα λεπτομέρειας. Μπορούμε, δηλαδή, να αλλάζουμε το επίπεδο της ιεραρχίας σε κάθε διάσταση, έτσι ώστε να προκύπτει διαφορετική όψη του κύβου δεδομένων. Για παράδειγμα, μπορούμε να δούμε τις προϊόντα, αντί ανά περιοχή, ανά πόλη. Στην Εικόνα 9 εμφανίζεται μια ιεραρχία εννοιών για το χαρακτηριστικό *Τοποθεσία*. Η έννοια της *Περιοχής* είναι η πιο γενική μέσα στην ιεραρχία. Στο δεύτερο επίπεδο της ιεραρχίας φαίνεται ότι μια περιοχή αποτελείται από έναν ή περισσότερους *Νομούς*. Το τρίτο και τέταρτο επίπεδο μας δείχνουν ότι σε ένα νομό ανήκουν μία ή περισσότερες *Πόλεις* και σε μια πόλη υπάρχουν μία ή περισσότερες *Διευθύνσεις*. Η ιεραρχία δείχνει ότι κάθε νομός περιέχεται πλήρως μέσα σε μία, και μόνο μία, περιοχή, και επίσης ότι κάθε πόλη είναι μέρος ακριβώς ενός νομού.

Εικόνα 11: Ιεραρχία "Τοποθεσία"

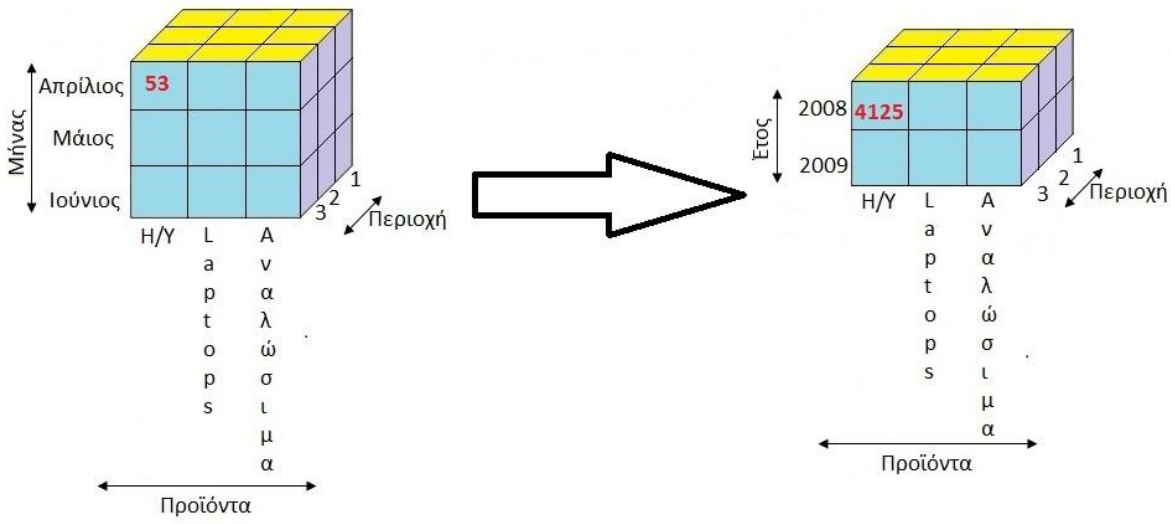
2.4 Πράξεις OLAP

Η τεχνολογία OLAP βοηθάει στην εύκολη διατύπωση αναλυτικών ερωτήσεων σε κύβους δεδομένων, καθώς και στη γρήγορη εκτέλεση τους.
Η OLAP, περιλαμβάνει τις παρακάτω βασικές αναλυτικές λειτουργίες:

- Roll-up
- Drill-down
- Slice
- Dice
- Pivot

Roll up (Σύμπτυξη)

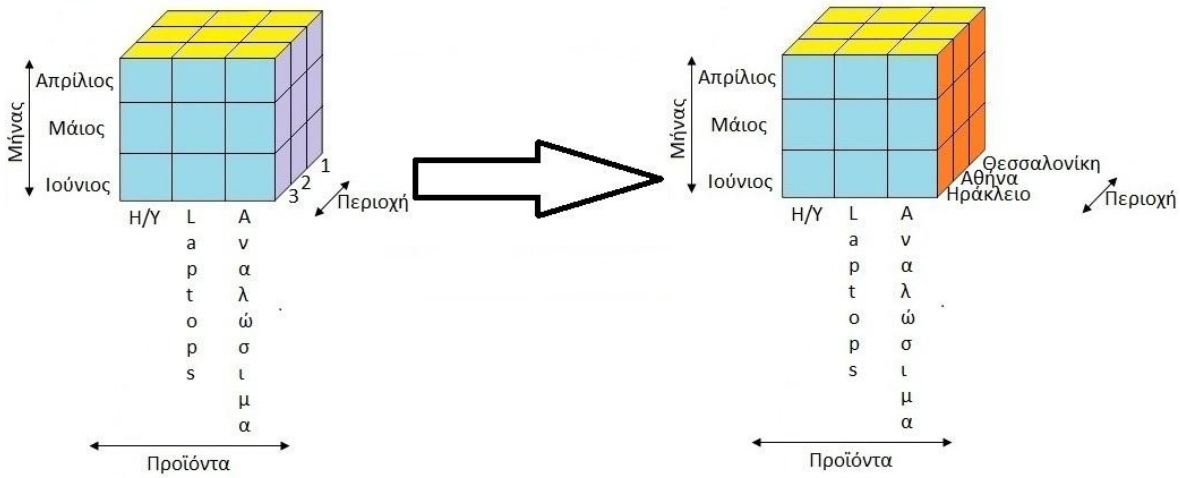
Η σύμπτυξη περιλαμβάνει την ομαδοποίηση των δεδομένων που μπορούν να συγκεντρωθούν και να υπολογιστούν σε μία ή περισσότερες διαστάσεις. Στην ουσία, παράγει ένα κύβο δεδομένων με μειωμένο επίπεδο λεπτομέρειας είτε με την επιλογή ανώτερου επιπέδου ιεραρχίας σε κάποιες διαστάσεις είτε με την αφαίρεση κάποιων διαστάσεων. Για παράδειγμα, στον κύβο του παραδείγματος, σύμπτυξη θα είχαμε είτε αν προβάλαμε έναν κύβο που αντί για μήνες θα είχε έτη, είτε αν αφαιρούσαμε μία από τις τρεις διαστάσεις. Αν αφαιρέσουμε τη διάσταση της τοποθεσίας, το τελικό αποτέλεσμα θα είναι ένα υπολογιστικό φύλλο που θα δείχνει τις πωλήσεις ανά μήνα και τύπο κατηγορίας.



Εικόνα 12: Η πράξη της σύμπτυξης (roll-up)

Drill-down (Ανάπτυξη)

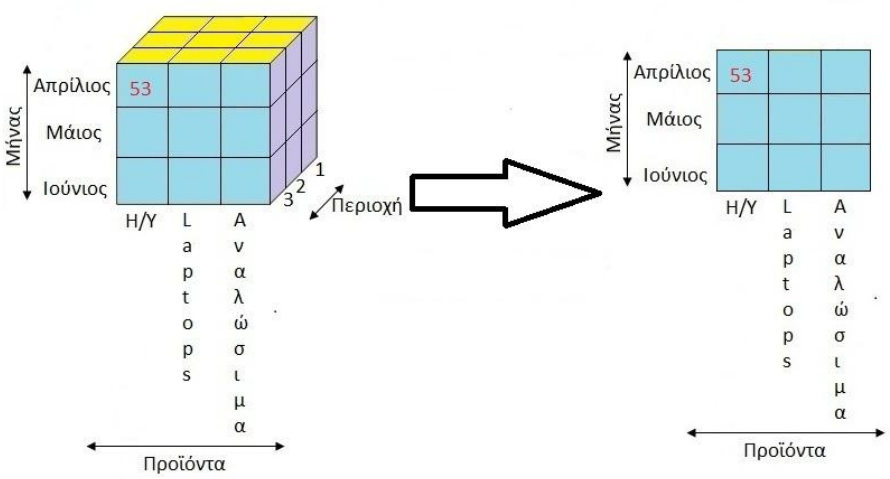
Το drill-down είναι μία τεχνική που επιτρέπει στους χρήστες να περιηγηθούν στις λεπτομέρειες. Στην ουσία, παράγει ένα κύβο δεδομένων με αυξημένο επίπεδο λεπτομέρειας είτε με την επιλογή κατώτερου επιπέδου ιεραρχίας σε κάποιες διαστάσεις είτε με την προσθήκη διαστάσεων. Στο παράδειγμα που χρησιμοποιούμε ανάπτυξη θα είχαμε αν στον κύβο προβάλουμε τη διάσταση Περιοχή σε επίπεδο Νομού ή Πόλης.



Εικόνα 13: Η πράξη της ανάπτυξης (drill-down)

Slice (Κόψιμο σε φέτες)

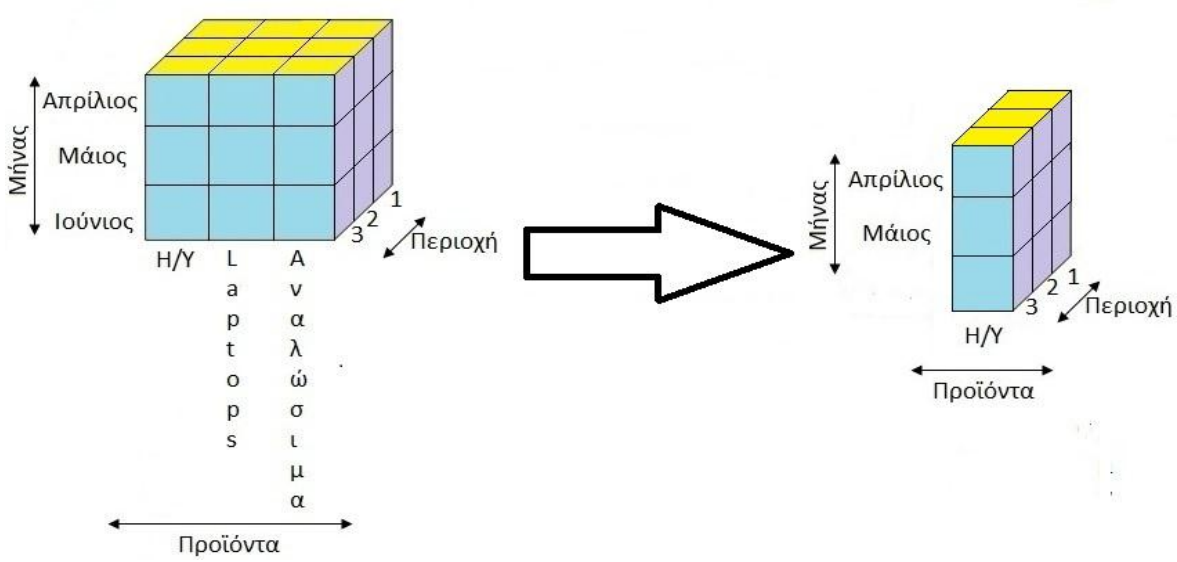
Παράγει υποκύβο με την επιλογή δεδομένων σε μία διάσταση ενός κύβου OLAP. Σε έναν τρισδιάστατο κύβο, η πράξη τεμαχισμού σε φέτες θα άφηνε δύο από τις τρεις διαστάσεις άθικτες, και η επιλογή δεδομένων στη διάσταση που απομένει θα δημιουργούσε έναν υποκύβο του αρχικού. Ένα ερώτημα για την πράξη slice, στον κύβο του παραδείγματος, θα μπορούσε να ζητάει να παραχθεί υπολογιστικό φύλλο με πληροφορίες για το μήνα και την περιοχή για όλα τα κελιά που αφορούν τα laptops.



Εικόνα 14: Η πράξη του κοψίματος σε φέτες (slice)

Dice (Τεμαχισμός σε κύβους)

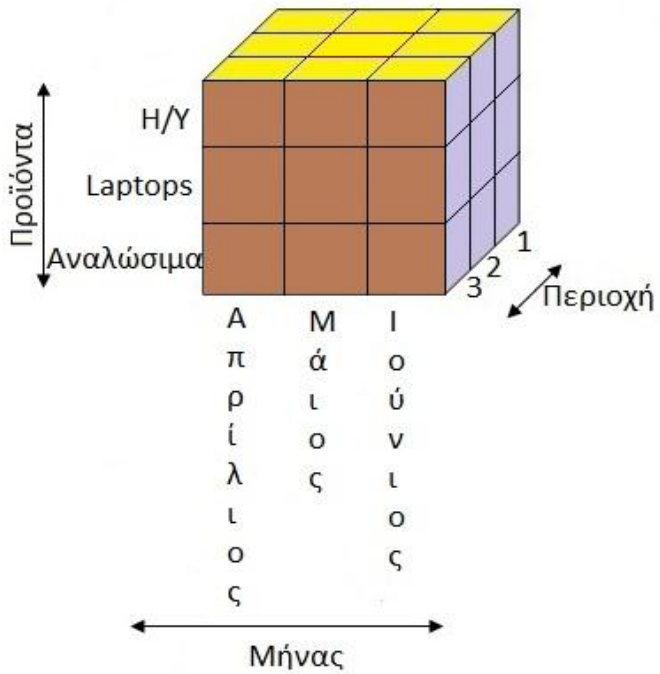
Εξάγει έναν υποκύβο από τον αρχικό κύβο, εκτελώντας μια πράξη επιλογής σε δύο ή περισσότερες διαστάσεις. Ένα ερώτημα για την πράξη dice, στον κύβο του παραδείγματος, θα μπορούσε τον προσδιορισμό τον μήνα με τις λιγότερες πωλήσεις Η/Υ για κάθε περιοχή.



Εικόνα 15: Η πράξη του τεμαχισμού σε κύβους (dice)

Pivot (Περιστροφή)

Παράγει κύβο με άλλη διάταξη των διαστάσεων. Στην εικόνα 3, βλέπουμε τον κύβο του παραδείγματός μας , στον οποίο έχουμε εφαρμόσει την πράξη της περιστροφής (στον οριζόντιο άξονα υπάρχει πλέον η διάσταση του χρόνου, και στον κάθετο, αυτή των προϊόντων).



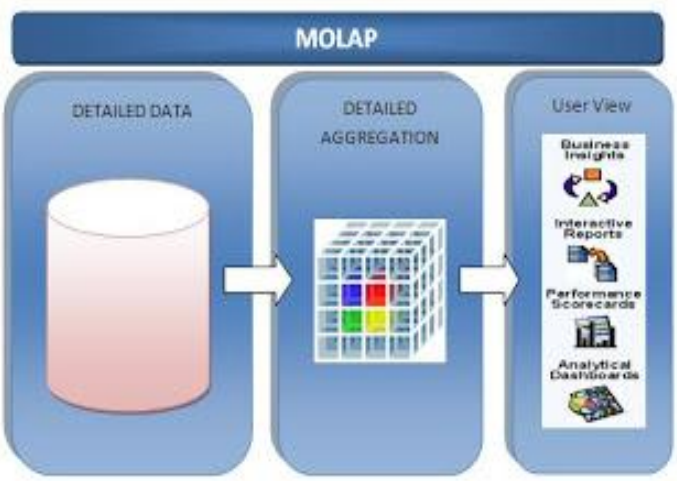
Εικόνα 16: Η πράξη της περιστροφής (pivot)

2.5 Τύποι Συστημάτων OLAP

Έχουν αναπτυχθεί διάφοροι τύποι συστημάτων OLAP όσον αφορά το φυσικό επίπεδο. Οι τρεις κυριότεροι τύποι είναι οι Multidimensional OLAP (MOLAP), Relational OLAP (ROLAP), και Hybrid OLAP (HOLAP). Υπάρχουν ακόμα οι Database OLAP και Web OLAP.

MOLAP

Αναφέρεται στην πολυδιάστατη αναλυτική επεξεργασία άμεσης επικοινωνίας. Αυτός είναι ο πιο παραδοσιακός τρόπος ανάλυσης OLAP. Στην περίπτωση αυτή το σύστημα OLAP υλοποιείται μέσω μιας εξειδικευμένης πολυδιάστατης βάσης δεδομένων.



Εικόνα 17: Σύστημα MOLAP

Μεγάλοι πολυδιάστατοι πίνακες σχηματίζουν τη δομή αποθήκευσης. Για παράδειγμα, για να αποθηκεύσουμε τον αριθμό των πωλήσεων 500 μονάδων για το προϊόν Α, τον μήνα Απρίλιο, στην περιοχή Π1, ο αριθμός των πωλήσεων 500 αποθηκεύεται σε έναν πίνακα που αντιπροσωπεύεται από τις τιμές (Προϊόν Α, Απρίλιος, Π1). Οι τιμές του πίνακα προσδιορίζουν τη θέση των κελιών. Αυτά τα κελιά είναι τομές της αξίας των χαρακτηριστικών της διάστασης. Αν παρατηρήσουμε πώς σχηματίζονται τα κελιά, θα διαπιστώσουμε ότι δεν περιέχουν όλα τα κελιά τιμές μετρήσεων. Για παράδειγμα αν ένα κατάστημα είναι κλειστό τις Κυριακές, τότε τα κελιά που αντιπροσωπεύουν τις Κυριακές θα είναι όλα null. Οι πίνακες που χρησιμοποιούνται σε αυτήν την περίπτωση δεν έχουν την ίδια έννοια με τους πίνακες στις γλώσσες προγραμματισμού. Στις γλώσσες προγραμματισμού οι πίνακες αποθηκεύονται στην κυρία μνήμη, ενώ στο σύστημα MOLAP αυτό μπορεί να μην είναι δυνατό εξ αιτίας του μεγέθους των πινάκων. Για αυτούς τους λόγους, χρησιμοποιούνται βελτιστοποιημένες δομές

πινάκων, οι οποίες κάνουν συμπίεση (εκμεταλλευόμενες την αραιότητα) και αποθήκευση στο δίσκο (βασισμένες στον κατακερματισμό).

Πλεονεκτήματα των συστημάτων MOLAP είναι:

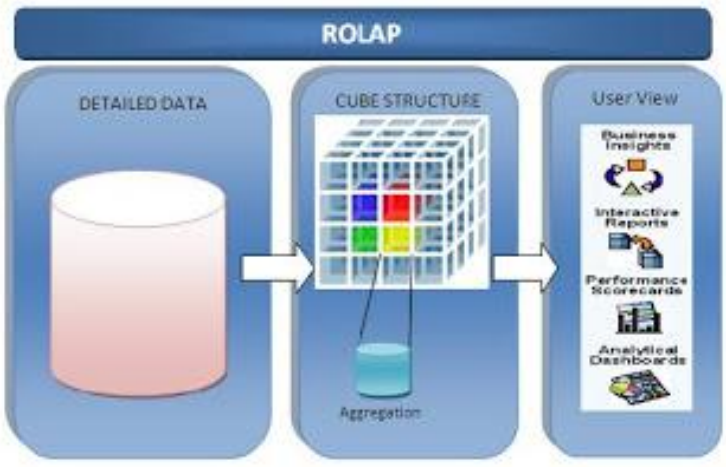
- Η πολύ γρήγορη απόδοση ερωτημάτων OLAP, Οι κύβοι MOLAP έχουν κατασκευαστεί για γρήγορη ανάκτηση δεδομένων, και είναι η βέλτιστη λύση για τις πράξεις του τεμαχισμό σε κύβους και κόψιμο σε φέτες.
- Ο μειωμένος αποθηκευτικός χώρος, λόγω της συμπίεσης. Ειδικά για λίγες διαστάσεις, η μείωση είναι δραστική.

Μειονεκτήματα των συστημάτων MOLAP είναι:

- Μπορεί να χειριστεί περιορισμένη ποσότητα δεδομένων. Επειδή όλοι οι υπολογισμοί εκτελούνται όταν ο κύβος έχει ήδη δημιουργηθεί, δεν είναι δυνατόν ο κύβος να περιλαμβάνει μια μεγάλη ποσότητα δεδομένων. Αυτό δεν σημαίνει ότι τα δεδομένα στον κύβο δεν μπορούν να προέρχονται από μία μεγάλη ποσότητα δεδομένων. Πράγματι, αυτό είναι δυνατό. Αλλά σε αυτή την περίπτωση, μόνο συνοπτικές πληροφορίες μπορούν να περιλαμβάνονται στον κύβο.
- Δεν υπάρχουν πρότυπα για την ανάπτυξη MOLAP συστημάτων. Ακόμα, δεν έχουν αναπτυχθεί κοινά αποδεκτές γλώσσες και μοντέλα για τη διατύπωση OLAP πράξεων σε συστήματα MOLAP.

ROLAP

Αναφέρεται στην σχεσιακή αναλυτική επεξεργασία άμεσης επικοινωνίας. Στην περίπτωση αυτή το σύστημα OLAP χρησιμοποιεί μια σχεσιακή Βάσης Δεδομένων για την αποθήκευση και διαχείριση του κύβου. Ένας πίνακας, με την έννοια των σχεσιακών Συστημάτων Διαχείρισης Βάσης Δεδομένων, αποθηκεύει τον πίνακα γεγονότων, και ξεχωριστοί



Εικόνα 18: Σύστημα ROLAP

πίνακες αποθηκεύουν τις διαστάσεις. Η ROLAP τείνει να χρησιμοποιείται για δεδομένα που έχουν ένα μεγάλο αριθμό χαρακτηριστικών και δεν μπορούν εύκολα να τοποθετηθούν σε μια δομή κύβου.

Πλεονεκτήματα των συστημάτων ROLAP είναι:

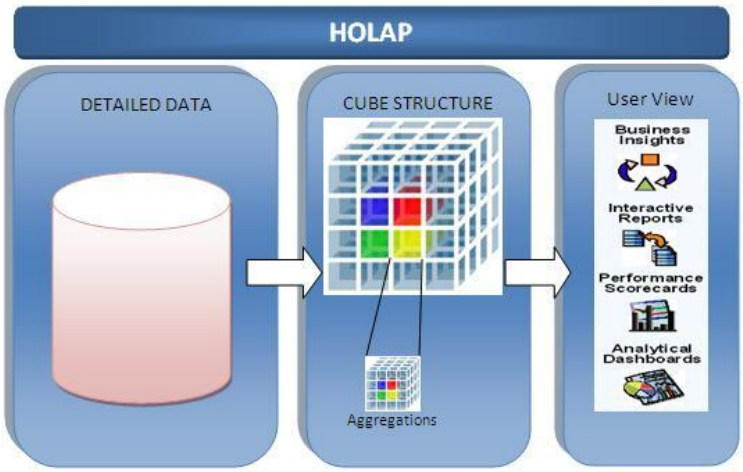
- Δεν απαιτείται η δημιουργία του κύβου κι έτσι μπορεί να χειριστεί μεγάλες ποσότητες δεδομένων. Ο περιορισμός του μεγέθους των δεδομένων στα συστήματα ROLAP είναι ο περιορισμός για το μέγεθος των δεδομένων που υπάρχει στις σχεσιακές βάσεις δεδομένων. Με άλλα λόγια, η ίδια ROLAP δεν θέτει κανένα περιορισμό για ποσότητα δεδομένων.
- Βασίζονται στην ενυπάρχουσα σχεσιακή τεχνολογία, για την οποία υπάρχουν πρότυπα. Επίσης, είναι ευκολότερη η διάχυση τους σε όσους έχουν εξοικείωση με τα σχεσιακά Συστήματα Διαχείρισης Βάσεων Δεδομένων, δηλαδή με αρκετό κόσμο.

Μειονεκτήματα των συστημάτων ROLAP είναι:

- Η μειωμένη ταχύτητα, συγκριτικά με τα συστήματα MOLAP, κατά την εκτέλεση των OLAP πράξεων. Επειδή κάθε αναφορά ROLAP είναι ουσιαστικά ένα ερώτημα SQL (ή πολλαπλά ερωτήματα SQL) στη σχεσιακή βάση δεδομένων, ο χρόνος απόκρισης του ερωτήματος μπορεί να είναι μεγάλος εάν το βασικό μέγεθος των δεδομένων είναι μεγάλο.
- Περιορίζεται από τις λειτουργίες SQL. Τα συστήματα ROLAP βασίζονται, κυρίως, στη δημιουργία SQL προτάσεων για ερωτήματα στη σχεσιακή βάση δεδομένων. Επειδή οι προτάσεις SQL δεν ταιριάζουν σε όλες τις ανάγκες (για παράδειγμα, είναι δύσκολο να εκτελεστούν πολύπλοκοι υπολογισμοί, χρησιμοποιώντας SQL), τα συστήματα ROLAP περιορίζονται από το τι μπορεί να κάνει η SQL.

HOLAP

Τα σύστημα HOLAP επιχειρούν να συνδυάσουν τα πλεονεκτήματα και τα χαρακτηριστικά των ROLAP και MOLAP συστημάτων, αποφεύγοντας τα μειονεκτήματα τους. Τα συστήματα HOLAP



Εικόνα 19: Σύστημα HOLAP

επιτρέπουν την αποθήκευση μέρος των δεδομένων όπως στην περίπτωση των συστημάτων MOLAP, και το υπόλοιπο μέρος των δεδομένων αποθηκεύεται όπως στην περίπτωση των συστημάτων ROLAP. Στην πρώτη κατηγορία αποθηκεύουμε το τμήμα του κύβου που αναλύεται συχνότερα, επιταχύνοντας έτσι τους χρόνους εκτέλεσης των OLAP πράξεων. Εφόσον ο υπόλοιπος κύβος βρίσκεται αποθηκευμένος σε ένα σχεσιακό Συστημάτων Διαχείρισης Βάσης Δεδομένων, επιτυγχάνουμε κλιμάκωση σε μεγάλους όγκους δεδομένων.

Τα περισσότερα εμπορικά συστήματα, σήμερα, είναι HOLAP.

Database OLAP

Αναφέρεται σε ένα σχεσιακό Σύστημα Διαχείρισης Βάσεων Δεδομένων που έχει σχεδιαστεί για να υποστηρίξει τις δομές OLAP και για να εκτελέσει υπολογισμούς OLAP.

Web OLAP

Αναφέρεται σε δεδομένα OLAP, τα οποία είναι προσβάσιμα από ένα πρόγραμμα περιήγησης στο διαδίκτυο.

2.6 Η γλώσσα ερωτημάτων MDX (MultiDimensional eXpressions)

Η MDX είναι μια γλώσσα ερωτημάτων για την OLAP. Η σύνταξη της μοιάζει αρκετά με αυτή της SQL, ωστόσο πρόκειται για εντελώς καινούρια γλώσσα μιας και η δομή των δεδομένων διαφέρει στους κύβους OLAP από τις σχεσιακές βάσεις δεδομένων. Η MDX χρησιμοποιήθηκε για πρώτη φορά στο πρόγραμμα OLE DB for OLAP της Microsoft, το 1997. Η γλώσσα παρέχει εξειδικευμένη σύνταξη για την υποβολή ερωτημάτων και το χειρισμό των πολυδιάστατων δεδομένων που είναι αποθηκευμένα σε κύβους OLAP. Όπως οι πίνακες και οι στήλες είναι ο πυρήνας των SQL ερωτημάτων, οι διαστάσεις, οι ιεραρχίες, και τα επίπεδα, είναι τα αντίστοιχα στη γλώσσα MDX. Όλα τα αντικείμενα στην MDX έχουν μοναδικά ονόματα.

Η βασική μορφή μιας έκφρασης MDX είναι:

```
SELECT {member selection} ON COLUMNS
FROM [cube name]
```

Η λέξη-κλειδί SELECT σηματοδοτεί την έναρξη του ερωτήματος και καθορίζει το τι θέλει ο χρήστης να επιλέξει. Η λέξη-κλειδί ON χρησιμοποιείται για την οργάνωση των επιλεγμένων δεδομένων. Τα δεδομένα που επιλέγονται από μια διάσταση τοποθετούνται σε έναν πίνακα. Στην περίπτωση του παραδείγματος, τα δεδομένα θα τοποθετηθούν σε μία στήλη του πίνακα. Στην πρόταση FROM, καθορίζεται ο κύβος που θα χρησιμοποιηθεί για την αναζήτηση των δεδομένων που καθορίζονται στην πρόταση SELECT.

Παραπάνω υπάρχει η βασική μορφή σύνταξης της MDX. Συχνά, όμως, είναι απαραίτητη η σύγκριση των πληροφοριών μιας διάσταση με μια άλλη. Η βασική μορφή του ενός δυσδιάστατου ερωτήματος είναι:

```
SELECT {member selection} ON COLUMNS,
       {member selection} ON ROWS
FROM [cube name]
```

Στα ερωτήματα μπορεί να υπάρχει και η πρόταση WHERE, προκειμένου να μειώσει το κομμάτι του κύβου που θα συμπεριληφθεί στο ερώτημα. Όταν δεν υπάρχει η πρόταση WHERE σε ένα ερώτημα, το σύνολο του κύβου λαμβάνεται υπόψη κατά τον υπολογισμό των αποτελεσμάτων. Όταν καθορίζεται μια πρόταση WHERE το πεδίο του ερωτήματος μειώνεται. Για το λόγο αυτό η πρόταση WHERE

ονομάζεται «κόφτης» (slicer), επειδή χρειάζεται μια «φέτα» από το συνολικό κύβο και εκτελεί τους υπολογισμούς μόνο σε αυτό το κομμάτι των δεδομένων.

Υπάρχουν κι άλλες λειτουργίες της MDX, που χρησιμοποιούνται για την πιο χρήσιμη ανάλυση των δεδομένων κύβου. Μια απ' αυτές είναι η πρόταση WITH MEMBER, που επιτρέπει στο χρήστη να δημιουργήσει ένα νέο υπολογισμένο μέλος (calculated member). Το νέο μέλος μπορεί να είναι είτε νέα μέτρηση είτε νέο μέλος μιας διάστασης. Σε μια πρόταση MDX, μπαίνει πριν το SELECT με την ακόλουθη σύνταξη:

```
WITH MEMBER parent.name AS 'expression'
```

Υπάρχουν τρία μέρη στην πρόταση:

- *parent*: Αναφέρεται στο γονέα του νέου μέλους. Επειδή οι διαστάσεις είναι οργανωμένες σε ιεραρχίες, όταν προστίθενται νέα μέλη στο δέντρο της ιεραρχίας θα πρέπει να καθορίζετε η θέση τους στο εσωτερικό της ιεραρχίας.
- *name*: Το όνομα που δίνει ο χρήστης στο νέο μέλος.
- *expression*: Το κύριο μέρος του ορισμού όπου καθορίζεται πως προέρχονται τα αποτελέσματα.

Επίλογος

Στο κεφάλαιο αυτό είδαμε τα σημαντικότερα στοιχεία για την αναλυτική επεξεργασία άμεσης επικοινωνίας, το βασικό της συστατικό που είναι ο κύβος OLAP και τις πράξεις που μπορούν να εφαρμοστούν σε αυτόν. Στη συνέχεια, θα δούμε πως μπορεί να υλοποιηθούν αυτά με παραδείγματα στα προγράμματα IBM Infosphere Warehouse και saiku.

Κεφάλαιο 3: Παραδείγματα κι εφαρμογές

Εισαγωγή

Στο κεφάλαιο αυτό θα δούμε πως μπορεί να υλοποιηθούν οι τεχνολογίες Data Warehousing και OLAP, με παραδείγματα τόσο στο εμπορικό πρόγραμμα IBM InfoSphere Warehouse, όσο και στο open source saiku.

3.1 Παράδειγμα στο IBM InfoSphere Warehouse

3.1.1 IBM InfoSphere Warehouse

Πρόκειται για τη σουίτα της IBM, που χρησιμοποιείται, εκτός των άλλων και για εφαρμογές Data Warehousing και OLAP.

Η πλατφόρμα προσφέρει ένα εμπορικό πρόγραμμα, που παρέχει την απόδοση, την επεκτασιμότητα, την αξιοπιστία και την επιτάχυνση που απαιτείται για την απλοποίηση δύσκολων ερωτημάτων και προσφέρει αξιόπιστες πληροφορίες στο χρήστη γρηγορότερα.

Χρησιμοποιήθηκε η έκδοση 9.7.2

3.1.1.1 Η βάση

Θα χρησιμοποιήσουμε τη Βάση Δεδομένων GSDB που περιέχει τα δεδομένα της φανταστικής εταιρίας «Sample Outdoor», η οποία πουλάει εξοπλισμό εξωτερικών χώρων. Η εταιρία πουλάει και διανέμει τα προϊόντα της μέσω τρίτων (καταστήματα λιανικής σε όλο τον κόσμο), αλλά και μέσω του ηλεκτρονικού καταστήματος της. Οι γραμμές παραγωγής περιλαμβάνουν: εξοπλισμό κατασκήνωσης, εξοπλισμό γκολφ, εξοπλισμό ορειβασίας, προϊόντα προστασίας σε εξωτερικούς χώρους και προσωπικά αξεσουάρ.

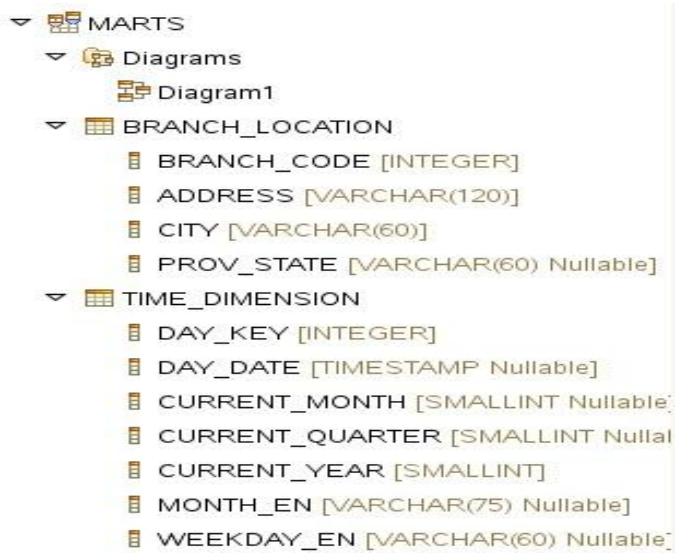
3.1.1.2 Design Studio

Στο παράδειγμα μας θα χρησιμοποιήσουμε το περιβάλλον Design Studio. Το Design Studio είναι ένα κομμάτι του Infosphere Warehouse που χρησιμοποιείται για εργασίες σχεδιασμού αποθήκης, συμπεριλαμβανομένης της μοντελοποίησης δεδομένων OLAP.

3.1.2 Παράδειγμα Data Warehousing

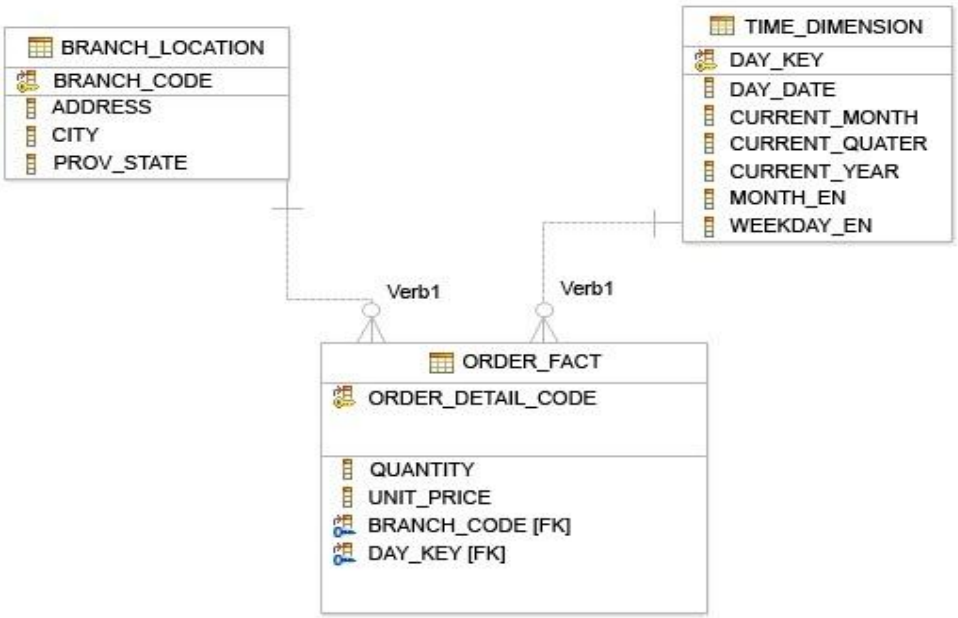
Θα χρησιμοποιήσουμε τα δεδομένα της βάσης δεδομένων GSDB και θα δημιουργήσουμε μια εφαρμογή αποθήκης δεδομένων, η οποία είναι ένα κέντρο δεδομένων (data mart) που θα επιτρέπει στους χρήστες να παρακολουθείτε τις πωλήσεις με την πάροδο του χρόνου στα διάφορα καταστήματα της εταιρείας Sample Outdoors.

Πρώτα απ' όλα, δημιουργούμε ένα data warehousing project κι ένα φυσικό μοντέλο δεδομένων. Στη συνέχεια, θα δημιουργήσουμε ένα καινούριο σχήμα, το Marts, το οποίο θα περιέχει τα αντικείμενα για το κέντρο δεδομένων. Στο σχήμα, θα εισάγουμε δύο πίνακες διαστάσεων, έναν για τα καταστήματα κι έναν για το χρόνο (Εικόνα 20).



Εικόνα 20: Πίνακες διαστάσεων σχήματος Marts

Ακολούθως, θα δημιουργήσουμε έναν πίνακα γεγονότων, με τον οποίον θα συνδέονται οι πίνακες διαστάσεων, και τους αντίστοιχους περιορισμούς ξένου κλειδιού, ώστε να έχουμε ένα ολοκληρωμένο σχήμα αστέρα. Στην εικόνα 21 υπάρχει το σχήμα αστέρα, όπου στον πίνακα γεγονότων ORDER_FACT, υπάρχουν τα ξένα κλειδιά από τους πίνακες διαστάσεων.



Εικόνα 21: Σχήμα αστέρα παραδείγματος

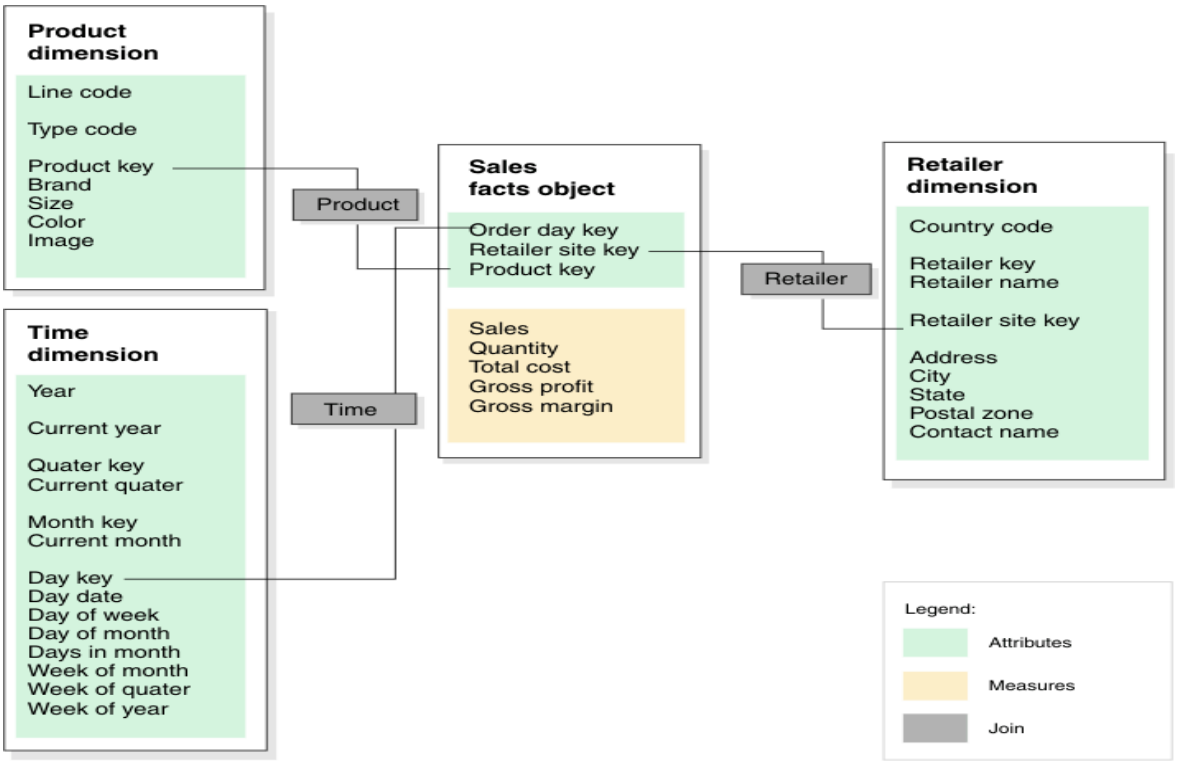
Το επόμενο βήμα είναι η δημιουργία ροών δεδομένων. Οι ροές δεδομένων (data flows) χρησιμοποιούνται για να εξάγουν δεδομένα από πηγές δεδομένων, όπως οι σχεσιακές βάσεις, να τα μετατρέψουν με τη χρήση τελεστών, και για να τα εισάγουν στον τελικό προορισμό, που μπορεί να είναι οι αποθήκες δεδομένων. Στο παράδειγμα μας, πηγή δεδομένων είναι η βάση GSDB και τελικός προορισμός το κέντρο δεδομένων της αποθήκης που σχεδιάζουμε. Στην εικόνα 3, βλέπουμε τη ροή δεδομένων που φτιάξαμε για την εισαγωγή δεδομένων από τον πίνακα «BRANCH» της GSDB στον πίνακα «BRANCH_LOCATION» του σχήματος MARTS. Επειδή δεν θέλουμε όλα τα δεδομένα του πίνακα «BRANCH», χρησιμοποιούμε τον τελεστή «Select List» για να επιλέξουμε μόνο όποια χρειαζόμαστε.

Εικόνα 22: Ροή δεδομένων για BRANCH_LOCATION

Με παρόμοιο τρόπο δημιουργούνται οι ροές δεδομένων και για τους άλλους πίνακες του σχήματος MARTS.

3.1.3 Παράδειγμα OLAP

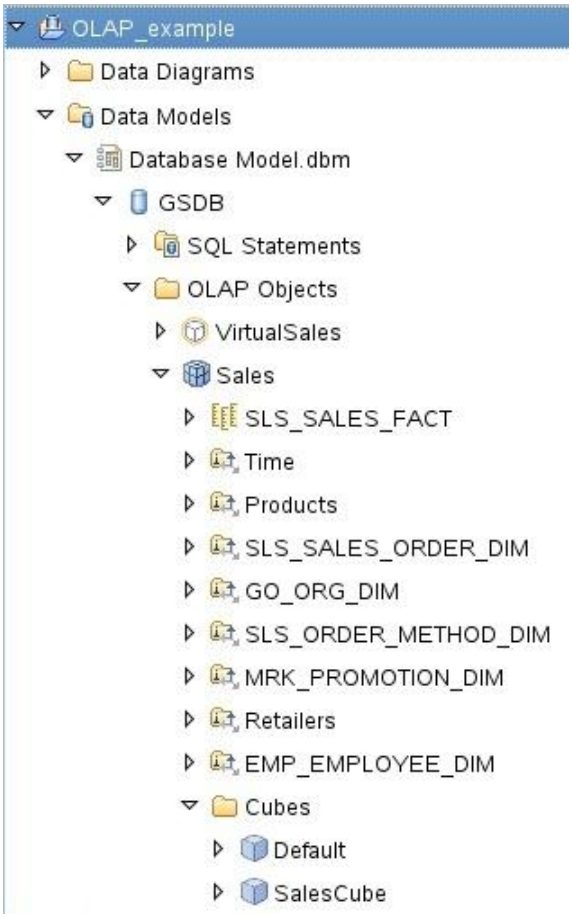
Αρχικά, δημιουργούμε ένα OLAP Data Design Project. Όταν το δημιουργούμε, αρχικά, περιέχει διάφορους κενούς φακέλους γι' αυτό και θα προσθέσουμε το φυσικό μοντέλο δεδομένων (physical data model), το μοντέλο του κύβου και τους κύβους. Ένα OLAP Data Design Project αποτελείται από φυσικά μοντέλα δεδομένων. Τα φυσικά μοντέλα δεδομένων περιέχουν μοντέλα κύβων, και τα μοντέλα κύβων περιέχουν κύβους. Το φυσικό μοντέλο θα το δημιουργήσουμε, με βάση το σχήμα αστέρα της Εικόνας 23.



Εικόνα 23: Σχήμα αστέρα για το παράδειγμα OLAP

Το "InfoSphere Warehouse" αποθηκεύει πληροφορίες σχετικά με τα σχεσιακά δεδομένα σε αντικείμενα μεταδεδομένων τα οποία μας παρέχουν μια νέα προοπτική των δεδομένων. Ορισμένα αντικείμενα μεταδεδομένων ενεργούν ως βάση για άμεση πρόσβαση σε σχεσιακά δεδομένα. Άλλα αντικείμενα μεταδεδομένων περιγράφουν τις σχέσεις μεταξύ των βασικών αντικειμένων μεταδεδομένων. Όλα τα αντικείμενα των μεταδεδομένων μπορούν να ομαδοποιηθούν ανάλογα με τις μεταξύ τους σχέσεις σε ένα αντικείμενο μεταδεδομένων που ονομάζεται μοντέλο κύβου. Ένα μοντέλο κύβου παρουσιάζει μια συγκεκριμένη ομαδοποίηση και τη διαμόρφωση των σχεσιακών πινάκων.

Το μοντέλο κύβου, για τα δεδομένα των πωλήσεων, θα το φτιάξουμε για να μπορούμε να ομαδοποιήσουμε και να καθορίσουμε τα δεδομένα μας. Η βάση δεδομένων GSDB περιέχει δεδομένα πωλήσεων, συμπεριλαμβανομένων πληροφοριών σχετικά με τα προϊόντα και τους λιανοπωλητές. Μαζί με το μοντέλο του κύβου, δημιουργείται αυτόματα κι ένας default κύβος. Ο default κύβος περιέχει ένα ευρύ φάσμα μεταδεδομένων και μετρήσεων. Για να περιορίσουμε αυτό το φάσμα, θα δημιουργήσουμε έναν άλλο κύβο. Ο καινούριος κύβος θα περιέχει δεδομένα μόνο για τα προϊόντα, τους λιανοπωλητές και το χρόνο. Με τη χρήση των κύβων ελαχιστοποιείται η συνολική επεξεργαστική ισχύ που είναι αναγκαία για ερωτήματα στο μοντέλο κύβου, επειδή τα ερωτήματα γίνονται σε υποσύνολα του μοντέλου.



Εικόνα 24: Δομή OLAP project

Ο κύβος που δημιουργήσαμε είναι αρχικά κενός. Θα τον τροποποιήσουμε, έτσι ώστε ο χρήστης να μπορεί να θέτει ερωτήματα στον κύβο. Αρχικά, προσθέτουμε τις διαστάσεις Time, Products, και Retailers. Διαμορφώνουμε τη διάσταση Time, έτσι ώστε να περιλαμβάνει τα επίπεδα έτους, τρίμηνου, μήνα, μέρας. Τέλος, θα δημιουργήσουμε μετρήσεις οι οποίες θα βοηθάνε το χρήστη να πάρει χρήσιμα

αποτελέσματα, όταν θέτει ερωτήματα στον κύβο. Μια τέτοια μέτρηση θα είναι αυτή που θα υπολογίζει το κόστος των προϊόντων.

Για παράδειγμα, αν θέλουμε τα δούμε το συνολικό κόστος των προϊόντων της εταιρίας το 2006, με την MDX, ο κώδικας του ερωτήματος θα είναι:

```
SELECT
CrossJoin({[Product]}, {[Measures].[Product Cost]}) ON COLUMNS,
{Hierarchize({[Time].[2006]})} ON ROWS
FROM [SalesCube]
```

Και τα αποτελέσματα:

```
===Axis Info===
Axis[0] = {(All, Product Cost)}
Axis[1] = {(2006)}

Cell[0]: 872630528.43
```

Ένα άλλο παράδειγμα από τον συγκεκριμένο κύβο είναι να προβάλλουμε, τους λιανοπωλητές που πωλούσαν προϊόντα γκολφ ανά έτος:

```
===Axis Info===
Axis[0] = {(2004), (2005), (2006), (2007)}
Axis[1] = {(Golf Retailers)}

Cell[0]: 45588707.98
Cell[1]: 71237904.29
Cell[2]: 100878698.67
Cell[3]: 78937877.03
```

3.2 Παραδείγματα στο Saiku

3.2.1 Saiku Analytics

Η εταιρία Saiku Analytics ιδρύθηκε το 2008. Προσφέρει ένα open source Business Intelligence (BI) προϊόν που ονομάζεται Saiku Analytics. Το Saiku Analytics είναι μια ελαφριά κι ευέλικτη web-based εφαρμογή που προσφέρει αναλύσεις OLAP, και είναι επεκτάσιμη, προσαρμόσιμη και εύκολο να ενσωματωθεί. Επιτρέπει στους χρήστες, γρήγορα και εύκολα, να αναλύσουν εταιρικά δεδομένα και να δημιουργήσουν εκθέσεις. Στα παραδείγματα, χρησιμοποιήθηκε η έκδοση 2.4.

Εικόνα 25: Saiku

3.2.1.1 Η βάση

Για τα παραδείγματα, χρησιμοποιήθηκε η Βάση Δεδομένων Foodmart και πιο συγκεκριμένα ο κύβος Sales, που περιέχει δεδομένα πωλήσεων.

3.2.2 Παράδειγμα

Θα δούμε ένα παράδειγμα το οποίο δείχνει πόσο κοστίζουν τα προϊόντα στα καταστήματα και ποια είναι τα έσοδα των καταστημάτων από την πώληση των προϊόντων αυτών. Θα χρησιμοποιήσουμε τις μετρήσεις Store Cost και Store Sales, αντίστοιχα. Θα προβάλουμε τα αποτελέσματα ανά πολιτεία (Store State) κι ανά κατηγορίες προϊόντων, για το έτος 1997.

Τα αποτελέσματα θα είναι τα παρακάτω:

Βλέπουμε, ότι το κόστος της κατηγορίας των ποτών, για τα καταστήματα της Καλιφόρνιας (CA) είναι $ 5662,27 και τα έσοδα τους από την αντίστοιχη κατηγορία $ 14203,24.

Ο κώδικας MDX, που παράγεται αυτόματα από το saiku, θα είναι:

```
SELECT
NON EMPTY Hierarchize(Union(CrossJoin([Store].[Store State].Members, {[Measures].[Store Cost]}), CrossJoin([Store].[Store State].Members, {[Measures].[Store Sales]}))) ON COLUMNS,
NON EMPTY CrossJoin([Product].[Product Family].Members, [Time].[Year].Members) ON ROWS
FROM [Sales]
```

Αν αφαιρέσουμε τον τελεστή NON EMPTY, στο ίδιο ερώτημα, θα έχουμε τα εξής αποτελέσματα:

		CA		OR		WA	
Product Family	Year	Store Cost	Store Sales	Store Cost	Store Sales	Store Cost	Store Sales
Drink	1997	5.662,27	14.203,24	4.836,35	12.137,29	8.978,61	22.495,68
Food	1997	45.980,35	115.193,17	40.967,45	102.564,67	76.322,92	191.277,75
Non-Consumable	1997	11.887,80	29.771,43	10.968,70	27.575,11	20.022,78	50.019,79

		BC		DF		
Product Family	Year	Store Cost	Store Sales	Store Cost	Store Sales	Sto Co
Drink	1997					
	1998					
Food	1997					
	1998					
Non-Consumable	1997					
	1998					

Κι ο αντίστοιχος κώδικας:

```
SELECT
Hierarchize(Union(CrossJoin([Store].[Store Country].Members, {[Measures].[Store Cost]}), CrossJoin([Store].[Store Country].Members, {[Measures].[Store Sales]}))) ON COLUMNS,
{Hierarchize({[Product].[Product Family].Members})} ON ROWS
FROM [Sales]
```

3.2.3 Πράξεις OLAP

Roll up

Ένα παράδειγμα σύμπτυξης θα ήταν να δούμε τον κύβο σε επίπεδο χώρας.

		USA	
Product Family	Year	Store Cost	Store Sales
Drink	1997	19.477,23	48.836,21
Food	1997	163.270,72	409.035,59
Non-Consumable	1997	42.879,28	107.366,33

Drill-down

Ένα παράδειγμα ανάπτυξης θα ήταν να δούμε τον κύβο σε επίπεδο οικονομικών τριμήνων.

			CA		OR		WA	
Product Family	Year	Quarter	Store Cost	Store Sales	Store Cost	Store Sales	Store Cost	Store Sales
Drink	1997	Q1	1.329,80	3.309,75	1.269,03	3.169,69	2.022,68	5.106,36
		Q2	1.320,77	3.329,80	1.177,38	2.961,28	2.230,32	5.623,50
		Q3	1.395,12	3.503,55	1.196,87	3.002,72	2.198,67	5.487,73
		Q4	1.616,59	4.060,14	1.193,09	3.003,60	2.526,93	6.278,09
Food	1997	Q1	10.393,84	26.044,84	11.505,69	28.736,14	18.547,32	46.480,34
		Q2	11.135,95	27.879,91	9.086,02	22.757,50	17.890,65	44.798,59
		Q3	11.334,08	28.523,85	10.451,31	26.251,48	18.759,67	47.032,27
		Q4	13.116,48	32.744,57	9.924,43	24.819,55	21.125,28	52.966,55
Non-Consumable	1997	Q1	2.707,45	6.820,61	3.306,36	8.264,46	4.670,08	11.696,16
		Q2	2.875,30	7.187,04	2.415,56	6.054,10	4.832,27	12.074,55
		Q3	2.943,63	7.366,65	2.625,61	6.626,26	4.999,92	12.477,38
		Q4	3.361,43	8.397,13	2.621,16	6.630,29	5.520,51	13.771,70

Slice

Αν θέλουμε να παραχθεί υπολογιστικό φύλλο με τις πληροφορίες των πολιτειών και του χρόνου για όλα τα κελιά που αφορούν την οικογένεια προϊόντων Ποτά (Drink), θα έχουμε το παρακάτω αποτέλεσμα: ΄

			CA		OR		WA	
Product Family	Year	Quarter	Store Cost	Store Sales	Store Cost	Store Sales	Store Cost	Store Sales
Drink	1997		5.662,27	14.203,24	4.836,35	12.137,29	8.978,61	22.495,68
		Q1	1.329,80	3.309,75	1.269,03	3.169,69	2.022,68	5.106,36
		Q2	1.320,77	3.329,80	1.177,38	2.961,28	2.230,32	5.623,50
		Q3	1.395,12	3.503,55	1.196,87	3.002,72	2.198,67	5.487,73
		Q4	1.616,59	4.060,14	1.193,09	3.003,60	2.526,93	6.278,09

Dice

Αν θέλουμε να βρούμε το τρίμηνο με τις υψηλότερες δαπάνες για κάθε πολιτεία ανά κατηγορία προϊόντων, θα έχουμε το αποτέλεσμα της εικόνας:

Product Family	MeasuresLevel	CA	OR	WA
Drink	Max	1.616,59	1.269,03	2.526,93
Food	Max	13.116,48	11.505,69	21.125,28
Non-Consumable	Max	3.361,43	3.306,36	5.520,51

Για να βρούμε το παραπάνω αποτέλεσμα θα πρέπει να γράψουμε τον εξής κώδικα σε MDX:

```
WITH MEMBER [Measures].[Max] AS MAX([Time].[Quarter].Members, [Measures].[Store Cost])
SELECT
NON EMPTY {Hierarchize({[Store].[Store State].Members})} ON COLUMNS,
NON EMPTY CrossJoin({{[Product].[Product Family].Members}}, [Measures].[Max]) ON ROWS
```

Βιβλιογραφία

[1] Νανόπουλος Α., & Μανωλόπουλος Ι., (2008), *Εισαγωγή στην εξόρυξη και τις αποθήκες δεδομένων*, Αθήνα: Εκδόσεις Νέων Τεχνολογιών

[2] Ramakrishnan R., & Gehrke J., (2002), *Συστήματα διαχείρισης βάσεων δεδομένων* (Τόμ. Β) (Δ. Δέρβος, Γ. Ευαγγελίδης, Μετάφ.), Θεσσαλονίκη: Εκδόσεις Τζιόλα (Το πρωτότυπο έργο δημοσιεύτηκε το 2000).

[3] Roiger R., & Geatz M., (2008), *Εξόρυξη Πληροφορίας, Ένας εισαγωγικός οδηγός με παραδείγματα* (Γ. Μαυρόπουλος, Μετάφ.), [χ.τ.]: Εκδόσεις Κλειδάριθμος. (Το πρωτότυπο έργο δημοσιεύτηκε το 2003).

[4] Baragoin C., Bercianos J., Komel J., Robinson G., Sawa R., & Schuinder E., (2001), *DB2 OLAP server: Theory and practices*, [χ.τ.]: International Business Machines Corporation. Ανακτήθηκε 3 Σεπτεμβρίου 2012 από IBM Redbooks Διαδικτυακός τόπος: http://www.redbooks.ibm.com/

[5] Borysowich C., (2007), *Star Schema Modelling (Data Warehouse)*, Ανακτήθηκε 18 Νοεμβρίου 2012, από http://www.toolbox.com/

[6] Celko J., (2006), *Analytics & OLAP in SQL*, United States of America: Elsevier Inc.

[7] Dunham M., (2003), Data Mining: Introductory and advanced topics, [χ.τ]: Prentice Hall

[8] Golfarelli M., & Rizzi S., (2009), *Data warehouse design: Modern principles and methodologies*, United States of America: The McGraw-Hill Companies, S.r.l.-Publishing Group Italia.

[9] Han J., & Kamber M., (2006), *Data Mining: Concepts and Techniques*, United States of America: Elsevier Inc.

[10] Kimball R., & Ross M., (2002), *The data warehouse toolkit: The complete guide to dimensional modeling* (2nd Ed.) United States of America: John Wiley & Sons, Inc.

[11] Nagabhushana S., (2006), *Data Warehousing: OLAP and data mining*, [χ.τ.]: New Age International (P) Ltd., Publishers

[12] Nolan C., (1999), *Manipulate and Query OLAP Data Using ADOMD and Multidimensional Expressions*, Ανακτήθηκε 8 Νοεμβρίου 2012, από http://www.microsoft.com/msj/0899/mdx/mdx.aspx

[13] Ponniah P., (2010), *Data Warehousing: Fundamentals for IT professionals*, New York: John Wiley & Sons, Inc.

[14] Robert W., & Koncilia C., (2007), *Data warehouses and OLAP: Concepts, architectures and solutions*, United States of America: Idea Group Inc.

[15] Rud O., (2009), *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. Hoboken, New Jersey: Wiley & Sons.

[16] Tan P., Steinbach M., & Kumar V., (2006), *Introduction to data mining*, United States of America: Pearson Education, Inc.

[17] Thierauf R., (1997), *On-Line analytical processing systems for business*, [χ.τ.]: Greenwood Publishing Group

[18] Thomsen E., (2002), *OLAP solutions: Building multidimensional information systems* (2nd Ed.), United States of America: John Wiley & Sons, Inc.

[19] Van der Velden M., (2010), *Why MDX?*, Ανακτήθηκε 8 Νοεμβρίου 2012, από http://blogs.simba.com/simba_technologies_ceo_co/2010/11/why-mdx.html

Άλλες πηγές

- http://pic.dhe.ibm.com/infocenter/db2luw/v9r7/index.jsp
- http://analytical-labs.com/index.html
- http://www.1keydata.com
- http://www.learnbi.com

- http://datawarehouse4u.info

Οδηγός χρήσης λογισμικού

Στην εργασία χρησιμοποιήθηκε το πρόγραμμα saiku, το οποίο έγινε λήψη από την ιστοσελίδα http://analytical-labs.com/.

Για τη χρήση του προγράμματος απαιτείται η ύπαρξη του J2SE runtime environment (JRE) και του apache tomcat server. Για το JRE χρειάζεται η έκδοση 5.0 ή νεώτερη και μπορεί να γίνει λήψη από την ιστοσελίδα http://java.sun.com/j2se (επίσης, μπορεί να γίνει λήψη της πλήρης πλατφόρμας Java, JDK, που περιλαμβάνει το JRE). Ο apache tomcat server μπορεί να ληφθεί είτε μαζί με το saiku είτε από την ιστοσελίδα http://tomcat.apache.org. Μετά για να ξεκινήσει το saiku, θα πρέπει να τρέξει το αρχείο start-saiku.bat για τα Windows ή το start-saiku.sh για unix. Τέλος για να χρησιμοποιήσουμε το πρόγραμμα, θα πρέπει να συνδεθούμε από έναν web browser στη διεύθυνση localhost:8080. Το πρόγραμμα τερματίζεται με την εντολή stop-saiku.bat για τα Windows ή το stop -saiku.sh για unix.