Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Πρόβλεψη Δικτυακών Επιθέσεων με χρήση τεχνικών Εξόρυξης Γνώσης: Υλοποίηση Ανιχνευτή Επιθέσεων** **Prediction of Internet Attack using Data Mining Techniques: Implementation of an Attack Detector**
Ονοματεπώνυμο Φοιτητή	**Κωνσταντίνος Σταυράκης**
Πατρώνυμο	**Κυριάκος**
Αριθμός Μητρώου	**ΜΠΠΛ/ 11034**
Επιβλέπων	**Χρήστος Δουληγέρης, Καθηγητής**

Ημερομηνία Παράδοσης **Οκτώβριος 2015**

Τριμελής Εξεταστική Επιτροπή

Χρήστος Δουληγέρης Καθηγητής	Παναγιώτης Κοτζανικολάου Επίκουρος Καθηγητής	Κων/νος. Πιτσάκης Λέκτορας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της μεταπτυχιακής διατριβής μου καθηγητή. κ. Χρ. Δουληγέρη για τις πολύτιμες συμβουλές και την καθοδήγηση που μου παρείχε.

Θερμές ευχαριστίες οφείλω, επίσης, στους: την επ. καθηγήτρια κα. Ν. Πολέμη και τον Δρα Σπ. Παπαστεργίου , οι οποίοι ήταν συνεπιβλέποντες της διατριβής, για την αμέριστη υποστήριξη και για τις πολύτιμες συμβουλές τους. Η συνδρομή τους ήταν καθοριστικής σημασίας για την εκπόνηση της διατριβής.

Τέλος, ευχαριστώ θερμά την οικογένεια μου για την έμπρακτη συμπαράσταση τους καθ' όλη την διάρκεια εκπόνησης της διατριβής.

Οκτώβριος 2015

Κωνσταντίνος Σταυράκης

Περίληψη

Η παρούσα μεταπτυχιακή διατριβή έχει ως θέμα την Εξόρυξη Γνώσης (data mining) από δεδομένα επιθέσεων σε υπολογιστικά συστήματα, βάσει των οποίων καθίσταται δυνατή η πρόβλεψη επιθέσεων. Δηλαδή θα επιχειρείται η εφαρμογή τεχνικών Εξόρυξης Γνώσης (data mining) στην Ασφάλεια με σκοπό την πρόβλεψη επιθέσεων.

Αυτός ο στόχος επιτυγχάνεται με την υλοποίηση ενός ανιχνευτή δικτυακών εισβολών. Συγκεκριμένα, ενός μοντέλου πρόβλεψης ικανού να διακρίνει τις ‘κακές’ συνδέσεις, οι οποίες ονομάζονται εισβολές ή επιθέσεις, από τις ‘καλές’ φυσιολογικές συνδέσεις.

Abstract

The subject of the current study is application of data mining techniques on data, for intrusion detection in computer networks. The training dataset consist of data regarding network connections. By analyzing that data and applying data mining techniques on that data, current study aims to be able to predict network attacks.

The aim of this study will be reached with the development of a network intrusion detector. More specific, the created model will be able to distinguish network attacks, which are called intrusions - from normal network connections.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΠΙΝΑΚΑΣ ΠΙΝΑΚΩΝ

Κεφάλαιο 1ο

1 Εισαγωγή

Ο στόχος της παρούσης διατριβής είναι η υλοποίηση ενός ανιχνευτή δικτυακών εισβολών, δηλαδή η κατασκευή ενός Πληροφοριακού Συστήματος Data Warehouse – Business Intelligence. Χρησιμοποιήθηκε το σύνολο δεδομένων (data set) που είναι το KDD Cup 1999 Data Set (1). Το εν λόγω σύνολο δεδομένων περιλαμβάνει μεγάλη ποικιλία από δικτυακές εισβολές (network intrusion) οι οποίες δημιουργήθηκαν έπειτα από προσομοίωση. Το Πληροφοριακό Σύστημα επεξεργάζεται τα στοιχεία με διαδικασίες Data Modeling, ETL (Extract, Transform, Load) μηχανισμών για το Loading των στοιχείων και το Transformation και την δημιουργία Κύβων Άμεσης Αναλυτικής Επεξεργασίας (On-line Analytical Processing OLAP Cubes) για την εξαγωγή (Publishing) των αποτελεσμάτων για την ανάλυση τους.

1.1 Περιγραφή του υπό μελέτη προβλήματος

Με τη μεγάλη αύξηση της χρήσης δικτύων υπολογιστών και την παράλληλη ανάλογη αύξηση του αριθμού των εφαρμογών που μπορούν να εκτελεστούν σε έναν υπολογιστή που βρίσκετε συνδεδεμένος σε ένα δίκτυο υπολογιστών, το θέμα της ασφάλειας των δικτύων υπολογιστών έχει εξελιχθεί σε σημαντικό ζήτημα για κάθε διαχειριστή δικτύου υπολογιστών. Ως εκ τούτου, υπάρχει ανάγκη πρόβλεψης των διαδικτυακών επιθέσεων .

1.2 Σκοπός και στόχοι της εργασίας

Οι στόχοι της εργασίας περιλαμβάνουν τα ακόλουθα:

1. Η ανάλυση εννοιών επί θεμάτων Ασφάλειας, Αποθήκες Δεδομένων και Εξόρυξης Γνώσης .

2. Η δημιουργία ενός ανιχνευτή δικτυακών εισβολών. Συγκεκριμένα, ενός μοντέλου πρόβλεψης ικανού να διακρίνει τις ‘κακές’ συνδέσεις, οι οποίες ονομάζονται εισβολές ή επιθέσεις, από τις ‘καλές’ φυσιολογικές συνδέσεις.

1.3 Παραδοτέα της εργασίας

Το τελικό παραδοτέο αποτελείται από ένα φάκελο που περιέχει τα εξής:

1. Το έντυπο κείμενο της εργασίας, το οποίο περιλαμβάνει αναλυτική περιγραφή των προσεγγίσεων που ακολουθήθηκαν σε κάθε μια από τις εργασίες συνοδευόμενα από τα κατάλληλα σχήματα (ΒΔ, ΑΔ κτλ), screenshots κτλ, επισκόπηση της σχετικής με το χώρο βιβλιογραφίας, τα αποτελέσματα της διατριβής.

2. Συνημμένα αρχεία που περιέχουν τα εξής: τα περιεχόμενα (backup) της αρχικής ΒΔ, τα script files της προπαρασκευής των δεδομένων, τα περιεχόμενα της ΑΔ, τα μοντέλα που προέκυψαν από την κατηγοριοποίηση, τα μοντέλα που προέκυψαν από την συσταδοποίηση και τα μοντέλα που προέκυψαν από την ανάλυση συσχετίσεων..

3. Η συλλογή πηγών και σχετικής βιβλιογραφίας για τη δημιουργία μίας βάσης γνώσης σχετικά με το θέμα.

1.4 Δομή διατριβής

Η Παρούσα διατριβή αποτελείται από 2 μέρη. Το πρώτο μέρος (κεφάλαια 2-5) αφορά το θεωρητικό τμήμα της διατριβής και περιγράφονται οι σχετικές έννοιες που χρησιμοποιούνται. Το δεύτερο μέρος (κεφάλαιο 6) αναλύεται ο Ανιχνευτής Δικτυακών Εισβολών για την πρόβλεψη διαδικτυακών επιθέσεων με χρήση τεχνικών εξόρυξης γνώσης (ΕΓ) , ο οποίος αποτελεί και την συμβολή της παρούσης διατριβής. Ουσιαστικά, πρόκειται για ένα σύστημα Επιχειρηματικής Ευφυΐας (Business Intelligence – BI)

Για την εξόρυξης γνώσης (ΕΓ) , χρησιμοποιήθηκε το σύνολο δεδομένων KDD Cup 1999 Data Set (1). Το εν λόγω σύνολο δεδομένων περιλαμβάνει μεγάλη ποικιλία από δικτυακές εισβολές (network intrusion) οι οποίες δημιουργήθηκαν έπειτα από προσομοίωση.

Στόχος αποτελεί η δημιουργία ενός ανιχνευτή δικτυακών εισβολών. Συγκεκριμένα, ενός μοντέλου πρόβλεψης ικανού να διακρίνει τις ‘κακές’ συνδέσεις, οι οποίες ονομάζονται εισβολές ή επιθέσεις, από τις ‘καλές’ φυσιολογικές συνδέσεις.

Σε πρώτη φάση, επιλύεται το πρόβλημα της προπαρασκευής δεδομένων (data preprocessing) ώστε να κρατούνται όσα παρουσιάζουν ενδιαφέρον για τους σκοπούς της ανάλυσης.

Σε δεύτερη φάση , κατασκευάζεται επ’ αυτών των δεδομένων μια αποθήκη δεδομένων - κύβο - ακολουθώντας το σχήμα αστέρα (star schema) ή χιονονιφάδας (snowflake schema) και εκτελούμε στην ΑΔ λειτουργίες OLAP (roll-up, drill-down, slice, dice και pivot).

Σε τρίτη φάση, εκτελούνται διάφορες λειτουργίες ΕΓ (classification, clustering, feature selection, association rules, outlier detection κλπ.) έχοντας προκαθορίσει το σκοπό της ανάλυσης. Η παραπάνω διαδικασία μπορεί να επαναληφθεί πολλές φορές μέχρι το αποτέλεσμα της ανάλυσης να είναι ικανοποιητικό και η αποκτηθείσα γνώση για τα δεδομένα να είναι επαρκής, άρα και άμεσα χρήσιμη (actionable) από τους υπεύθυνους αποφάσεων των εκάστοτε εφαρμογών.

Η υλοποίηση του Ανιχνευτή Διαδικτυακών Επιθέσεων αποτελείται από τις ακόλουθες εργασίες:

1η εργασία (υλοποίηση ΒΔ): Επεξεργασία μεταφοράς δεδομένων από το text αρχείο και εισαγωγή των δεδομένων σε μια κατάλληλα σχεδιασμένη ΒΔ.

2η εργασία (προπαρασκευή δεδομένων): Από το σύνολο δεδομένων (dataset), επιλογή των δεδομένων θα επιλέξετε τα δεδομένα που θα χρησιμοποιηθούν για αναλυτική επεξεργασία, και προπαρασκευαστική εργασία (επιλογή, καθαρισμό, μετασχηματισμό, δειγματοληψία κλπ.)

3η εργασία (υλοποίηση της ΑΔ): Από το σύνολο δεδομένων, κατασκευή αποθήκης δεδομένων - κύβο - με κατάλληλες διαστάσεις, ιεραρχίες και μέτρα.

4η εργασία (αναλυτική επεξεργασία - OLAP):επίδειξη κύβου για αναλυτική επεξεργασία των δεδομένων. Παρουσίαση παραδειγμάτων λειτουργιών OLAP (roll-up, drill-down, slice, dice, pivot), επεξήγηση των αποτελεσμάτων και πιθανά συμπεράσματα.

5η εργασία (Σημαντικά χαρακτηριστικά feature selection): Περιγραφή σεναρίων που αναδεικνύουν ποια γνωρίσματα είναι σημαντικά .

6η εργασία (Κατηγοριοποίηση/ Πρόβλεψη-classification/prediction): Σενάρια κατηγοριοποίησης και πρόβλεψης. Περιγραφή διαδικασίας και εξήγηση των αποτελεσμάτων που προκύπτουν.

7η εργασία Συσταδοποίηση(clustering).

8η εργασία **Κανόνες συσχέτισης(association mining),**

Μέρος Ι: Θεωρητικό υπόβαθρο

Κεφάλαιο 2ο

2 Ασφάλεια δικτύων υπολογιστών

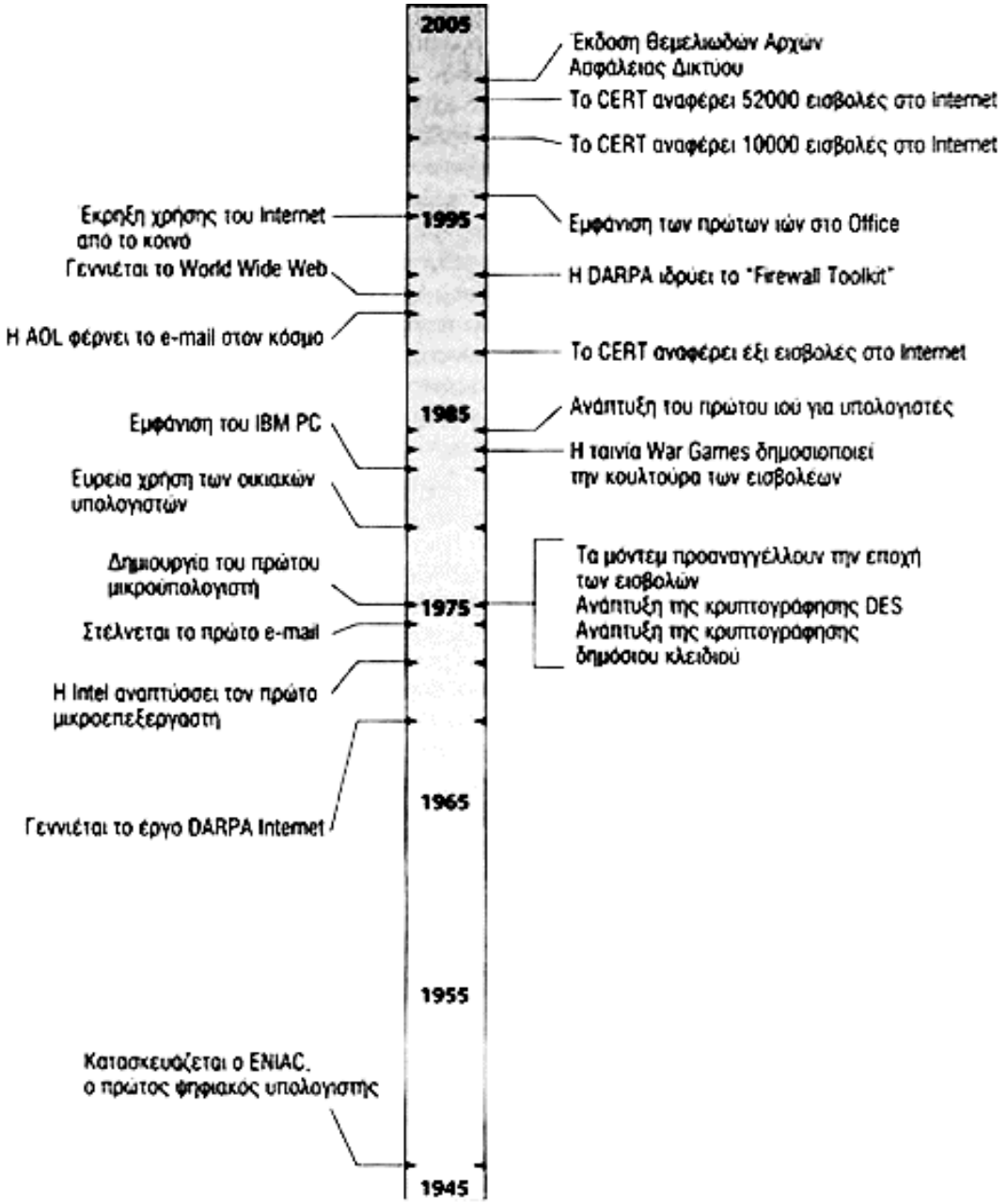
Η έννοια της ασφάλειας Δικτύου Υπολογιστών σχετίζεται με την ικανότητα μιας επιχείρησης ή ενός οργανισμού να προστατεύει τις πληροφορίες του από τυχόν αλλοιώσεις και καταστροφές, καθώς και από μη εξουσιοδοτημένη χρήση των πόρων του. Εκτός αυτού, θεωρείται ως η δυνατότητα ενός δικτύου ή συστήματος πληροφοριών να αντισταθεί, σε δεδομένο επίπεδο αξιοπιστίας, σε τυχαία συμβάντα ή κακόβουλες ενέργειες που θέτουν σε κίνδυνο τη διάθεση, την επαλήθευση ταυτότητας, την ακεραιότητα και την τήρηση του απορρήτου των δεδομένων που έχουν αποθηκευτεί ή μεταδοθεί καθώς και τις συναφείς υπηρεσίες που παρέχονται είτε είναι προσβάσιμες μέσω των δικτύων και συστημάτων αυτών (2) (3) (4) (5) (6) (7).

2.1 Ιστορική Εξέλιξη

Κατά τη διάρκεια της περιόδου 1955-1965 η ασφάλεια των υπολογιστών περιοριζόταν μόνο στο να περιορίζεται ένας δυσαρεστημένος υπάλληλος για να μην προκαλέσει καταστροφές και στο να μην έχουν οι ανταγωνιστές πρόσβαση στον υπολογιστή της επιχείρησης (8), (7).

Την περίοδο 1965-1975 καθώς άρχισαν οι κεντρικοί υπολογιστές να γίνονται πιο ισχυροί και ο αριθμός των χρηστών που συνδεόταν σε αυτούς έφτασαν τις χιλιάδες, το θέμα της υπευθυνότητα έγινε πιο σημαντικό. Η εισβολή εκείνη την εποχή ήταν σε επίπεδο φημών, περί κακόβουλων προγραμματιστών, που έκαναν παράνομες ενέργειες - όπως να γράφουν κώδικα που έπαιρνε τα δεκαδικά ψηφία τραπεζικών συναλλαγών και τα κατέθετε στο δικό τους λογαριασμό ή να γράφουν συστήματα "πίσω πόρτας" στον κώδικά τους για να μπορούν να μπαίνουν σε συστήματα.

Η έλλειψη πραγματικής ασφάλειας εμφανίστηκε στην περίοδο 1975-1985, όταν οι εταιρείες άρχισαν να παρέχουν απομακρυσμένη προσπέλαση σε χρήστες τερματικών, μέσω μόντεμ που εργαζόταν χρησιμοποιώντας το δημόσιο τηλεφωνικό δίκτυο. Η IBM ανέπτυξε τον αλγόριθμο Data Encryption Standard (DES) για την κυβέρνηση των Η.Π.Α το 1975. Σχεδόν ταυτόχρονα, οι Whitfield Diffie και Martin Hellman ανέπτυξαν την έννοια της κωδικοποίησης δημόσιου κλειδιού (Public Key Encryption, PKE), η οποία επέλυσε το πρόβλημα της ασφαλούς ανταλλαγής κλειδιού. Το 1977, οι Rivest, Shamir και Adelman υλοποίησαν την PKE στον ιδιοταγή αλγόριθμο κρυπτογράφησης RSA, που ήταν τα θεμέλια της σημερινής ασφάλειας δικτύων. Ωστόσο,το πρόσφατο ενδιαφέρον για την ασφάλεια τροφοδοτήθηκε από το έγκλημα του Kevin Mitnick. Ο Κέβιν Μίτνικ διέπραξε το μεγαλύτερο έγκλημα σε Ιστορίας των Ηνωμένων Πολιτειών το 1979. Οι απώλειες ήταν ογδόντα εκατομμυρίων δολάρια στις ΗΠΑ από την πνευματική ιδιοκτησία του πηγαίου κώδικα από διάφορες εταιρείες . Εκ τότε, ασφάλεια των πληροφοριών ήρθε στο προσκήνιο.



Εικόνα 1. Ιστορία της Ασφάλειας των Υπολογιστών (9)

2.2 Βασικοί ορισμοί

Η έννοια της ασφάλειας των δικτύων υπολογιστών συνδέεται στενά με τρεις βασικές έννοιες (5) (6) (3) (2) (4).

- **Διαθεσιμότητα (Availability)**

Διαθεσιμότητα ονομάζεται η ιδιότητα του να είναι προσπελάσιμες και χωρίς αδικαιολόγητη καθυστέρηση οι υπηρεσίες ενός δικτύου υπολογιστών όταν τις χρειάζεται μια εξουσιοδοτημένη οντότητα. Για τους σκοπούς της ασφάλειας, μας απασχολεί βασικά η παρεμπόδιση κακόβουλων επιθέσεων που αποσκοπούν στο να παρακωλύσουν την πρόσβαση των νόμιμων χρηστών σε ένα πληροφοριακό σύστημα. Αυτές οι επιθέσεις ονομάζονται επιθέσεις άρνησης παροχής υπηρεσιών.

- **Εμπιστευτικότητα (Confidentiality)**

Εμπιστευτικότητα σημαίνει πρόληψη μη εξουσιοδοτημένης αποκάλυψης πληροφοριών, δηλαδή, πρόληψη από μη εξουσιοδοτημένη ανάγνωση. Επομένως, τα δεδομένα που διακινούνται μεταξύ των υπολογιστών ενός δικτύου, αποκαλύπτονται μόνο σε εξουσιοδοτημένα άτομα. Άλλες εκφάνσεις της εμπιστευτικότητας είναι:

1. Ι**διωτικότητα**, προστασία των δεδομένων προσωπικού χαρακτήρα, δηλαδή αυτών που αφορούν συγκεκριμένα πρόσωπα και

2. Μ**υστικότητα**, προστασία των δεδομένων που ανήκουν σε έναν οργανισμό ή μια επιχείρηση.

- **Ακεραιότητα (Integrity)**

Η *ακεραιότητα* μπορεί να οριστεί γενικότερα ως η απαίτηση να είναι τα πράγματα όπως πρέπει να είναι. Στην πληροφορική, ακεραιότητα σημαίνει πρόληψη μη εξουσιοδοτημένης μεταβολής πληροφοριών, δηλαδή, πρόληψη από μη εξουσιοδοτημένη εγγραφή ή διαγραφή, συμπεριλαμβανομένης και της μη εξουσιοδοτημένης δημιουργίας δεδομένων.

2.3 Μέθοδοι Επίθεσης

- ***Denial-of-Service (DoS):*** Αποστολή περισσοτέρων αιτήσεων σύνδεσης από όσες μπορεί να επεξεργαστεί ένας server

- ***Μη εξουσιοδοτημένη πρόσβαση (Unauthorized access attacks)***: διάφοροι τρόποι επίθεσης που εμπεριέχουν την ανάκτηση του δικαιώματος εισόδου, εκτέλεσης εντολών, ή ανάκτησης πληροφορίας σε ένα μηχάνημα που δεν παρέχει τέτοιες υπηρεσίες στον επιτιθέμενο<

- ***Password attacks***: Αποτελεί την μέθοδο εύρεσης ενός password,είτε με επαναληπτικό τρόπο δοκιμάζοντας όλους τους δυνατούς συνδυασμούς, είτε με αποκρυπτογράφηση του password δοκιμάζοντας όλους τους δυνατούς συνδυασμούς των πιθανών κλειδιών κρυπτογράφησης

- ***Trojan Horses***: Είναι ένα πρόγραμμα που περιέχει η εγκαθιστά μία «κακόβουλη» (malicious) εφαρμογή

- ***Network packet sniffers***: Είναι ένα πρόγραμμα ή μηχάνημα το οποίο μπορεί να υποκλέψει κίνηση που μεταφέρεται από ένα δίκτυο.[3][4]

2.3.1.1 Επιθέσεις κατά της εμπιστευτικότητας

Παράδειγμα τέτοιων επιθέσεων είναι:

- Μη εξουσιοδοτημένη φυσική πρόσβαση στο υπολογιστικό κέντρο

- Μη εξουσιοδοτημένη λογική πρόσβαση σε σύστημα.

- Είσοδος κακόβουλου λογισμικού όπως είναι οι Δούρειοι ίπποι (Trojans).

- κ.ο.κ

2.3.1.2 Επιθέσεις κατά της ακεραιότητας

Παράδειγμα τέτοιων επιθέσεων είναι:

- Μη εξουσιοδοτημένη λογική πρόσβαση σε σύστημα.
- Είσοδος κακόβουλου λογισμικού όπως είναι οι ιοί και οι Δούρειοι ίπποι (Trojans).
- Λανθασμένη είσοδος δεδομένων από το προσωπικό
- κ.ο.κ.

Βλέπετε ότι κάποιες επιθέσεις είναι κοινές σε περισσότερες κατηγορίες.

2.3.1.3 Επιθέσεις κατά της διαθεσιμότητας

Παράδειγμα τέτοιων επιθέσεων είναι:

- Βλάβη υπολογιστή.
- Επίθεση άρνησης υπηρεσίας.
- Φυσική καταστροφή.
- Κ.ο.κ

2.4 Τρόποι Αποφυγής Επιθέσεων

- **Έλεγχος γνησιότητας της ταυτότητας** (identification and authentication) των χρηστών , των προγραμμάτων ή των μηχανημάτων καθώς και των εξουσιοδοτήσεων που αυτά διαθέτουν για την προσπέλαση των προστατευμένων πόρων του συστήματος με συνδυασμένη χρήση συνθηματικών και ψηφιακών πιστοποιητικών

- **Προστασία της εμπιστευτικότητας των δεδομένων** (data confidentiality) , δηλαδή προστασία ενάντια σε μη εξουσιοδοτημένες αποκαλύψεις πληροφοριών

- Αποφυγή συστημάτων με **"single points of failure"**

- **Firewall**, το οποίο είναι ένα πρόγραμμα ή ένα μηχάνημα που μπορεί να χρησιμοποιηθεί σαν διαχωριστικό μεταξύ των δύο αυτών δικτύων

- **Κωδικοποίηση / Κρυπτογράφηση**

- **Πρωτόκολλα Ασφαλείας**, που αναφέρονται στα επίπεδα Πρόσβασης Δικτύου,Internet,Μεταφοράς και Εφαρμογής

- **Ενημέρωση** σχετικά με τα *operating systems* και κάποια *patches*

- **Αποφυγή τοποθέτησης δεδομένων** *σε σημεία οπού δεν είναι κατανοητά*

Κεφάλαιο 2ο

3 Εισαγωγή στις Αποθήκες και στην Εξόρυξη Δεδομένων

Η εποχή μας αναφέρεται και ως η “εποχή της πληροφορίας”. Ένας από τους λόγους για αυτόν το χαρακτηρισμό, είναι η δυνατότητα συλλογής μεγάλων όγκων δεδομένων. Μερικά χαρακτηριστικά παραδείγματα είναι τα εξής (10), (11), (12):

- Δεδομένα επιχειρήσεων: περιλαμβάνουν δεδομένα πωλήσεων (π.χ., ραβδοκωδικοί, ηλεκτρονικό επιχειρείν, κ.λπ.), συναλλαγών (π.χ., ATM), οικονομικών ενεργειών (π.χ., πωλήσεις, αγορές, χρηματιστήριο), ενδο-επιχειρηματικών ενεργειών (π.χ., προμήθειες, απογραφές). Τέτοιου είδους δεδομένα δημιουργούν συλλογές της τάξης αρκετών terabytes.

- Επιστημονικά δεδομένα: Ίσως δεν είναι υπερβολή να ισχυρισθούμε ότι σήμερα η επιστημονική έρευνα καθοδηγείται από τα δεδομένα (data driven). Αν τον 18ο αιώνα αρκούσαν μερικές εκατοντάδες αστρονομικές παρατηρήσεις για να διατυπωθεί ένας φυσικός νόμος (π.χ., βαρυτικός), σήμερα οι επιστήμονες προκειμένου να αναπτύξουν νέες θεωρίες, πρέπει να βασισθούν σε πολύ μεγάλο όγκο μετρήσεων, ώστε να ερμηνεύσουν πολύπλοκα φαινόμενα. Για παράδειγμα, η μελέτη των κλιματολογικών αλλαγών απαιτεί τη συλλογή πολλών terabytes δεδομένων, από αισθητήρες, δορυφόρους, κ.α., σε παγκόσμια κλίμακα και σε μεγάλη λεπτομέρεια ως προς το χρόνο. Μία άλλη επιστημονική περιοχή που βασίζεται στη μελέτη μεγάλων συλλογών δεδομένων είναι η βιολογία. Για παράδειγμα, η μελέτη της αλληλεπίδρασης των γονιδίων και η συσχέτισή τους με ασθένειες μπορεί να οδηγήσει σε νέες θεραπείες.

- Δεδομένα μορφής κειμένου: Λόγω των νέων τρόπων επικοινωνίας (π.χ. email, παγκόσμιος ιστός), τα δεδομένα με τη μορφή κειμένου πολλαπλασιάζονται με γρήγορους ρυθμούς. Ως συνέπεια προέκυψαν νέες ανάγκες και προβλήματα. Το χαρακτηριστικότερο παράδειγμα είναι η ανάγκη για εντοπισμό σελίδων του παγκόσμιου ιστού, οι οποίες περιέχουν ενδιαφέρουσα πληροφορία. Ένα άλλο παράδειγμα είναι η ανάγκη για φιλτράρισμα του συνήθως μεγάλου αριθμού emails που δέχεται κάποιος, και ειδικά η αποφυγή του spam.

- Δεδομένα του παγκόσμιου ιστού: Υπάρχουν πολλοί τρόποι θεώρησης του περιεχομένου του παγκόσμιου ιστού πέραν από την απλοϊκή θεώρησή τους ως κειμένων. Η ύπαρξη υπερκειμένου (μέσω συνδέσμων), πολυμεσικής πληροφορίας (όπως, εικόνες, ήχος, βίντεο), δεδομένων καταγραφής (π.χ., server logs), καθιστούν τον παγκόσμιο ιστό ως μία τεράστια και περίπλοκη συλλογή δεδομένων, η οποία αυξάνεται ραγδαία.

- Άλλες πηγές δεδομένων: Για να δοθεί απλώς η εικόνα της πολλαπλότητας και της διαφορετικότητας των διαφόρων διαθέσιμων συλλογών δεδομένων, αναφέρουμε επιγραμματικά κάποιες από αυτές: αθλητικά δεδομένα (NBA, σκάκι), δεδομένα ασφάλειας (κάμερες παρακολούθησης), βιομηχανικά δεδομένα (σχέδια, μηχανές), γεωγραφικά δεδομένα (πολεοδομικά, δορυφορικά).

Η τεχνολογία των βάσεων δεδομένων (databases) αναπτύχθηκε για την αποτελεσματική οργάνωση συλλογών δεδομένων. Η καθιέρωση της τεχνολογίας αυτής οδήγησε στην περαιτέρω ανάπτυξη του αριθμού και του μεγέθους των συλλογών δεδομένων, καθώς συστηματοποιήθηκε η διαδικασία. Το αποτέλεσμα ήταν η δημιουργία ενός νέου προβλήματος: είναι πλέον διαθέσιμες πολύ μεγάλες συλλογές ετερογενών βάσεων δεδομένων. Για παράδειγμα, είναι δυνατόν σε μία εταιρεία να διαθέτουμε μία βάση δεδομένων για το τμήμα πωλήσεων και μία άλλη για το τμήμα μάρκετινγκ. Ενδεχομένως, οι δύο βάσεις να αναπτύχθηκαν σε διαφορετικές χρονικές περιόδους και να ακολουθήθηκαν διαφορετικές αρχές

σχεδιασμού και συμβάσεις (π.χ., διαφορετικός τρόπος καταγραφής των πωλούμενων προϊόντων από αυτά που προωθούνται διαφημιστικά). Αυτό καθιστά δύσκολη τη σύνδεση των δεδομένων τους, ώστε να προκύψουν χρήσιμα συμπεράσματα, όπως αν η διαφημιστική καμπάνια ενός προϊόντος όντως οδήγησε σε αύξηση των πωλήσεών του.

Επομένως, προκύπτει ότι είναι επιτακτική η επίλυση του προβλήματος της ετερογένειας των βάσεων δεδομένων σε έναν οργανισμό/εταιρεία. Παρά ταύτα, όμως, τίθεται το εξής ερώτημα: ποια είναι εξ αρχής η χρησιμότητα των δεδομένων; Με άλλα λόγια, για ποιο λόγο συλλέγουμε όλα αυτά τα δεδομένα; Η προφανής απάντηση είναι ότι στις περισσότερες περιπτώσεις είμαστε υποχρεωμένοι να το κάνουμε. Για παράδειγμα, μία τράπεζα οφείλει να καταγράφει τις συναλλαγές των πελατών της, μία εταιρεία οφείλει να διατηρεί αρχεία πωλήσεων, κ.λπ. Αυτή είναι η αναγκαία θεώρηση των συλλογών δεδομένων: τις δημιουργούμε γιατί οφείλουμε να το κάνουμε. Αυτή η θεώρηση, όμως, οδήγησε στην παρομοίωση των συλλογών δεδομένων με ‘τάφους δεδομένων”. Ο λόγος είναι ότι συλλέγουμε δεδομένα απλώς για να τα αποθηκεύσουμε σε μία συλλογή δεδομένων, σε περίπτωση που χρειασθεί να τα ανακτήσουμε, π.χ., για ενημέρωση τραπεζικών λογαριασμών, φορολογικούς ελέγχους κ.α.

Ένα προσφιλές παράδειγμα στον Ελληνικό χώρο είναι ο δημόσιος τομέας, όπου διάφορα υπουργεία και οργανισμοί έχουν τις δικές τους βάσεις δεδομένων τους, αλλά λόγω της ετερογένειας τους δεν υπάρχει η δυνατότητα επικοινωνίας μεταξύ τους, με συνέπεια τις χρονοβόρες διαδικασίες που ο πολίτης υφίσταται.

Ο λόγος που η εποχής μας αποκαλείται ως η “εποχή της πληροφορίας” είναι ότι έγινε αντιληπτή η ισοδυναμία της πληροφορίας με ισχύ. Σωστές πληροφορίες μπορεί να οδηγήσουν σε αποφάσεις που μπορούν να επιφέρουν πλεονεκτήματα, π.χ., αύξηση πωλήσεων, νέες επιστημονικές ανακαλύψεις. Για τη λήψη σωστών αποφάσεων πρέπει να ανατρέξουμε στα ίδια τα δεδομένα, να τα κατανοήσουμε, και να ανακαλύψουμε πληροφορίες που αυτά εμπεριέχουν, οι οποίες θα μας οδηγήσουν σε ορθές αποφάσεις. Για παράδειγμα, έστω ότι για ένα νέο τύπο φορητού υπολογιστή αρχικά είχε υποτεθεί ότι η κατάλληλη αγοραστική ομάδα είναι έφηβοι ηλικίας 15-19 ετών από οικογένειες με υψηλά εισοδήματα. Μετά από ένα διάστημα, διαπιστώνεται ότι οι πωλήσεις του δεν είναι καθόλου ικανοποιητικές. Αν από δεδομένα πωλήσεων και δεδομένα μάρκετινγκ (π.χ., ερωτηματολόγια) μπορεί να προκύψει η πληροφορία, ότι για το προϊόν αυτό έγινε λανθασμένη αγοραστική τοποθέτηση, και ότι η κατάλληλη αγοραστική ομάδα είναι άνδρες ηλικιών 25-30 με υψηλό μορφωτικό επίπεδο, τότε αυτό μπορεί να οδηγήσει σε καλύτερο μάρκετινγκ και αύξηση των πωλήσεων. Σε ένα άλλο παράδειγμα, από τα δεδομένα μέτρησης του CO2 στην ατμόσφαιρα στις 12 Σεπτεμβρίου του 2001 παρατηρήθηκε απότομη μείωση των επιπέδων του σε παγκόσμιο επίπεδο. Ο συνδυασμός με δεδομένα πτήσεων οδήγησε στην πληροφορία ότι οι νυκτερινές πτήσεις (που το προηγούμενο βράδυ δεν έγιναν λόγων των τραγικών γεγονότων στις 11 Σεπτεμβρίου 2001) επιβαρύνουν σημαντικά το φαινόμενο του θερμοκηπίου. Αυτή η πληροφορία λειτούργησε σε επανασχεδιασμό/νέες τιμολογήσεις των νυκτερινών πτήσεων.

Παρά την υπερβολική διαθεσιμότητα δεδομένων σήμερα, το μέγεθος και η ετερογένειά τους καθιστούν δύσκολη την εύρεση πληροφορίας που μπορεί να οδηγήσει σε λήψη αποφάσεων. Για παράδειγμα, στον Κέπλερ χρειάσθηκαν ευάριθμες σελίδες με μετρήσεις κινήσεων πλανητών, για να διατυπώσει το νόμο της ελλειπτικής τους κίνησης. Η ανακάλυψη μίας νέας γονιδιακής θεραπείας, σήμερα, απαιτεί εξέταση πολυάριθμων μετρήσεων που βρίσκονται αποθηκευμένες σε ετερογενείς βάσεις (π.χ., δεδομένα φαρμακευτικών εταιριών, κλινικών μετρήσεων σε νοσοκομεία, ψηφιακών βιβλιοθηκών σε εργαστήρια γενετικής). Επιπλέον, μιλώντας με απλουστευτικούς όρους και ξεπερνώντας το πρόβλημα της ετερογένειας, ας θεωρήσουμε τα δεδομένα με την κλασική σχεσιακή μορφή τους, δηλαδή, ως πίνακες. Διάφορες μεθοδολογίες, όπως η στατιστική, μπορούν να βοηθήσουν την ανίχνευση πληροφοριών σε πίνακες λίγων γραμμών και στηλών. Με την αύξηση του αριθμού τόσο των στηλών (χιλιάδες) όσο και των γραμμών (δισεκατομμύρια), κάτι τέτοιο είναι πρακτικά αδύνατον. Χρειαζόμαστε νέες μεθοδολογίες για την εύρεση της πληροφορίας εντός μεγάλων και ετερογενών βάσεων δεδομένων. Διαφορετικά, οι βάσεις δεδομένων μας θα εξακολουθήσουν να είναι απλώς “τάφοι” των δεδομένων μας [9],10], [11] ,[12] .

4 Εξόρυξη Δεδομένων

4.1 Ορισμός Εξόρυξης Γνώσης και Δεδομένων

Η εξόρυξη γνώσης από δεδομένα (data mining) ή πιο απλά η εξόρυξη γνώσης είναι μια νέα δυναμική τεχνολογία που βοηθάει τις επιχειρήσεις να εστιάσουν στην σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (data warehouses). Οι τεχνικές της είναι σε θέση να αναζητήσουν και να βρουν γρήγορα και λεπτομερειακά βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (patterns). Ουσιαστικά, η εξόρυξη γνώσης είναι μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων (13), (11), (12).

«Εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής υπονοούμενης και εν πολλοίς άγνωστης αλλά ενδεχομένως χρήσιμης γνώσης υπό την μορφή συσχετίσεων προτύπων και τάσεων, μέσω της εξέτασης ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, την στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση.» (13).

Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι ο στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού ποικίλουν σε πολύ υψηλό βαθμό. Η εξόρυξη γνώσης περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν, την στατιστική ανάλυση (statistical analysis), τα δένδρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), την εξαγωγή κανόνων (rule induction) και την γραφική οπτικοποίηση (graphic visualization).

Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση εργαλείων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων μέσω σχέσεων μεταξύ των χαρακτηριστικών των βάσεων δεδομένων.

Η ανακάλυψη γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases KDD) αναφέρεται στη διεργασία εξόρυξης γνώσης από τις μεγάλες αποθήκες δεδομένων. Ο όρος εξόρυξη δεδομένων χρησιμοποιείται ως συνώνυμο της ανακάλυψης γνώσης από βάσεις δεδομένων, καθώς επίσης και για αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και την εξαγωγή της από διάφορα σύνολα δεδομένων. Για να διαφοροποιηθούμε μεταξύ της διαδικασίας και των εργαλείων, θα χρησιμοποιήσουμε τον πρώτο όρο, KDD, για να περιγράψουμε ολόκληρη τη διαδικασία ανάλυσης ενός συνόλου δεδομένων, και το δεύτερο όρο, την εξόρυξη δεδομένων, για να αναφερθούμε κυρίως στις μεθόδους και τις τεχνικές που χρησιμοποιούνται στη διαδικασία ανάλυσης. Πολλοί ερευνητές θεωρούν τον όρο εξόρυξη δεδομένων μη αντιπροσωπευτικό της διαδικασίας που αντιπροσωπεύει, υποστηρίζοντας ότι ο όρος εξόρυξη γνώσης θα ήταν μια πιο κατάλληλη περιγραφή • Εντούτοις, ένας τέτοιος όρος μπορεί να μην δίνει έμφαση στην ανάλυση και την εξαγωγή των προτύπων από μεγάλα σύνολα δεδομένων. ο όρος εξόρυξη δεδομένων (Data Mining) είναι αυτός που έχει επικρατήσει και χαρακτηρίζει τη διαδικασία της εύρεσης δομών γνώσης οι οποίες περιγράφουν με ακρίβεια μεγάλα σύνολα πρωτογενών δεδομένων. οι δομές αυτές αναδεικνύουν γνώση (συσχετίσεις ή κανόνες) που είναι κρυμμένοι μέσα στα δεδομένα και δεν μπορούν να εξαχθούν από τον άνθρωπο-χρήστη της βάσης με «γυμνό» μάτι. οι προκύπτουσες δομές είναι πλούσιες σε σημασιολογία και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων (14).

4.2 Η ανακάλυψη γνώσης από βάσεις δεδομένων (KDD) σε σχέση με την εξόρυξη δεδομένων.

Η ανακάλυψη γνώσης από μία βάση δεδομένων (KDD) αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων. Ένας γενικός ορισμός, που παρουσιάζει με περισσότερη σαφήνεια την έννοια του όρου KDD δόθηκε από τους Frawley, Piatesky-Shaphiro & Matheus (13), σύμφωνα με τον οποίο:

«KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα».

Για την κατανόηση του προαναφερθέντος ορισμού, θα εξετάζονται ακολούθως λεπτομερώς οι βασικές έννοιες των όρων στους οποίους είναι βασισμένος (13):

1. Τα δεδομένα περιγράφουν οντότητες ή συσχετίσεις του πραγματικού κόσμου, F. Παραδείγματος χάριν θα μπορούσε να είναι ένα σύνολο εγγραφών που αναφέρονται στις συναλλαγές τραπεζών, οι οποίες να περιέχουν τις τιμές τριών ιδιοτήτων (π.χ. τμήμα, εισόδημα, κατάσταση δανείου).
2. Ένα πρότυπο είναι μια έκφραση E σε μια γλώσσα L η οποία περιγράφει ένα υποσύνολο δεδομένων F£CF εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του. Σε αυτή την περίπτωση το πρότυπο θεωρείται υποσύνολο του F και αφαίρεση (abstraction) του F.
3. Η διαδικασία KDD είναι μια διαδικασία πολλαπλών βημάτων, η οποία περιλαμβάνει την προ-επεξεργασία των δεδομένων, την αναζήτηση των προτύπων και την αξιολόγηση της εξαγόμενης γνώσης.
4. Εγκυρότητα. Το εξαγόμενο πρότυπο θα πρέπει να είναι συνεπές σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας. Το ζήτημα της εγκυρότητας αποτελεί ένα από τα βασικά προβλήματα και αντικείμενο έρευνας στην εξόρυξη δεδομένων.
5. Πιθανά χρήσιμο. Η εξαγωγή των προτύπων θα πρέπει να ακολουθείται από μερικές χρήσιμες διεργασίες όπως η αξιολόγηση τους από κάποιες συναρτήσεις χρησιμότητας. Επίσης, θα ήταν χρήσιμο να εμπλουτιστεί η σημασιολογία τους, διατηρώντας όσο το δυνατόν περισσότερη γνώση από τα αρχικά δεδομένα η οποία μπορεί να φανεί χρήσιμη για τη λήψη αποφάσεων. Παραδείγματος χάριν, σε περίπτωση μίας βάσης δεδομένων που αφορά σε δάνεια, σαν χρήσιμη διαδικασία θα μπορούσε να θεωρηθεί αυτή που θα δίνει μια ένδειξη της αναμενόμενης αύξησης στα κέρδη. Συνδέεται επίσης με τον ακόλουθα κανόνα απόφασης: ««Εάν έσοδα < $t, τότε ο πελάτης δεν μπορεί να πάρει δάνειο».
6. Τελικά κατανοητό. Ο στόχος της εξόρυξης γνώσης είναι να προσδιοριστούν τα πρότυπα και να γίνουν κατανοητά, ώστε να μπορούν να οδηγήσουν ακόμη και τους μη ειδικούς σε χρήσιμα συμπεράσματα και αποφάσεις..

4.3 Σύγκριση διαφόρων τεχνολογιών δεδομένων

Γίνεται κατανοητό ότι σε σύγκριση με την απλή ανάκτηση από βάσεις δεδομένων με τη βοήθεια της γλώσσας SQL, τεχνολογίες όπως η OLAP και η εξόρυξη δεδομένων εξάγουν χρήσιμη πληροφορία. Αν και η εξόρυξη δεδομένων δεν απαιτεί το σχηματισμό προηγούμενων υποθέσεων, η εξόρυξη και η OLAP μπορούν να λειτουργήσουν συμπληρωματικά, π.χ., να γίνει μία αρχική ανάλυση με OLAP, ώστε να αποκτηθεί μία γνώση των ιδιοτήτων των δεδομένων, και στη συνέχεια μια πιο στοχευόμενη ανάλυση με εξόρυξη δεδομένων. Ο Πίνακας 1 συνοψίζει τα χαρακτηριστικά των προαναφερθεισών τεχνολογιών συγκριτικά με τις συμβατικές (σχεσιακές ΒΔ) [10].

Τεχνολογία	Ερωτήματα	Χαρακτηριστικά
Αρχεία	“Ποιο το σύνολο πωλήσεων τα τελευταία 5 έτη;”	Ανάκτηση στατικών δεδομένων
Σχεσιακές ΒΔ	“Δώσε το σύνολο πωλήσεων στη Μακεδονία φέτος.”	Ανάκτηση δυναμικών δεδομένων σε επίπεδο εγγραφής
Αποθήκες δεδομένων	“Δώσε το σύνολο πωλήσεων στη Μακεδονία φέτος. Ανάλυσε κατά νομό και μήνα.”	Ανάκτηση δυναμικών, ετερογενών δεδομένων, σε πολλαπλά επίπεδα
Εξόρυξη δεδομένων	“Σε ποιες περιοχές θα αυξηθούν οι πωλήσεις το επόμενο 3μηνο;”	Ανακάλυψη πληροφορίας, πρόβλεψη

Πίνακας 1. Σύγκριση διαφόρων τεχνολογιών δεδομένων

Οι αποθήκες δεδομένων προήρθαν από την εξέλιξη της τεχνολογίας των βάσεων δεδομένων. Η επιστημονική περιοχή της εξόρυξη δεδομένων βασίζεται σε προγενέστερες, όπως η στατιστική, η μηχανική μάθηση, οι βάσεις δεδομένων κ.α. Η εξόρυξη δεδομένων συνδυάζει μεθόδους από όλες αυτές τις περιοχές, με την ειδοποιό διαφορά ότι θεωρεί μεγάλες συλλογές δεδομένων. Για παράδειγμα, στη στατιστική αναπτύσσονται θεωρητικά μοντέλα, τα οποία, καθοδηγούμενα από τον αναλυτή, εφαρμόζονται σε μικρά δείγματα (λίγες γραμμές και στήλες) και έχουν ακριβή αποτελέσματα. Η εφαρμογή τους σε πολλά και περίπλοκα δεδομένα δεν είναι πάντοτε εφικτή, τόσο επειδή ο χρόνος εκτέλεσης τέτοιων μεθόδων είναι αποτρεπτικός όσο και επειδή ο αναλυτής δεν είναι σε θέση να καθοδηγήσει τη διαδικασία.

Η εξόρυξη δεδομένων έχει κάποιους κοινούς στόχους αλλά και αρκετές διαφορές ως προς τη μηχανική μάθηση:

- Η εξόρυξη δεδομένων στοχεύει στην εύρεση πληροφορίας, που θα κατανοηθεί από αναλυτές και θα τους οδηγήσει σε αποφάσεις. Αντίθετα, στη μηχανική μάθηση, η εύρεση πληροφορίας σκοπεύει στη βελτίωση της απόδοσης κάποιας τεχνητής οντότητας, όπως ενός ρομπότ, ενός πράκτορα (agent) κ.α.

- Η εξόρυξη δεδομένων αντιμετωπίζει μεγάλους όγκους δεδομένων, ενώ αντίθετα στη μηχανική μάθηση χρησιμοποιούνται (συνήθως) μικρά δείγματα. Ως αποτέλεσμα, οι αλγόριθμοι της εξόρυξης δεδομένων αποσκοπούν στην ελαχιστοποίηση της χρονικής πολυπλοκότητας και στην κλιμάκωση σε δεδομένα στη δευτερεύουσα μνήμη.

Η εξόρυξη δεδομένων εφαρμόζεται σε in-vivo συλλογές δεδομένων, οι οποίες συχνά δεν σχεδιάσθηκαν για την εξόρυξή τους. Αντιθέτως, η μηχανική μάθηση εφαρμόζεται συνήθως σε

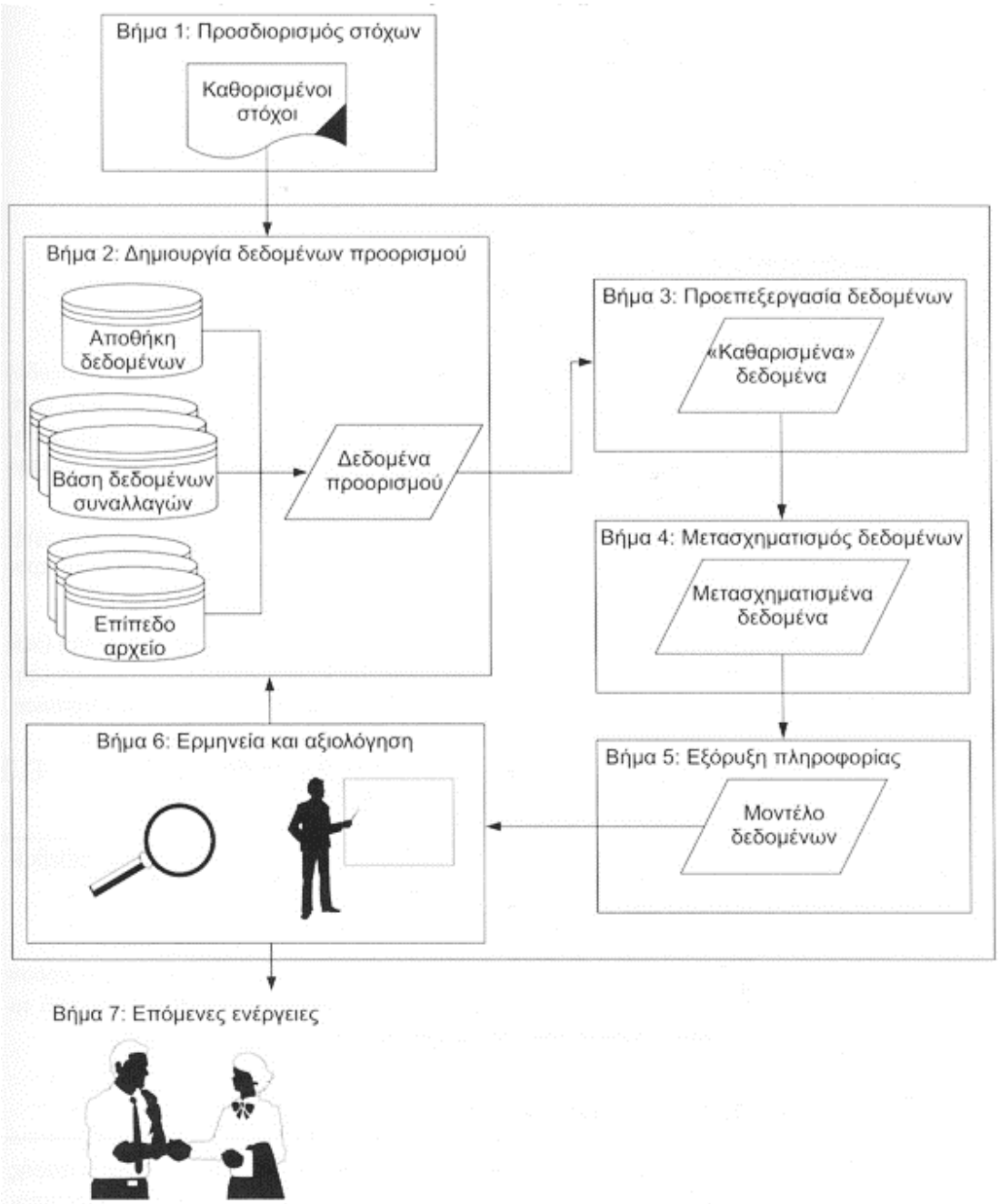
in-vitro δεδομένα που συλλέχθηκαν ώστε να εφαρμοσθεί σε αυτά κάποιος αλγόριθμος μάθησης.

4.4 Διαδικασία Εξόρυξης Γνώσης

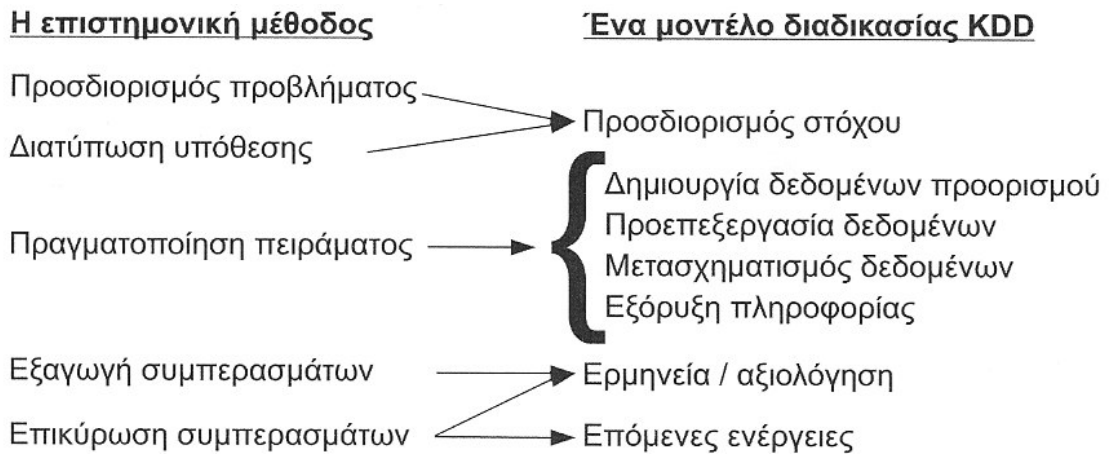
Η ανακάλυψη γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases-KDD) είναι μια αλληλεπιδραστική, επαναληπτική διαδικασία που προσπαθεί να εξάγει υπονοούμενες, ενδεχομένως χρήσιμες γνώσεις από δεδομένα, οι οποίες δεν ήταν γνωστές νωρίτερα. Υπάρχουν πολλές παραλλαγές αυτού που είναι πλέον γνωστό ως μοντέλο διαδικασίας KDD. Αυτές οι παραλλαγές περιγράφουν τη διαδικασία KDD σε 4 μέχρι και 12 βήματα. Αν και ο αριθμός των βημάτων μπορεί να διαφέρει, οι περισσότερες περιγραφές είναι συνεπείς ως προς το περιεχόμενο. Προτιμούμε να χαρακτηρίζουμε τη διαδικασία KDD ως προσέγγιση επτά βημάτων για την ανακάλυψη γνώσης. Το μοντέλο διαδικασίας KDD επτά βημάτων φαίνεται στην Εικόνα 5. Ακολουθεί μια σύντομη περιγραφή του κάθε βήματος (11), (10), (15) (16) (17) (12):

1. **Προσδιορισμός στόχου.** Το επίκεντρο αυτού του βήματος είναι η κατανόηση του πεδίου εφαρμογής που εξετάζεται για εξόρυξη γνώσης. Διατυπώνουμε ξεκάθαρα τι πρόκειται να επιτευχθεί. Μπορούμε επίσης να διατυπώσουμε υποθέσεις για πιθανά ή επιθυμητά αποτελέσματα.
2. **Δημιουργία ενός συνόλου δεδομένων προορισμού.** Με την βοήθεια ενός ή περισσότερων ειδικών και εργαλείων ανακάλυψης γνώσης, επιλέγεται ένα αρχικό σύνολο δεδομένων που πρόκειται να αναλυθεί.
3. **Προεπεξεργασία δεδομένων.** Χρησιμοποιούνται οι διαθέσιμοι πόροι προς διαχείριση των δεδομένων , τα οποία έχουν θόρυβο. Διαχείριση των τιμών δεδομένων που λείπουν και πληροφοριών χρονικής αλληλουχίας.
4. **Μετασχηματισμός δεδομένων.** Προστίθενται και εξαλείφονται χαρακτηριστικά και στιγμιότυπα από τα δεδομένα προορισμού. Λήψη απόφασης ποιες μεθόδους θα χρησιμοποιηθούν για την κανονικοποίηση, μετατροπή, και εξομάλυνση δεδομένων.
5. **Εξόρυξη γνώσης.** Δημιουργείται το βέλτιστο μοντέλο αναπαράστασης των δεδομένων με την εφαρμογή ενός ή περισσότερων αλγορίθμων εξόρυξης γνώσης .
6. **Ερμηνεία και αξιολόγηση.** Εξέταση των αποτελεσμάτων (την έξοδο) του Βήματος 5 για να καθορισθεί αν αυτό που ανακαλύφθηκε είναι χρήσιμο και ενδιαφέρον. Αποφασίζουμε αν θα επαναλάβουμε προηγούμενα βήματα με τη χρήση καινούργιων χαρακτηριστικών ή στιγμιοτύπων.
7. **Αξιοποίηση εξορυχθείσας γνώσης .** Αν η γνώση που ανακαλύφθηκε θεωρείται χρήσιμη, τότε ενσωματώνεται και εφαρμόζεται άμεσα σε κατάλληλα προβλήματα.

Η μέθοδος KDD αποτελεί την εφαρμογή της επιστημονικής μεθόδου (scientific method) στην εξόρυξη πληροφορίας. Το 1620, ο Sir Francis Bacon εξήγησε για πρώτη φορά την επιστημονική μέθοδο στο βιβλίο *Novum Organum.*



Εικόνα 2. Μοντέλο διαδικασίας KDD 7 βημάτων



Εικόνα 3. Εφαρμογή της επιστημονικής μεθόδου στην Εξόρυξη Γνώσης

Παρουσίαση της επιστημονικής μέθοδο ως διαδικασία τεσσάρων βημάτων:

1. Προσδιορισμός του προβλήματος που πρέπει να επιλυθεί.
2. Διατύπωση υπόθεσης.
3. Πραγματοποίηση ενός ή περισσότερων πειραμάτων για την επιβεβαίωση ή τη διάψευση της υπόθεσης.
4. Εξαγωγή και επικύρωση συμπερασμάτων.

4.4.1 Βήμα 1: Προσδιορισμός στόχου

Ένας βασικός αντικειμενικός σκοπός του προσδιορισμού στόχου είναι ο σαφής καθορισμός του τι πρέπει να επιτευχθεί. Το πρώτο βήμα είναι, από πολλές απόψεις, το πιο δύσκολο αφού πρέπει να ληφθούν αποφάσεις σχετικά με την κατανομή πόρων και τη μέτρηση της επιτυχίας. Οποτεδήποτε είναι δυνατό, οι γενικοί στόχοι θα πρέπει να διατυπώνονται με τη μορφή συγκεκριμένων αντικειμενικών σκοπών. Στη συνέχεια ακολουθεί ένας μερικός κατάλογος πραγμάτων που πρέπει να ληφθούν υπόψη σε αυτό το στάδιο:

- Δίνεται μια σαφής δήλωση του προβλήματος, καθώς και μια λίστα κριτηρίων για τη μέτρηση της επιτυχίας ή της αποτυχίας. Μπορούν να διατυπωθούν μία ή περισσότερες υποθέσεις για πιθανά ή επιθυμητά αποτελέσματα.
- Επιλέγεται το εργαλείο ή το σύνολο εργαλείων εξόρυξης πληροφορίας. Η επιλογή αυτή εξαρτάται από πολλούς παράγοντες, στους οποίους συγκαταλέγονται το απαιτούμενο επίπεδο ερμηνείας των δεδομένων και το αν η εκ- μάθηση είναι καθοδηγούμενη, μη καθοδηγούμενη, ή ένας συνδυασμός των δύο τεχνικών.
- Γίνεται εκτίμηση του κόστους του έργου. Καταρτίζεται ένα σχέδιο για τη διαχείριση των ανθρώπινων πόρων.
- Δίνεται μια ημερομηνία ολοκλήρωσης του έργου/παράδοσης του προϊόντος.
- Λαμβάνονται υπόψη νομικά θέματα που μπορεί να προκύψουν από την εφαρμογή των αποτελεσμάτων της διαδικασίας ανακάλυψης.
- Παρέχεται ένα σχέδιο κατάλληλο για τη συντήρηση ενός συστήματος που βρίσκεται σε λειτουργία. Καθώς γίνονται διαθέσιμα νέα δεδομένα, ένα σημαντικό ζήτημα είναι η ύπαρξη μιας μεθοδολογίας για την ενημέρωση του μοντέλου λειτουργίας.

4.4.2 Βήμα 2: Δημιουργία ενός συνόλου δεδομένων προορισμού

Βασική θεωρείται η ύπαρξη ενός βιώσιμου συνόλου δεδομένων για την επιτυχία οποιουδήποτε έργου εξόρυξης πληροφορίας. Στην Εικόνα 5.1 απεικονίζεται η εξαγωγή δεδομένων προορισμού από τρεις βασικές πηγές - μια αποθήκη δεδομένων, μία ή περισσότερες βάσεις δεδομένων συναλλαγών, ή ένα ή περισσότερα επίπεδα αρχεία. Πολλά εργαλεία εξόρυξης πληροφορίας απαιτούν να είναι αποθηκευμένα τα δεδομένα εισόδου σε μορφή επιπέδου αρχείου ή σε μορφή λογιστικού φύλλου. Αν τα αρχικά δεδομένα είναι αποθηκευμένα σε επίπεδο

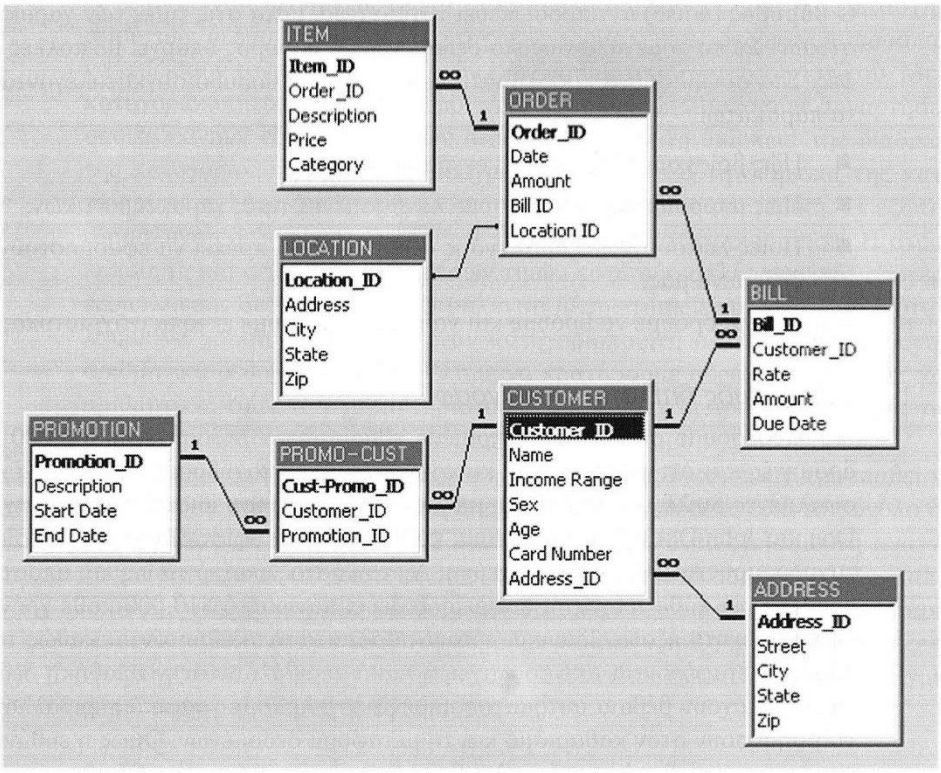
αρχείο, τότε η δημιουργία των αρχικών δεδομένων προορισμού γίνεται άμεσα. Ας εξετάσουμε τις άλλες πιθανότητες.

Τα συστήματα διαχείρισης βάσεων δεδομένων (database management systems-DBMS) αποθηκεύουν και χειρίζονται δεδομένα συναλλαγών. Οι εφαρμογές ενός DBMS είναι σε θέση να ενημερώνουν γρήγορα και να ανασύρουν πλη- ροφορίες από μια αποθηκευμένη βάση δεδομένων. Τα δεδομένα ενός DBMS συχνά ακολουθούν τη δομή του σχεσιακού μοντέλου. Μια **σχεσιακή βάση δεδομένων** (relational database) αντιπροσωπεύει τα δεδομένα ως συλλογή πινάκων οι οποίοι περιέχουν γραμμές και στήλες. Κάθε στήλη ενός πίνακα αντιπροσωπεύει ένα χαρακτηριστικό, ενώ σε κάθε γραμμή του πίνακα αποθηκεύονται οι πληροφορίες μίας εγγραφής δεδομένων. Οι επιμέρους γραμμές ονομάζονται **πλειάδες** (tuples). Όλες οι πλειάδες ενός σχεσιακού πίνακα προσδιορίζονται με μοναδικό τρόπο από ένα συνδυασμό ενός ή περισσοτέρων χαρακτηριστικών του πίνακα.

Ένας βασικός σκοπός του σχεσιακού μοντέλου είναι η μείωση του πλεονασμού των δεδομένων έτσι ώστε να είναι εφικτή η γρήγορη πρόσβαση στις πληροφορίες της βάσης δεδομένων. Με ένα σύνολο κανονικών μορφών που δεν επιτρέπουν τον πλεονασμό δεδομένων καθορίζονται κανόνες μορφοποίησης για σχεσιακούς πίνακες. Αν ένας σχεσιακός πίνακας περιέχει πλεονάζοντα δεδομένα, ο πλεονασμός καταργείται με την ανάλυση του πίνακα σε δύο ή περισσότερες σχεσιακές δομές. Αντίθετα, ο σκοπός της εξόρυξης γνώσης είναι η αποκάλυψη του εγγενούς πλεονασμού στα δεδομένα. Κατά συνέπεια, συνήθως απαιτούνται μία ή περισσότερες σχεσιακές λειτουργίες σύζευξης για την αναδόμηση των δεδομένων σε μια μορφή που επιδέχεται εξόρυξη πληροφορίας.

Η υποθετική βάση δεδομένων προώθησης πωλήσεων μέσω πιστωτικών καρτών που ορίζεται στον πίνακα 5. Τα χαρακτηριστικά του πίνακα: *income range, magazine promotion, watch promotion, life insurance promotion, credit card insurance, sex,* και *age.* Τα δεδομένα του πίνακα 5 δεν μπορούν καν να θεωρηθούν ως βάση δεδομένων, αλλά αντιπροσωπεύουν μια δομή επίπεδου αρχείου που έχει εξαχθεί από μια βάση δεδομένων, όπως αυτή που φαίνεται στην Εικόνα 5. Η βάση δεδομένων πιστωτικών καρτών της υποθετικής εταιρείας Acme περιέχει πίνακες με πληροφορίες χρεώσεων πιστωτικών καρτών και παραγγελίες, εκτός από τις πληροφορίες για την προώθηση προϊόντων μέσω καρτών.

Η εικόνα 3 δίνει, επίσης, τη δυνατότητα εξαγωγής δεδομένων από πολλές βάσεις δεδομένων ή πολλά αρχεία. Αν τα δεδομένα προορισμού πρόκειται να εξαχθούν από περισσότερες από μία πηγές, η διαδικασία μεταφοράς μπορεί να είναι κουραστική. Σκεφτείτε ένα απλό παράδειγμα στο οποίο αποθηκεύεται το γένος των πελατών σε μια επιχειρησιακή βάση δεδομένων με την κωδικοποίηση *male = 1, female = 2.* Σε μια δεύτερη βάση δεδομένων, το γένος αποθηκεύεται με την κωδικοποίηση *male = M* και *female = F.* Η κωδικοποίηση για *male* και *female* θα πρέπει να είναι συνεπής σε όλες τις εγγραφές των δεδομένων προορισμού, διότι διαφορετικά τα δεδομένα θα έχουν μικρή χρησιμότητα. Η διαδικασία επιβολής αυτής της συνέπειας κατά τη μετα- φορά δεδομένων είναι μια μορφή *μετασχηματισμού δεδομένων* (data transformation).



Εικόνα 4. Η ΒΔ πιστωτικών καρτών της εταιρείας Acme

Τέλος, μια τρίτη δυνατότητα για τη συγκομιδή δεδομένων προορισμού είναι η αποθήκη δεδομένων.

4.4.3 Βήμα 3: Προεπεξεργασία δεδομένων

Το μεγαλύτερο μέρος της προεπεξεργασίας δεδομένων αφορά τον **καθαρισμό** των δεδομένων (data cleaning), με τον οποίο αντιμετωπίζονται τα προβλήματα του θορύβου και των πληροφοριών που λείπουν. Σε ιδανικές συνθήκες, το μεγαλύτερο μέρος της προεπεξεργασίας δεδομένων γίνεται προτού τα δεδομένα αποθηκευτούν μόνιμα σε μια δομή, όπως μια αποθήκη δεδομένων.

Δεδομένα με θόρυβο

Ο **θόρυβος** (noise) αντιπροσωπεύει τυχαία σφάλματα στις τιμές των χαρακτηριστικών. Σε πολύ μεγάλα σύνολα δεδομένων, ο θόρυβος υπάρχει με πολλές μορφές. Στα συνηθισμένα προβλήματα δεδομένων με θόρυβο συγκαταλέγονται και τα παρακάτω:

Εντοπισμός διπλότυπων εγγραφών

Ας υποθέσουμε ότι μια συγκεκριμένη εβδομαδιαία έκδοση έχει 100,000 συνδρομητές και ότι το 0.1% των καταχωρίσεων της ταχυδρομικής λίστας έχει εσφαλμένες διπλές καταχωρίσεις με μια παραλλαγή του ίδιου ονόματος (πχ. Jon Doe και John Doe). Κατά συνέπεια, κάθε εβδομάδα υφίστανται επεξεργασία και ταχυδρομούνται 100 επιπλέον τεύχη. Αν το κόστος επεξεργασίας και αποστολής είναι $2 για κάθε τεύχος, η εταιρεία ξοδεύει πάνω από $10,000 κάθε χρόνο σε αδικαιολόγητα έξοδα. Ιδανικά, τέτοια σφάλματα ανακαλύπτονται καθώς τα δεδομένα μεταφέρονται από το επιχειρησιακό περιβάλλον στην αποθήκη δεδομέ- νων. Υπάρχουν βέβαια αυτοματοποιημένα εργαλεία με γραφικό περιβάλλον για να βοηθήσουν στον καθαρισμό και τη μεταφορά δεδομένων. Όμως η ευθύνη για τη μεταφορά δεδομένων εξακολουθεί να είναι στα χέρια του ειδικού της αποθήκης δεδομένων.

Εντοπισμός λανθασμένων τιμών χαρακτηριστικών

Ο εντοπισμός λαθών σε δεδομένα κατηγοριών αποτελεί πρόβλημα στα μεγάλα σύνολα δεδομένων. Μερικά εργαλεία εξόρυξης πληροφορίας δίνουν μια σύνοψη των τιμών συχνότητας και των ποσοστών προβλεψιμότητας για χαρακτηριστικά κατηγοριών. Θα πρέπει να θεωρούμε τις τιμές χαρακτηριστικών με ποσο- στά προβλεψιμότητας κοντά στο Ο ως υποψήφια λάθη.

Μηδενική τιμή για ένα χαρακτηριστικό όπως είναι η πίεση ή το βάρος αποτελεί προφανές λάθος. Τέτοια λάθη συμβαίνουν συχνά όταν λείπουν δεδομένα και αντιστοιχίζονται προεπιλεγμένες τιμές για να συμπληρώσουν τα κενά στοιχεία. Σε μερικές περιπτώσεις, τέτοια λάθη μπορούν να εντοπιστούν μόνο αν εξεταστούν οι τιμές του μέσου και της τυπικής απόκλισης των κατηγοριών. Όμως, αν το σύνολο δεδομένων είναι μεγάλο και υπάρχουν μόνο μερικές εσφαλμένες τιμές, ο εντοπισμός τέτοιων σφαλμάτων μπορεί να είναι δύσκολος. Μερικά εργαλεία ανάλυσης δεδομένων επιτρέπουν στο χρήστη να εισάγει ένα διάστημα έγκυρων τιμών για τα αριθμητικά δεδομένα. Στιγμιότυπα με τιμές χαρακτηριστικών έξω από τα όρια του έγκυρου διαστήματος επισημαίνονται ως υποψήφια λάθη.

Εξομάλυνση δεδομένων

Η εξομάλυνση δεδομένων είναι μια διαδικασία καθαρισμού και μετασχηματισμού των δεδομένων. Αρκετές τεχνικές εξομάλυνσης δεδομένων προσπαθούν να μειώσουν τον αριθμό των τιμών ενός αριθμητικού χαρακτηριστικού. Μερικοί κατηγοριοποιητές, όπως τα νευρωνικά δίκτυα, χρησιμοποιούν συναρτήσεις οι οποίες κάνουν εξομάλυνση των δεδομένων κατά τη διάρκεια της διαδικασίας της κατηγοριοποίησης. Όταν πραγματοποιείται κατά τη διάρκεια της κατηγοριοποίησης, λέμε ότι η εξομάλυνση δεδομένων είναι εσωτερική. Η εξωτερική εξομάλυνση δεδομένων διεξάγεται πριν από την κατηγοριοποίηση. Η στρογγυλοποίηση και ο υπολογισμός μέσων τιμών είναι δύο απλές τεχνικές εξωτερικής εξομάλυνσης δεδομένων. Η εξομάλυνση μέσης τιμής είναι κατάλληλη όταν θέλουμε να χρησιμοποιήσουμε έναν κατηγοριοποιητή , ο οποίος δεν υποστηρίζει αριθμητικά δεδομένα και θα θέλαμε να κρατήσουμε αδρές πληροφορίες για τις αριθμητικές τιμές των χαρακτηριστικών. Τότε, όλες οι αριθμητικές τιμές γνωρισμάτων αντικαθίστανται από τον αντίστοιχο μέσο όρο της κατηγορίας. [20], [21].

Μια άλλη συνηθισμένη τεχνική εξομάλυνσης δεδομένων προσπαθεί να εντοπίσει και πιθανώς να εξαλείψει μη τυπικά στιγμιότυπα από το σύνολο δεδομένων. Είναι πάντα χρήσιμο να αναγνωρίζουμε τα έκτοπα στιγμιότυπα που περιέχονται στα δεδομένα, αλλά η απομάκρυνσή τους από τη βάση δεδομένων θα μπορούσε να αποδειχθεί αντιπαραγωγική.

Ελλιπή δεδομένα

Τα ελλιπή στοιχεία αποτελούν ένα πρόβλημα που μπορεί να αντιμετωπιστεί με διάφορους τρόπους. Στις περισσότερες περιπτώσεις, οι τιμές που λείπουν υποδηλώνουν χαμένες πληροφορίες. Για παράδειγμα, μια απούσα τιμή για το χαρακτηριστικό *age* υποδηλώνει σίγουρα ένα στοιχείο δεδομένων το οποίο υπάρχει, αλλά είναι άγνωστο. Όμως μια απούσα τιμή για το χαρακτηριστικό *salary* μπορεί να εκληφθεί ως ένα στοιχείο δεδομένων που δεν καταχωρίστηκε, αλλά θα μπορούσε επίσης να αντιστοιχεί και σε έναν άνεργο. Μερικές τεχνικές εξόρυξης πληροφορίας είναι σε θέση να αντιμετωπίσουν απευθείας το πρόβλημα των τιμών που λείπουν. Όμως πολλοί κατηγοριοποιητές απαιτούν να έχουν τιμή όλα τα χαρακτηριστικά. Ακολουθούν κάποιες πιθανές επιλογές για την αντιμετώπιση δεδομένων που λείπουν *προτού* αυτά τροφοδοτηθούν σε έναν αλγόριθμο εξόρυξης πληροφορίας.

- **Απόρριψη εγγραφών από τις οποίες λείπουν κάποιες τιμές.** Αυτή η μέθοδος είναι η πλέον κατάλληλη όταν μόνο από ένα μικρό ποσοστό του συνολικού αριθμού των στιγμιοτύπων λείπουν κάποια δεδομένα, και μπορούμε να είμαστε σίγουροι ότι οι απούσες τιμές πραγματικά αντιπροσωπεύουν χαμένες πληροφορίες.
- **Για δεδομένα πραγματικών τιμών, αντικατάσταση των τιμών που λείπουν με τη μέση τιμή της κατηγορίας.** Στις περισσότερες περιπτώσεις, αυτή είναι μια λογική προσέγγιση για τα αριθμητικά χαρακτηριστικά. Επιλογές, όπως η αντικατάσταση αριθμητικών δεδομένων που δεν είναι στη διάθεσή μας με μηδενική τιμή ή κάποια αυθαίρετα μεγάλη ή μικρή τιμή, είναι γενικά κακές επιλογές.

- **Αντικατάσταση των απουσών τιμών χαρακτηριστικών με τις αντίστοιχες τιμές άλλων παρόμοιων στιγμιότυπων.** Αυτή η τεχνική είναι κατάλληλη τόσο για όλα τα αριθμητικά χαρακτηριστικά όσο και για τα χαρακτηριστικά κατηγοριών.

Μερικές τεχνικές εξόρυξης πληροφορίας επιτρέπουν να λείπουν κάποιες τιμές από τα στιγμιότυπα. Στη συνέχεια αναφέρονται τρεις τρόποι με τους οποίους οι τεχνικές εξόρυξης πληροφορίας αντιμετωπίζουν τα δεδομένα που δεν υπάρχουν κατά την εκμάθηση:

1. **Αγνοούν τις απούσες τιμές.** Πολλοί αλγόριθμοι εξόρυξης πληροφορίας, συμπεριλαμβανομένων των νευρωνικών δικτύων και των κατηγοριοποιητών Bayes , χρησιμοποιούν αυτή την προσέγγιση.
2. **Αντιμετωπίζουν τις τιμές που λείπουν ως συγκρίσεις ισότητας.** Αυτή η προσέγγιση είναι επικίνδυνη στα δεδομένα με πολύ θόρυβο αφού στιγμιότυπα που είναι ανόμοια μπορεί να εμφανίζονται ως όμοια.
3. **Αντιμετωπίζουν τις τιμές που λείπουν ως συγκρίσεις ανισότητας.** Αυτή είναι μια απαισιόδοξη προσέγγιση, αλλά ίσως είναι η πιο κατάλληλη. Δύο παρόμοια στιγμιότυπα από τα οποία λείπουν διάφορες τιμές θα αναφερθούν ως ανόμοια.

Τέλος, μια ελαφρώς διαφορετική προσέγγιση είναι η χρήση της καθοδηγούμενης εκμάθησης για τον προσδιορισμό πιθανών τιμών για τα δεδομένα που μας λείπουν. Όταν το χαρακτηριστικό που λείπει είναι χαρακτηριστικό κατηγορίας, τότε το ορίζουμε ως χαρακτηριστικό εξόδου. Τα στιγμιότυπα με γνωστές τιμές για το συγκεκριμένο χαρακτηριστικό χρησιμοποιούνται για την κατασκευή ενός μοντέλου κατηγοριοποίησης. Κατόπιν, το μοντέλο που δημιουργείται χρησιμοποιείται για την ταξινόμηση των στιγμιοτύπων από τα οποία λείπουν τιμές. Για αριθμητικά δεδομένα μπορούμε να χρησιμοποιήσουμε ένα εργαλείο εξόρυξης εκτίμησης, όπως ένα νευρωνικό δίκτυο, και να εφαρμόσουμε την ίδια στρατηγική (18), .

4.4.4 Βήμα 4: Μετασχηματισμός των δεδομένων

ο μετασχηματισμός των δεδομένων μπορεί να πάρει πολλές μορφές και είναι απαραίτητος για πολλούς λόγους. Στις παρακάτω ενότητες δίνουμε μια περιγραφή μερικών γνωστών μετασχηματισμών δεδομένων.

Κανονικοποίηση δεδομένων

Ένας συνηθισμένος μετασχηματισμός δεδομένων προβλέπει την αλλαγή των αριθμητικών τιμών έτσι ώστε να βρίσκονται μέσα σε ένα καθορισμένο διάστημα. Κατηγοριοποιητές όπως τα νευρωνικά δίκτυα αποδίδουν καλύτερα με αριθμητικά δεδομένα προσαρμοσμένα σε ένα διάστημα μεταξύ Ο και 1. Η μέθοδος της κανονικοποίησης ταιριάζει ιδιαίτερα σε κατηγοριοποιητές βασισμένους σε αποστάσεις, επειδή χαρακτηριστικά με μεγάλο εύρος τιμών είναι λιγότερο πιθανό να υπερτερούν χαρακτηριστικών με μικρότερο αρχικό εύρος. Τέσσερις συ-νηθισμένες μέθοδοι κανονικοποίησης είναι:

- **Δεκαδική κλιμάκωση.** Στη δεκαδική κλιμάκωση, κάθε αριθμητική τιμή διαιρείται με την ίδια δύναμη του δέκα. Για παράδειγμα, αν γνωρίζουμε ότι οι τιμές ενός χαρακτηριστικού κυμαίνονται μεταξύ -1000 και 1000, μπορούμε να αλλάξουμε το διάστημα σε -1 και 1, διαιρώντας κάθε τιμή με το 1000.
- **Κανονικοποίηση ελαχίστου-μεγίστου.** Η κανονικοποίηση ελαχίστου- μεγίστου (Min-Max) ενδείκνυται όταν είναι γνωστές η ελάχιστη και η μέγιστη τιμή ενός χαρακτηριστικού. Ο μαθηματικός τύπος είναι:

$$newValue = \frac{originalValue - oldMin\{newMax - newMin) + newMin}{oldMax - oldMin}$$

όπου 01 *oldMax* και *oldMin* αντιπροσωπεύουν τις αρχικές μέγιστες και ελά- χιστες τιμές του χαρακτηριστικού. Οι *NewMax* και *NewMin* καθορίζουν τις νέες μέγιστες και

ελάχιστες τιμές. Η *newValue* αντιπροσωπεύει το μετα- σχηματισμό της *originalValue.* Αυτός ο μετασχηματισμός είναι ιδιαίτερα χρήσιμος στα νευρωνικά δίκτυα, όπου το επιθυμητό διάστημα είναι το [0,1]. Στην περίπτωση αυτή, ο μαθηματικός τύπος ανάγεται στη μορφή:

$$newValue = \frac{originalValue - oldMin}{oldMax - oldMin}$$

- Κανονικοποίηση με χρήση των τιμών Z-score. Η κανονικοποίηση Z- score μετατρέπει μια τιμή σε ένα τυποποιημένο αποτέλεσμα, αφαιρώντας από αυτήν τη μέση τιμή (μ) του χαρακτηριστικού και διαιρώντας την με την τυπική απόκλιση (σ) του χαρακτηριστικού. Πιο συγκεκριμένα:

$$newValue = \frac{originalValue - \mu}{\sigma}$$

  Αυτή η τεχνική είναι ιδιαίτερα χρήσιμη όταν η μέγιστη και ελάχιστη τιμή δεν είναι γνωστές.

- ■ **Λογαριθμική κανονικοποίηση.** Ο λογάριθμος βάσης *b* ενός αριθμού *n* είναι ο εκθέτης στον οποίο θα πρέπει να υψωθεί ο *b* για να ισούται με τον αριθμό *n.* Για παράδειγμα, ο λογάριθμος βάσης 2 του 64 είναι 6 επειδή $2^6 = 64$. Η αντικατάσταση ενός συνόλου τιμών με τους λογάριθμούς τους έχει ως αποτέλεσμα την προσαρμογή της κλίμακας του διαστήματος τιμών χωρίς απώλεια πληροφορίας.

Μετατροπή τύπων δεδομένων

Πολλά εργαλεία εξόρυξης πληροφορίας, συμπεριλαμβανομένων των νευρωνικών δικτύων και μερικών στατιστικών μεθόδων (17), (18), δεν μπορούν να επεξεργαστούν δεδομένα κατηγοριών. Κατά συνέπεια, η μετατροπή δεδομένων κατηγοριών σε αριθμητικά ισοδύναμα είναι μια συνήθης μετατροπή δεδομένων. Από την άλλη, μερικές τεχνικές εξόρυξης πληροφορίας δεν είναι σε θέση να επεξεργαστούν αριθμητικά δεδομένα στην αρχική μορφή τους. Για παράδειγμα, οι περισσότεροι αλγόριθμοι δέντρων αποφάσεων διακριτοποιούν τις αριθμητικές τιμές ταξινομώντας τα δεδομένα και θεωρώντας εναλλακτικές δυαδικές διαιρέσεις των στοιχείων δεδομένων.

Επιλογή χαρακτηριστικών και στιγμιότυπων

Οι κατηγοριοποιητές ποικίλλουν ως προς τη δυνατότητά τους να χειριστούν με- γάλο όγκο δεδομένων. Μερικοί αλγόριθμοι εξόρυξης πληροφορίας έχουν πρόβλημα όταν υπάρχει μεγάλος αριθμός στιγμιοτύπων, ενώ άλλοι δεν μπορούν να αναλύσουν δεδομένα που περιέχουν περισσότερα από λίγα μόνο χαρακτηριστικά. Επίσης, πολλοί αλγόριθμοι εξόρυξης πληροφορίας δεν είναι σε θέση να δια- κρίνουν τα σχετικά χαρακτηριστικά από τα άσχετα. Αυτό είναι πρόβλημα, αφού έχει αποδειχθεί ότι ο αριθμός των στιγμιοτύπων εκπαίδευσης που απαιτείται για την κατασκευή ακριβών μοντέλων καθοδηγούμενης εκμάθησης επηρεάζεται άμεσα από τον αριθμό των άσχετων χαρακτηριστικών που περιέχονται στα δεδομένα. Για να ξεπεραστούν αυτά τα προβλήματα, θα πρέπει να αποφασίσουμε ποια χαρακτηριστικά και ποια στιγμιότυπα θα χρησιμοποιηθούν όταν κατασκευάζουμε τα δικά μας μοντέλα εξόρυξης πληροφορίας. Στη συνέχεια αναφέρεται ένας πιθανός αλγόριθμος , ο οποίος διευκολύνει την επιλογή χαρακτηριστικών:

1. Αν δίνονται *N* χαρακτηριστικά, δημιουργία συνόλου *S* όλων των πιθανών συνδυασμών χαρακτηριστικών.

2. Αφαίρεση του πρώτου συνδυασμού χαρακτηριστικών του συνόλου *S* και δημιουργία ενός μοντέλου εξόρυξης γνώσης *M* χρησιμοποιώ ντας αυτά τα χαρακτηριστικά.

3. Μέτρηση της ορθότητας του μοντέλου *M.*

4. Μέχρι να μείνει κενό το σύνολο *S.*

    i Αφαίρεση του επόμενο συνδυασμού χαρακτηριστικών από το *S* και κατασκευή ενός μοντέλου εξόρυξης γνώσης χρησιμοποιώ-ντας τον επόμενο συνδυασμό που περιέχει το *S.*
    ii Σύγκριση της ορθότητας του καινούριου μοντέλου με το αποθηκευμένο μοντέλο *M.* Ονομασία του καλύτερου από τα δύο μοντέλο ως μοντέλο Μ και αποθήκευση αυτού ως βέλτιστο.

5. Το μοντέλο Μ είναι το μοντέλο που επιλέγεται .

Αυτός ο αλγόριθμος δίνει σίγουρα ένα βέλτιστο αποτέλεσμα. Το πρόβλημα με τον αλγόριθμο είναι η πολυπλοκότητά του. Αν έχουμε ένα σύνολο *η* χαρακτηριστικών, ο συνολικός αριθμός των συνδυασμών τους είναι 2"-1. Η εργασία της δημιουργίας και του ελέγχου όλων των πιθανών μοντέλων για οποιοδήποτε σύνολο δεδομένων με περισσότερα από λίγα μόνο χαρακτηριστικά είναι αδύνατη. Ας εξετάσουμε μερικές τεχνικές που μπορούν να εφαρμοστούν.

Απαλοιφή χαρακτηριστικών

Γενικά, οι αλγόριθμοι εξόρυξης πληροφορίας δεν έχουν καλή απόδοση όταν τα δεδομένα περιέχουν πολλά χαρακτηριστικά τα οποία δεν μπορούν να προβλέψουν τη συμμετοχή στις κατηγορίες. Μερικοί στατιστικοί, αλλά και κάποιοι μη στατιστικοί κατηγοριοποιητές διαθέτουν τεχνικές επιλογής χαρακτηριστικών που συμπεριλαμβάνονται στη διαδικασία κατασκευής του μοντέλου. Οι κατηγοριοποιητές με ενσωματωμένη επιλογή χαρακτηριστικών έχουν μικρότερη πιθανότητα να πάσχουν από τα συμπτώματα των συνόλων δεδομένων που περιέχουν χαρακτηριστικά μικρής αξίας στην πρόβλεψη. Δυστυχώς, πολλοί αλγόριθμοι εξόρυξης γνώσης, συμπεριλαμβανομένων των νευρωνικών δικτύων και των κατηγοριοποιητών αλγορίθμων πλησιέστερου γείτονα, δίνουν τον ίδιο συντελεστή βάρους σε όλα τα χαρακτηριστικά κατά τη διάρκεια της φάσης κατασκευής του μοντέλου. Με αυτούς τους κατηγοριοποιητές, η επιλογή χαρακτηριστικών πρέπει να γίνει πριν αρχίσει η διαδικασία εξόρυξης πληροφορίας. Μπορούν να γίνουν πολλές ενέργειες για να καθοριστούν τα χαρακτηριστικά που θα πρέπει να απαλειφθούν:

1. Χαρακτηριστικά εισόδου που σχετίζονται στενά με άλλα χαρακτηριστικά εισόδου είναι περιττά. Τα περισσότερα εργαλεία εξόρυξης πληροφορίας παράγουν καλύτερα μοντέλα όταν ορίζεται ως τιμή εισόδου μόνο ένα χαρακτηριστικό μέσα από ένα σύνολο στενά σχετιζόμενων χαρακτηριστικών.
2. Σε δεδομένα κατηγοριών, υποψήφιο για απαλοιφή μπορεί να είναι οποιοδήποτε χαρακτηριστικό περιέχει την τιμή v, με τιμή προβλεψιμότητας πεδίου ορισμού μεγαλύτερη από μια επιλεγμένη κατώτερη τιμή. Αυτό ισχύει επειδή τα περισσότερα στιγμιότυπα του πεδίου ορισμού θα έχουν τη v, ως τιμή τους για το συγκεκριμένο χαρακτηριστικό. Καθώς αυξάνει η τιμή προβλεψιμότητας πεδίου ορισμού της μειώνεται η δυνατότητά της να αποτελεί διακριτικό των επιμέρους κατηγοριών.
3. Όταν η εκμάθηση είναι καθοδηγούμενη, η σημαντικότητα των αριθμητικών χαρακτηριστικών μπορεί να καθοριστεί με τη σύγκριση της μέσης τιμής και της τυπικής απόκλισης της κάθε κατηγορίας. Θυμηθείτε ότι η μετρική της σημαντικότητας αριθμητικών χαρακτηριστικών που χρησιμοποιείται από το ESX υπολογίζει τις τιμές σημαντικότητας των χαρακτηριστικών συγκρίνοντας τυποποιημένες διαφορές μεταξύ των μέσων τιμών των κατηγοριών.

Οι δύο πρώτες τεχνικές μπορούν να εφαρμοστούν είτε στην καθοδηγούμενη εκμάθηση είτε στη μη καθοδηγούμενη συσταδοποίηση. Δυστυχώς, στη μη καθοδηγούμενη συσταδοποίηση, η σημαντικότητα των αριθμητικών χαρακτηριστικών δεν μπορεί να υπολογιστεί επειδή δεν υπάρχουν προκαθορισμένες κατηγορίες. Όμως μπορούμε να δοκιμάσουμε υποσύνολα πιθανών επιλογών χαρακτηριστικών και να χρησιμοποιήσουμε το κατάλληλο μέτρο ποιότητας συστάδων για να μας βοηθήσει να καθορίσουμε ένα βέλτιστο σύνολο αριθμητικών χαρακτηριστικών.

Σε μια ενδιαφέρουσα προσέγγιση επιλογής χαρακτηριστικών χρησιμοποιείται η γενετική εκμάθηση. Η μέθοδος είναι ελκυστική επειδή, με την ενσωμάτωση μιας συνάρτησης αξιολόγησης, απαλείφουμε τα συνδυαστικά προβλήματα που παρατηρούνται όταν δοκιμάζουμε άκριτα όλα τα πιθανά ενδεχόμενα, αλλά ταυτόχρονα μπορούμε να επιτύχουμε ικανοποιητικά αποτελέσματα. Η χρησιμότητα αυτής της προσέγγισης φαίνεται ιδιαίτερα όταν υπάρχουν πολλά άσχετα χαρακτηριστικά στα δεδομένα. Μπορούμε να εξηγήσουμε καλύτερα τη μέθοδο με ένα παράδειγμα.

Ας θεωρήσουμε τη βάση δεδομένων προώθησης πωλήσεων μέσω πιστωτικών καρτών. Και πάλι υποθέτουμε ότι το χαρακτηριστικό εξόδου είναι το *life insurance promotion.* Στον πίνακα 5 υπάρχει ένας αρχικός πληθυσμός τριών στοιχείων. Κάθε στοιχείο μάς λέει ποια χαρακτηριστικά πρέπει να χρησιμοποιήσουμε κατά την κατασκευή του σχετικού μοντέλου εκμάθησης. Το “1” αναφέρεται σε χαρακτηριστικά εισόδου, ενώ με “0” προσδιορίζονται τα μη χρησιμοποιούμενα χαρακτηριστικά. Η τεχνική είναι η ακόλουθη:

1. Επιλέξτε κατάλληλα δεδομένα εκπαίδευσης και δεδομένα ελέγχου.
2. Χρησιμοποιήστε μια διαδικασία τυχαίας επιλογής για να αρχικοποιήσετε έναν πληθυσμό στοιχείων.
3. Κατασκευάστε ένα μοντέλο καθοδηγούμενης εκμάθησης για κάθε στοιχείο του πληθυσμού. Κάθε μοντέλο κατασκευάζεται με τα χαρα- κτηριστικά που καθορίζονται από το αντίστοιχο στοιχείο του πληθυ- σμού. Για παράδειγμα, το μοντέλο εκμάθησης για το πρώτο στοιχείο του πληθυσμού χρησιμοποιεί τα χαρακτηριστικά εισόδου: *income range, credit card insurance, sex,* και *age.*
4. Αξιολογήστε κάθε στοιχείο εφαρμόζοντας το αντίστοιχο μοντέλο στα δεδομένα ελέγχου. Μια πιθανή συνάρτηση αξιολόγησης είναι η ακρίβεια του μοντέλου στα δεδομένα ελέγχου.
5. Αν ικανοποιείται η συνθήκη τερματισμού, επιλέξτε ένα στοιχείο από τον πληθυσμό για να κατασκευάσετε ένα τελικό μοντέλο καθοδηγούμενης εκμάθησης από τα δεδομένα εκπαίδευσης.
6. Αν δεν ικανοποιείται η συνθήκη τερματισμού, εφαρμόστε γενετικούς τελεστές για να τροποποιήσετε ένα ή περισσότερα στοιχεία του πληθυσμού και επαναλάβετε τα Βήματα 3-5.

Αν και είναι εγγυημένο ότι αυτή η τεχνική συγκλίνει, η σύγκλιση δεν είναι απαραίτητα βέλτιστη. Για το λόγο αυτό, μπορεί να χρειαστούν πολλές εκτελέσεις του αλγορίθμου για να φτάσουμε στο επιθυμητό αποτέλεσμα.

Population Element	**Income Range**	**Magazine Promotion**	**Watch Promotion**	**Credit Card Insurance**	**Sex**	**Age**
1	1	0	0	1	1	1
2	0	0	0	1	0	1
3	0	0	0	0	1	1

Πίνακας 5. **Ένας αρχικός πληθυσμός για επιλογή χαρακτηριστικών με την μέθοδο της γενετικής εκμάθησης**

Δημιουργία χαρακτηριστικών

Μερικές φορές, χαρακτηριστικά με μικρή αξία στην πρόβλεψη μπορούν να συνδυαστούν με άλλα χαρακτηριστικά για να σχηματίσουν νέα με υψηλό βαθμό δυνατότητας πρόβλεψης. Για παράδειγμα, σκεφτείτε τη βάση δεδομένων που αποτελείται από δεδομένα μετοχών. Στα πιθανά χαρακτηριστικά συμπεριλαμβάνονται η τρέχουσα τιμή της κάθε μετοχής, η διακύμανση της τιμής της σε χρονικό διάστημα 12 μηνών, ο ρυθμός ανάπτυξης της εταιρείας, τα κέρδη ανά

τρίμηνο, η κεφαλαιοποίηση, ο τομέας της εταιρείας, και άλλα παρόμοια. Τα χαρακτηριστικά τιμής και κερδών παρουσιάζουν κάποια αξία για την πρόβλεψη σε σχέση με τον καθορισμό μιας μελλοντικής τιμής στόχου. Όμως έχει διαπιστωθεί ότι ο λόγος τιμής-κερδών (λόγος P/E) είναι πιο χρήσιμος. Ένα δεύτερο χαρακτηριστικό που δημιουργείται και είναι πιθανό να προβλέψει αποτελεσματικά τη μελλοντική τιμή μιας μετοχής είναι ο λόγος P/E της μετοχής διαιρεμένος με το ρυθμό ανάπτυξης της εταιρείας. Στη συνέχεια αναφέρουμε μερικούς μετασχηματισμούς που εφαρμόζονται συνήθως για τη δημιουργία καινούργιων χαρακτηριστικών;

- Δημιουργία ενός νέου χαρακτηριστικού του οποίου κάθε τιμή αντιπροσωπεύει το λόγο της τιμής ενός χαρακτηριστικού προς την τιμή ενός δεύτερου χαρακτηριστικού.
- Δημιουργία ενός νέου χαρακτηριστικού του οποίου οι τιμές είναι διαφορές μεταξύ των τιμών δύο άλλων υφισταμένων χαρακτηριστικών.
- Δημιουργία ενός νέου χαρακτηριστικού του οποίου οι τιμές υπολογίζονται ως ποσοστιαία αύξηση ή μείωση δύο άλλων υφισταμένων χαρακτηριστικών. Αν δίνονται δύο τιμές V_1 και V_2 με $V_1 < V_2$, η ποσοστιαία αύξηση της V_2 σε σχέση με τη V_1 υπολογίζεται ως το πηλίκο:

$$PercentIncrease\{V_2, V_1) = \frac{V_2 - V_1}{V_1}$$

Αν $V_1 < V_2$, , αφαιρούμε την τιμή V_2 από τη V_1 και διαιρούμε με τη V_1, το οποίο δίνει την ποσοστιαία μείωση της V_2 σε σχέση με τη V_1.

Νέα χαρακτηριστικά που αντιπροσωπεύουν διαφορές και ποσοστιαίες αυξήσεις ή μειώσεις είναι ιδιαίτερα χρήσιμα στην ανάλυση χρονοσειρών. Η ανάλυση χρονοσειρών (time-series analysis) μοντελοποιεί αλλαγές συμπεριφοράς μέσα σε ένα χρονικό διάστημα. Γι' αυτό, είναι σημαντικά τα χαρακτηριστικά που δημιουργούνται από τον υπολογισμό διαφορών μεταξύ ενός χρονικού διαστήματος και του επόμενου χρονικού διαστήματος.

Επιλογή στιγμιοτύπων

Συχνά τα δεδομένα που χρησιμοποιούνται για τη φάση εκπαίδευσης της καθοδηγούμενης εκμάθησης επιλέγονται τυχαία από ένα σύνολο στιγμιοτύπων. Το μόνο κριτήριο που επηρεάζει την τυχαία διαδικασία είναι ότι τα στιγμιότυπα επιλέγονται έτσι ώστε να είναι εγγυημένη η αντιπροσώπευση κάθε κατηγορίας εννοιών που πρέπει να συμπεριληφθεί στην εκμάθηση. Οι αλγόριθμοι δέντρων αποφάσεων προχωρούν ένα ακόμα βήμα επιλέγοντας αρχικά ένα τυχαίο υποσύνολο των επιλεγμένων στιγμιοτύπων εκπαίδευσης για να κατασκευάσουν έναν αρχικό κατηγοριοποιητή. Στη συνέχεια, ο κατηγοριοποιητής ελέγχεται στα υπόλοιπα στιγμιότυπα εκπαίδευσης. Τα στιγμιότυπα που κατηγοριοποιούνται λανθασμένα από το δέντρο αποφάσεων προστίθενται στο υποσύνολο των δεδομένων εκπαίδευσης. Η διαδικασία επαναλαμβάνεται μέχρι να εξαντληθεί το σύνολο στιγμιοτύπων εκπαίδευσης ή να έχει κατασκευαστεί ένας κατηγοριοποιητής που κατηγοριοποιεί σωστά όλα τα δεδομένα εκπαίδευσης.

Εξαίρεση σε αυτόν τον κανόνα αποτελούν οι **κατηγοριοποιητές που βασίζονται σε στιγμιότυπα** (instance-based classifiers). Αυτοί δεν δημιουργούν γενικευμένα μοντέλα κατηγοριοποίησης. Αντί γι' αυτό, αποθηκεύουν ένα υποσύνολο αντιπροσωπευτικών στιγμιοτύπων της κάθε κατηγορίας. Το εκάστοτε νέο στιγμιότυπο ταξινομείται με βάση τη σύγκριση των τιμών των χαρακτηριστικών του με τις τιμές των αποθηκευμένων στιγμιοτύπων. Το άγνωστο στιγμιότυπο *i* τοποθετείται σε εκείνη την κατηγορία της οποίας τα αντιπροσωπευτικά στιγμιότυπα μοιάζουν περισσότερο με το *i.* Είναι προφανές ότι τα στιγμιότυπα που επιλέγονται να αντιπροσωπεύουν την κάθε κατηγορία καθορίζουν την ακρίβεια πρόβλεψης του μοντέλου.

Οι τιμές τυπικότητας στιγμιοτύπων μπορούν να χρησιμοποιηθούν για την επιλογή ενός βέλτιστου συνόλου αντιπροσωπευτικών στιγμιοτύπων από κάθε κατηγορία. Έχουμε εκτελέσει πολλά πειράματα χρησιμοποιώντας την τυπικότητα στιγμιοτύπων για να μας βοηθήσει να επιλέξουμε δεδομένα εκπαίδευσης για την καθοδηγούμενη εκμάθηση. Τα πειράματά μας

δείχνουν ότι μπορεί να επιτευχθεί βέλτιστη ακρίβεια κατηγοριοποίησης ενός δείγματος ελέγχου με όλους τους τύπους των κατηγοριοποιητών σχηματίζοντας σύνολα στιγμιοτύπων εκ-παίδευσης που περιέχουν μια υπερ-σταθμισμένη επιλογή από πολύ τυπικά και λιγότερο τυπικά στιγμιότυπα εκπαίδευσης.

Η μη καθοδηγούμενη συσταδοποίηση μπορεί επίσης να ωφεληθεί από την επιλογή στιγμιοτύπων. Μια απλή τεχνική είναι ο καθορισμός μιας τιμής τυπικότητας για κάθε στιγμιότυπο του πεδίου ορισμού. Καθώς δεν υπάρχουν προκαθορισμένες κατηγορίες, κάθε τιμή τυπικότητας υπολογίζεται σε σχέση με όλα τα στιγμιότυπα του πεδίου ορισμού. Αν γίνει απαλοιφή των περισσότερο παράτυπων στιγμιοτύπων του πεδίου ορισμού, τα συστήματα μη καθοδηγούμενης εκ- μάθησης μπορούν να σχηματίσουν καλά ορισμένες συστάδες. Αφού σχηματιστούν αυτές οι ποιοτικές συστάδες, τα παράτυπα στιγμιότυπα μπορούν να τροφοδοτηθούν στο σύστημα συσταδοποίησης. Το μοντέλο συσταδοποίησης είτε θα δημιουργήσει καινούργιες συστάδες με τα στιγμιότυπα ή θα τα τοποθετήσει σε υπάρχουσες συστάδες.

4.4.5 **Βήμα 5: Εξόρυξη πληροφορίας**

Η πειραματική και επαναληπτική φύση της ανακάλυψης γνώσης είναι περισσότερο προφανής στα Βήματα 5 και 6 της διαδικασίας ανακάλυψης γνώσης. Στη συνέχεια φαίνεται ένα συνηθισμένο σενάριο κατασκευής ενός καθοδηγούμενου ή μη καθοδηγούμενου μοντέλου εκμάθησης:

1. Επιλογή δεδομένων εκπαίδευσης και δεδομένων ελέγχου από το σύνολο των διαθεσίμων στιγμιοτύπων.
2. Ορισμός ενός συνόλου χαρακτηριστικών εισόδου.
3. Στην καθοδηγούμενη εκμάθηση, επιλογή ενός ή περισσότερων χαρακτηριστικών εξόδου.
4. Επιλογή τιμών για τις παραμέτρους εκμάθησης.
5. Χρήση του εργαλείου εξόρυξης πληροφορίας για την κατασκευή ενός γενικευμένου μοντέλου των δεδομένων.

Εφόσον ολοκληρωθεί η εξόρυξη πληροφορίας, θα γίνει αξιολόγηση του μοντέλου (Βήμα 6). Αν δεν προκύψει αποδεκτό αποτέλεσμα, τα βήματα που μόλις περιγράψαμε μπορούν να επαναληφθούν πολλές φορές. Γι' αυτό, ο συνολικός αριθμός των πιθανών μοντέλων εκμάθησης που δημιουργείται από ένα σύνολο δεδομένων είναι απεριόριστος. Ευτυχώς, η φύση της πειραματικής διαδικασίας, σε συνδυασμό με το γεγονός ότι οι τεχνικές εξόρυξης πληροφορίας είναι σε θέση να δημιουργήσουν αποδεκτά μοντέλα με ατελή δεδομένα, αυξάνει την πιθανότητα επιτυχίας.

4.4.6 Βήμα 6: Ερμηνεία και αξιολόγηση

Με την ερμηνεία και την αξιολόγηση καθορίζεται αν ένα μοντέλο εκμάθησης είναι αποδεκτό και μπορεί να εφαρμοστεί σε προβλήματα έξω από το χώρο ενός πειραματικού περιβάλλοντος. Αν προκύψουν αποδεκτά αποτελέσματα, αυτή είναι η φάση όπου η γνώση που έχει αποκτηθεί μεταφράζεται σε όρους κατανοητούς για τους χρήστες.

Η ερμηνεία και η αξιολόγηση μπορούν να πάρουν πολλές μορφές, μερικές από τις οποίες περιλαμβάνουν:

- **Στατιστική ανάλυση.** Μια τέτοια ανάλυση είναι χρήσιμη για τον καθορισμό σημαντικών διαφορών μεταξύ της απόδοσης των διαφόρων μοντέλων εξόρυξης πληροφορίας που δημιουργήθηκαν με χρήση διακριτών συνόλων χαρακτηριστικών και στιγμιοτύπων.
- **Ευρετική ανάλυση.** Οι *ευρετικοί κανόνες* είναι γενικοί κανόνες οι οποίοι δίνουν αρκετά καλές λύσεις σε προβλήματα. Τα περισσότερα εργαλεία εξόρυξης γνώσης προσφέρουν αριθμητικά υπολογισμένες ευρετικές λύσεις για να μας βοηθήσουν να αποφασίσουμε κατά πόσο έχει κάποια αξία η γνώση που έχει ανακαλυφθεί. Δύο παραδείγματα περιλαμβάνουν τον ευρετικό κανόνα ομοιότητας κατηγοριών που

υπολογίζεται από το ESX και τον υπολογισμό του αθροίσματος των τετραγώνων των σφαλμάτων, ο οποίος σχετίζεται με τον αλγόριθμο K-Means.

- ■ **Πειραματική ανάλυση.** Οι τεχνικές νευρωνικών δικτύων και ο αλγόριθμος K-Means κατασκευάζουν ελαφρώς διαφορετικά μοντέλα κάθε φορά που εφαρμόζονται με το ίδιο σύνολο παραμέτρων και τα ίδια δεδομένα. Άλλες μέθοδοι κατασκευάζουν διακριτά μοντέλα με ελαφρές παραλλαγές στην επιλογή των δεδομένων ή τη ρύθμιση των παραμέτρων. Για το λόγο αυτό, ο πειραματισμός με διάφορες επιλογές χαρακτηριστικών ή στιγμιοτύπων, καθώς επίσης και οι εναλλακτικές ρυθμίσεις παραμέτρων, μπορούν να δώσουν αποτελέσματα που διαφέρουν σημαντικά.
- **Ανάλυση που διενεργείται από άνθρωπο.** Κατά την ανάλυση αποτελεσμάτων, ο ανθρώπινος παράγοντας έρχεται να μας υπενθυμίζει ότι εμείς έχουμε τον έλεγχο της πειραματικής διαδικασίας. Σε τελική ανάλυση, εμείς πρέπει να αποφασίσουμε αν η γνώση που αποκτήθηκε από μια διαδικασία εξόρυξης πληροφορίας μπορεί να εφαρμοστεί επιτυχώς σε καινούργια προ- βλήματα.

4.4.7 **Βήμα 7: Αξιοποίηση εξορυχθείσας γνώσης**

Ο απώτερος σκοπός της εξόρυξης πληροφορίας είναι η εφαρμογή της γνώσης εξορύχθηκε . Από την πετυχημένη εφαρμογή της διαδικασίας ανακάλυψης γνώσης μπορούν να προκόψουν πολλές πιθανές ενέργειες:

- Η δημιουργία μιας αναφοράς ή ενός τεχνικού άρθρου σχετικά με αυτά που ανακαλύφθηκαν.
- Η αλλαγή της θέσης των προϊόντων λιανικής ή η τοποθέτηση σε κοντινή θέση επιλεγμένων προϊόντων που είναι σε προσφορά.
- Η αποστολή διαφημιστικών πληροφοριών σε ένα στοχευμένο δείγμα ενός πληθυσμού πελατών.
- Η ενσωμάτωση ενός αναπτυγμένου μοντέλου εκμάθησης ως σύστημα ε-μπροσθοφυλακής (front-end system) σχεδιασμένο να εντοπίζει απάτες με πιστωτικές κάρτες.
- ■ Η χρηματοδότηση μιας καινούργιας επιστημονικής μελέτης με βάση τα στοιχεία που μάθαμε από μια διαδικασία ανακάλυψης γνώσης.

Οι δυνατότητες περιορίζονται μόνο από τη δυνατότητα να συλλογής, να προεπεξεργασίας και να ανάλυσης δεδομένων με αποτελεσματικό τρόπο.

Κεφάλαιο 4ο

5 Αποθήκες Δεδομένων και OLAP

5.1 Ανάγκη Ανάπτυξης Αποθηκών Δεδομένων

Τα Συστήματα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) αναπτύχθηκαν στη δεκαετία του 1970 και άκμασαν τη δεκαετία του 1980. Επικρατέστερος τύπος είναι τα Σχεσιακά ΣΔΒΔ, τα οποία έχουν ως βασικό στόχο την επεξεργασία συναλλαγών. Ο όρος On-Line Transaction Processing (OLTP) αποδίδει το βασικό τρόπο λειτουργίας των Σχεσιακών ΣΔΒΔ. Παραδείγματα διαδικασιών OLTP είναι η καταγραφή αεροπορικών κρατήσεων μέσω του Παγκόσμιου Ιστού, πωλήσεων αγαθών στα ταμεία με σάρωση ραβδοκωδικών, τραπεζικών αναλήψεων/καταθέσεων στα ATM, πληροφοριών τηλεφωνημάτων (τηλεφωνικοί αριθμοί, διάρκεια) από τηλεπικοινωνιακές εταιρείες, κ.λπ. Στόχοι των διαδικασιών OLTP είναι η αξιοπιστία των συναλλαγών (δεν χάνονται δεδομένα) και η αποτελεσματικότητα (χειρισμός χιλιάδων συναλλαγών ανά δευτερόλεπτο). Συνοπτικά, λοιπόν, τα Σχεσιακά ΣΔΒΔ διαχειρίζονται τις καθημερινές λειτουργίες ενός οργανισμού με σκοπό τη διεκπεραίωση συναλλαγών. Για αυτό το λόγο, στο πλαίσιο ενός οργανισμού, τα Σχεσιακά ΣΔΒΔ ονομάζονται και *Επιχειρησιακές Βάσεις Δεδομένων* (Operational Databases) (10) (11) (12), (19) (20), (21).

Η μεγάλη αποδοχή που γνώρισαν τα Σχεσιακά ΣΔΒΔ οδήγησε στη δημιουργία μεγάλων Επιχειρησιακών Βάσεων Δεδομένων, οι οποίες, όμως, κατέληξαν να περιέχουν αναξιοποίητους τεράστιους όγκους δεδομένων. Για αυτό τον λόγο οι Επιχειρησιακές Βάσεις Δεδομένων παρομοιαζόταν ως “τάφοι δεδομένων” . Στα μέσα της δεκαετίας του 1990 έγινε αντιληπτό ότι το περιεχόμενο των Επιχειρησιακών Βάσεων Δεδομένων αποτελεί πολύτιμο υλικό για ανάλυση, με σκοπό την *υποστήριξη αποφάσεων* (decision support). Για παράδειγμα, σε μία επιχείρηση η ανάλυση των πωλήσεων της προηγούμενης διετίας μπορεί να αναδείξει ποια προϊόντα είναι επιτυχημένα και ποια όχι, ή ακόμη, ποιες αγοραστικές ομάδες ανταποκρίθηκαν στα νέα προϊόντα, ώστε να ληφθούν αποφάσεις για ανάπτυξη νέων στρατηγικών προώθησης πωλήσεων με διαφημιστικά και εκπτώσεις. Αμέσως γίνεται αντιληπτό το πλεονέκτημα που προσφέρει η αξιοποίηση των Επιχειρησιακών Βάσεων Δεδομένων, η οποία έχει ως αποτέλεσμα την ανάπτυξη της λεγόμενης *Επιχειρησιακής Νοημοσύνης* (Business Intelligence).

Όμως δυστυχώς η τεχνολογία των Σχεσιακών ΣΔΒΔ και τα εργαλεία που αυτά παρέχουν (π.χ., γλώσσα SQL, επεξεργασία και βελτιστοποίηση ερωτημάτων), δεν αφορούν τους σκοπούς της ανάλυσης των περιεχομένων των Επιχειρησιακές Βάσεις Δεδομένων. Αυτό συμβαίνει για διάφορους λόγους που συνοψίζονται ως εξής:

- Τις περισσότερες φορές τα δεδομένα των Επιχειρησιακών Βάσεων δεν έχουν καλή ποιότητα, δηλαδή υπάρχουν ελλιπή στοιχεία, θόρυβος και ασυνέπειες. Επομένως, τα δεδομένα δεν προσφέρονται προς άμεση αξιοποίηση, αφού απαιτούν καθαρισμό.

- Στα πλαίσια ενός οργανισμού, οι Επιχειρησιακές Βάσεις Δεδομένων είναι συνήθως ανεξάρτητες. Για παράδειγμα, σε μία επιχείρηση υπάρχουν τμήματα με ανεξάρτητες Επιχειρησιακές Βάσεις, όπως τα τμήματα πωλήσεων, μάρκετινγκ, προσωπικού και προμηθειών. Επειδή αυτές οι Επιχειρησιακές Βάσεις δεν έχουν σχεδιασθεί για να συνεργάζονται, δεν αλληλοενημερώνονται, με αποτέλεσμα τα δεδομένα τους να είναι ετερογενή και να προ- κύπτουν συνωνυμίες ή αμφισημίες. Για παράδειγμα, ένας πελάτης μπορεί να είναι καταχωρισμένος με διαφορετικά στοιχεία στη βάση του τμήματος πωλήσεων από ότι στη βάση του τμήματος μάρκετινγκ. Επομένως, πριν αναλυθούν τα δεδομένα, πρέπει να ολοκληρωθούν και να ομογενοποιηθούν. Επίσης, από την πληθώρα των δεδομένων που υπάρχουν στις

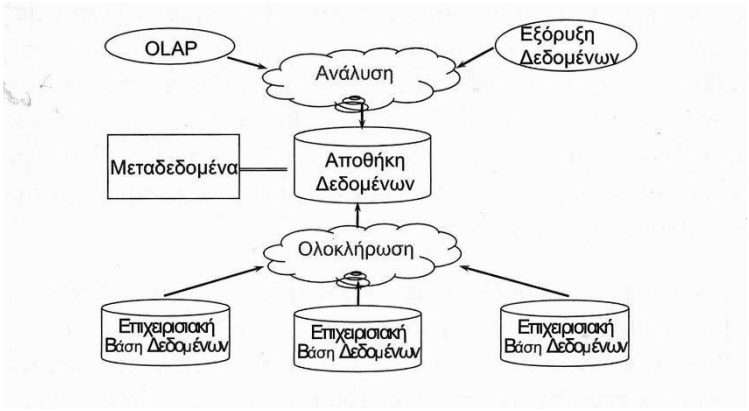
Επιχειρησιακές Βάσεις, πρέπει να επιλεγούν μόνο τα δεδομένα που είναι χρήσιμα για τους σκοπούς της ανάλυσης.

• Μέσω διαδικασιών ενημέρωσης (εισαγωγές/διαγραφές), οι Επιχειρησιακές Βάσεις διατηρούν δεδομένα μόνο για την τρέχουσα κατάσταση. Στο προηγούμενο παράδειγμα της επιχείρησης, στη βάση του τμήματος προμηθειών διατηρούνται μόνο όσοι προμηθευτές συνεργάζονται μαζί της αυτή τη στιγμή. Όμως, για τους σκοπούς της ανάλυσης μπορεί να χρειασθούν δεδομένα και για προμηθευτές που συνεργαζόταν στο παρελθόν, ώστε, π.χ., να συγκριθούν οι τιμές των προμηθευόμενων προϊόντων και να φανεί αν η αύξηση τιμών οδήγησε σε μείωση πωλήσεων. Αυτό που χρειάζεται είναι να διατηρούνται ιστορικά δεδομένα, όχι δηλαδή μόνο αυτά που αποτυπώνουν την τρέχουσα κατάσταση.

• Η ανάλυση των δεδομένων δεν είναι εύκολο να επιτευχθεί με εργαλεία όπως η γλώσσα SQL. Ο λόγος είναι ότι προκύπτουν περίπλοκα ερωτήματα που δεν είναι εύκολο να συνταχθούν. Επιπλέον, τα Σχεσιακά ΣΔΒΔ στο φυσικό επίπεδο δεν είναι σχεδιασμένα για να ανταποκρίνονται στις απαιτήσεις τέτοιων περίπλοκων ερωτημάτων. Συνεπώς, χρειάζονται νέα εργαλεία με τα οποία η ανάλυση των δεδομένων θα γίνεται εύκολα για το χρήστη και αποδοτικά από πλευράς χρόνου εκτέλεσης.

• Τέλος, στα Σχεσιακά ΣΔΒΔ τα δεδομένα οργανώνονται με βάση μεθοδολογίες, όπως το Διάγραμμα Οντοτήτων-Συσχετίσεων (ΔΟΣ) και η κανονικοποίηση. Το ΔΟΣ και η κανονικοποίηση εκπληρώνουν τις απαιτήσεις της αποδοτικής OLTP, αλλά παράγουν βάσεις δεδομένων που είναι περί- πλοκές στο εννοιολογικό επίπεδο (conceptual level). Για τους σκοπούς της ανάλυσης των δεδομένων, απαιτούνται διαφορετικές τεχνικές σχεδιασμού και οργάνωσης των δεδομένων στο εννοιολογικό επίπεδο.

• Η ανάγκη για την ικανοποίηση των προαναφερθέντων απαιτήσεων οδήγησε στην ανάπτυξη των *αποθηκών δεδομένων* (data warehouses). Με τρόπους που θα αναλυθούν στη συνέχεια, η τεχνολογία των αποθηκών δεδομένων προσφέρει (α) ολοκλήρωση ετερογενών πηγών δεδομένων, και (β) μια πλατφόρμα για αποδοτική ανάλυση ιστορικών δεδομένων. Επομένως, μία αποθήκη δεδομένων αποτελεί μία ολοκληρωμένη συλλογή δεδομένων που επιλέγονται από τις Επιχειρησιακές Βάσεις, ενώ στη συνέχεια η συλλογή αυτή αναλύεται με διαδικασίες όπως η On-line Analytical Processing (OLAP) ή η εξόρυξη δεδομένων. Λόγω των διαφορετικών απαιτήσεων, η αποθήκη δεδομένων διατηρείται σε ξεχωριστό υπολογιστικό σύστημα από τις επιχειρησιακές βάσεις, από τις οποίες αντιγράφονται και ολοκληρώνονται τα δεδομένα. Συνεπώς, αναλόγως της συχνότητας ενημέρωσης της αποθήκης ως προς τις μεταβολές των δεδομένων των επιχειρησιακών βάσεων, η αποθήκη μπορεί να μην περιέχει τα πλέον ενημερωμένα δεδομένα. Όμως, στις περισσότερες εφαρμογές αυτό δεν αποτελεί πρόβλημα, καθώς η ανάλυση των δεδομένων δεν επηρεάζεται από λίγες πρόσφατες μεταβολές. Οι προαναφερθείσες έννοιες συνοψίζονται στην εικόνα 2.



Εικόνα 2. Βασικές λειτουργίες Αποθήκης Δεδομένων

• Η τεχνολογία των αποθηκών δεδομένων προσέλκυσε γρήγορα το επιχειρηματικό ενδιαφέρον λόγω των πλεονεκτημάτων που προσφέρει. Οι πρώτες προσπάθειες ανάπτυξης αποθηκών δεδομένων ξεκίνησαν εντατικά στα μέσα της δεκαετίας του 1990, οπότε οι αποθήκες δεδομένων εξελίχθηκαν σε αγορά με οικονομικό κύκλο της τάξης των 2 δις δολαρίων. Ωστόσο, οι πρώτες αυτές προσπάθειες είχαν ποσοστό επιτυχίας μόλις 20%, επειδή αρχικά δεν αναγνωρίσθηκαν οι προαναφερθέντες παράγοντες. Όμως, πολύ σύντομα η κατάσταση άλλαξε. Στα τέλη της δεκαετίας του 1990, το 95% των 1000 επιχειρήσεων του Fortune ανέπτυσσαν αποθήκες δεδομένων, οπότε η αγορά των αποθηκών δεδομένων ανήρθε οικονομικά στο ύψος των 7 δις δολαρίων. Εκτιμήθηκε ότι σε 3 χρόνια από την ανάπτυξη μίας αποθήκης δεδομένων, επιτυγχάνεται απόσβεση σε ποσοστό 400%. Αυτό επιβεβαιώνεται από επιτυχημένες περιπτώσεις εφαρμογής, όπως στην αλυσίδα υπεραγορών Walmart με 2000 υποκαταστήματα. Η ανάπτυξη αποθήκης δεδομένων βοήθησε τη Walmart να βελτιστοποιήσει τις διαδικασίες προμήθειας προϊόντων και να μειώσει το κόστος αγοράς τους κατά 20%. Για να γίνει κατανοητή η κλίμακα μίας τέτοιας εφαρμογής, αναφέρεται ενδεικτικά ότι ο όγκος των δεδομένων στην αποθήκη δεδομένων της Walmart ανέρχεται στα 24 TB, σε ένα σύστημα 96 κόμβων με 900 επεξεργαστές και 2700 δίσκους.

5.2 Ορισμός Αποθήκης Δεδομένων

Ως *αποθήκη δεδομένων* ορίζεται μία προσανατολισμένη προς το θέμα, ολοκληρωμένη, χρονικά μεταβαλλόμενη και μη πτητική συλλογή δεδομένων με σκοπό την υποστήριξη λήψης αποφάσεων.

Ο όρος *προσανατολισμένη προς το θέμα* (subject-oriented) σημαίνει ότι η αποθήκη δεδομένων οργανώνεται γύρω από βασικές επιχειρηματικές έννοιες. Για παράδειγμα, σε μία επιχείρηση βασικές έννοιες είναι οι πωλήσεις, τα προϊόντα, οι πελάτες, κ.λπ. Αυτές οι έννοιες ορίζονται από όσους εμπλέκονται στην ανάλυση των δεδομένων, και μπορεί να μην διακρίνονται εντός των διαδικασιών OLTP στις επιχειρησιακές βάσεις δεδομένων. Για παράδειγμα, τα δεδομένα ενός πελάτη μπορεί να βρίσκονται διεσπαρμένα σε πολλούς πίνακες και σε διαφορετικές βάσεις μεταξύ των τμημάτων. Όμως, για τους σκοπούς της ανάλυσης δεδομένων, επιθυμούμε τη συνολική ανάδειξη της έννοιας του πελάτη, ασχέτως πως αυτή αναπαρίσταται στις επιχειρησιακές βάσεις κατά διεσπαρμένο τρόπο. Επιπλέον, οι έννοιες διατηρούνται απλές και κατανοητές επιλέγοντας μόνο τα δεδομένα που σχετίζονται με τη λήψη αποφάσεων.

Ο όρος *ολοκληρωμένη* (integrated) σημαίνει ότι τα δεδομένα που αφορούν στο ίδιο θέμα ορίζονται με ίδιο τρόπο. Αυτό γίνεται μέσω της ολοκλήρωσης των ετερογενών επιχειρησιακών βάσεων δεδομένων. Η ολοκλήρωση εφαρμόζεται μέσω μετασχηματισμών κατά τη φόρτωση των δεδομένων, όπου φροντίζουμε να υπάρχει συνέπεια στην ονομασία ιδιοτήτων, τιμών, και μετρήσεων. Για παράδειγμα, έστω ότι η ημερομηνίες στο τμήμα προμηθειών αναπαρίστανται ως Ημέρα/Μήνας/ΕΕΕΕ, ενώ στο τμήμα αγορών ως ΗΗ/ΜΜ/ΕΕ. Μετά την ολοκλήρωση, επιλέγουμε τη μορφή Ημέρα/Μήνας/ΕΕΕΕ. Επίσης, έστω ότι στο τμήμα πωλήσεων μετρούμε το βάρος των προϊόντων σε γραμμάρια, ενώ στο τμήμα μάρκετινγκ σε κιλά. Μετά την ολοκλήρωση, επιλέγουμε να μετρούμε το βάρος μόνο σε κιλά.

Ο όρος ***χρονικά μεταβαλλόμενη*** (time-variant) σημαίνει ότι με το πέρασμα του χρόνου αλλάζουν τα δεδομένα που σχετίζονται με ένα θέμα, οπότε διατηρούμε ιστορικά δεδομένα σε βάθος χρόνου ακόμα και ετών. Αυτό σημαίνει ότι κάθε δεδομένο συνοδεύεται από στοιχεία που αφορούν στο χρόνο. Για παράδειγμα, αν ένας πελάτης αλλάξει διεύθυνση, διατηρούμε την παλαιά διεύθυνση και το χρονικό διάστημα για το οποίο ίσχυε.

Ο όρος ***μη πτητική*** (non volatile) σημαίνει ότι δεν συμβαίνουν διαγραφές στην αποθήκη δεδομένων. Όπως αναφέρθηκε, όταν συμβαίνουν μεταβολές, οι παλαιές τιμές διατηρούνται ως ιστορικά δεδομένα.

Για την καλύτερη κατανόηση της έννοιας των αποθηκών δεδομένων, στον Πίνακα 2 παρουσιάζονται οι ειδοποιοί διαφορές σε σχέση με τις επιχειρησιακές βάσεις δεδομένων (Σχεσιακό ΣΔΒΔ). Πάντως συνοπτικά μπορούμε να ισχυρισθούμε ότι οι επιχειρησιακές βάσεις δεδομένων “τρέχουν” μία επιχείρηση, ενώ η αποθήκη δεδομένων τη βελτιστοποιεί [10].

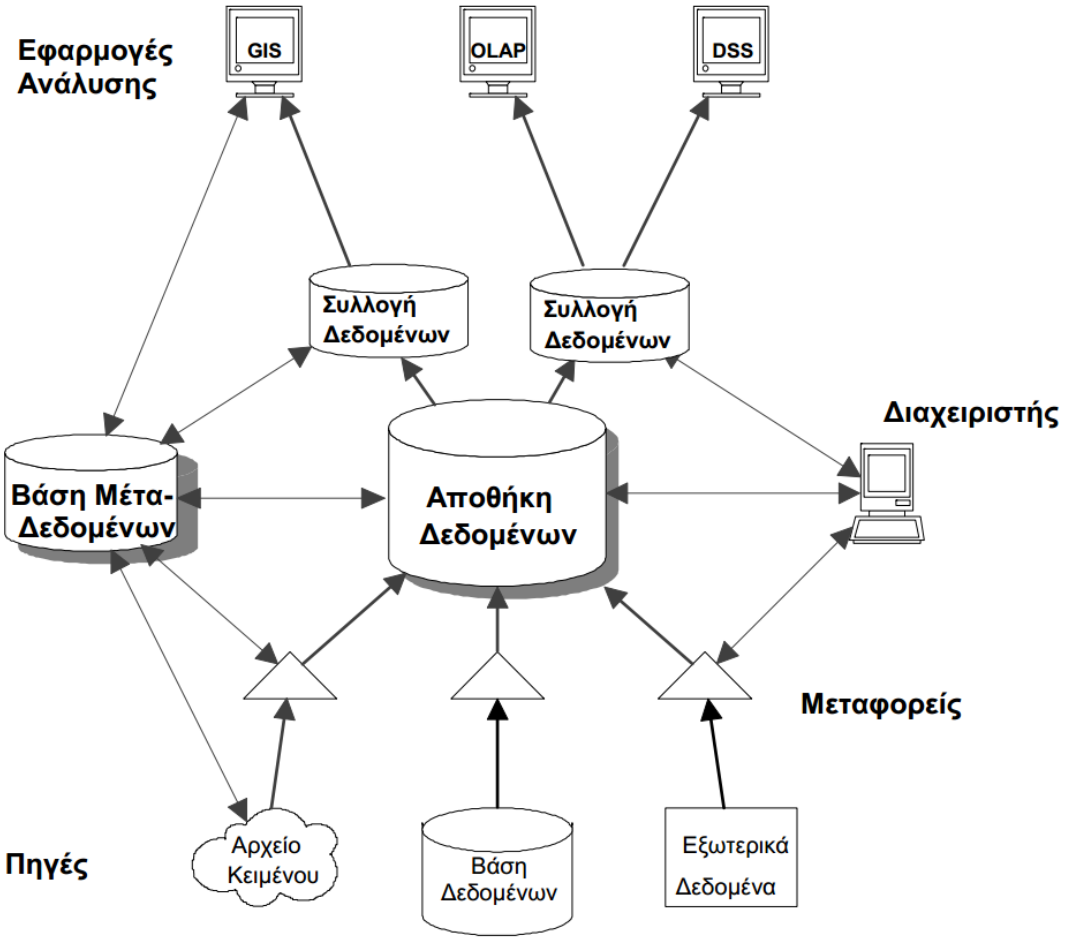
Χαρακτηριστικό	Σχεσιακό ΣΔΒΔ	Αποθήκη Δεδομένων
Σκοπός	“τρέξιμο” καθημερινών διεργασιών	υποστήριξη αποφάσεων
Λειτουργία	διεκπεραίωση συναλλαγών	εξαγωγή πληροφορίας
Χρήστες	κατώτεροι εργαζόμενοι (π.χ., ταμίες), DBAs	υψηλόβαθμα στελέχη, αναλυτές γνώσης
Αρι,θμός χρηστών	(μέχρι) χιλιάδες	(μέχρι) εκατοντάδες
Δεδομένα	τρέχοντα, απομονωμένα	ιστορικά, ολοκληρωμένα
Μέγεθος	100 MB-100 GB	100 GB-10 TB
Σχεδιασμός	διάγραμμα 0-Σ, κανονικοποίηση	μοντελοποίηση διαστάσεων, απο-κανονικοποίηση
Χρήση	επαναληπτική	ad-hoc
Προσπέλαση	ανάγνωση/εγγραφή	(κυρίως) ανάγνωση
Ενημέρωση	συνεχής	περιοδική
Μονάδα εργασίας	σύντομες, απλές συναλλαγές	περίπλοκα ερωτήματα
Χρόνοι διεκπεραίωσης	millisec-sec	λεπτά-ώρες

Αριθμός προσπελασμένων εγγραφών	δεκάδες	εκατομμύρια
Μονάδα απόδοσης	συναλλαγές ανά δευτερόλεπτο	χρόνος απόκρισης ερωτημάτων
ACID	ναι	όχι
Κατάλογοι	B-δένδρα	κατάλογοι bitmap

Πίνακας 2. Διαφορές μεταξύ Σχεσιακών ΣΔΒΔ και αποθηκών δεδομένων

5.3 Αρχιτεκτονική Αποθήκης Δεδομένων

Η επιλογή της αρχιτεκτονική μίας Αποθήκης Δεδομένων πρέπει να ικανοποιεί τις συγκεκριμένες ανάγκες του οργανισμού για τις οποίες δημιουργήθηκε και να εξασφαλίζει τη διαθεσιμότητα και την αποδοτικότητα του συστήματος. Η εικόνα 4 παρουσιάζει μία γενική αρχιτεκτονική ενός συστήματος Αποθήκης Δεδομένων. Στην εικόνα σημειώνονται τα βασικά δομικά στοιχεία μίας Αποθήκης Δεδομένων, η διασύνδεση των στοιχείων τους, καθώς και η ροή των δεδομένων (22), (19), (23), (21), .



Εικόνα 5. : Γενική Αρχιτεκτονική Αποθήκης Δεδομένων

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

- **Πηγές**: Κάθε πηγή από την οποία η Αποθήκη Δεδομένων αντλεί δεδομένα.
- **Μεταφορείς - Μετατροπείς**: Εφαρμογές που εκτελούν τις διαδικασίες μεταφοράς των δεδομένων από τις πηγές στην Αποθήκη Δεδομένων.
- **Αποθήκη Δεδομένων, Συλλογές Δεδομένων**: Τα συστήματα που αποθηκεύονται τα δεδομένα που παρέχονται προς τους χρήστες.
- **Βάση Μετα-Δεδομένων**: Σύστημα αποθήκευσης πληροφορίας σχετικά με τη δομή και λειτουργία του συστήματος.
- **Διαχειριστής**: Εφαρμογή που παρέχει δυνατότητα διαχείρισης του συστήματος
- **Εφαρμογές Ανάλυσης**: Εφαρμογές που έχουν πρόσβαση στην Αποθήκη Δεδομένων. Συνήθως είναι συστήματα στήριξης αποφάσεων.

5.3.1 Πηγές και Μεταφορείς - Μετατροπείς

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

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

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

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

5.3.2 Αποθήκη Δεδομένων, Συλλογές Δεδομένων

Οι Αποθήκες Δεδομένων και οι Συλλογές Δεδομένων, όπως φαίνεται στο εικόνα 4, υλοποιούνται με τη χρήση Σχεσιακών Συστημάτων Διαχείρισης Βάσεων Δεδομένων. Τα δεδομένα αποθηκεύονται σε σχεσιακές βάσεις δεδομένων, ενώ πρόσβαση σε αυτά παρέχεται

από μία γλώσσα διαχείρισης δεδομένων που είναι επέκταση της SQL. Εναλλακτική της χρήσης σχεσιακών συστημάτων είναι η χρήση των *Πολυδιάστατων Συστημάτων Αναλυτικής Επεξεργασίας (Multidimensional OLAP servers),* που αποθηκεύουν και διαχειρίζονται δεδομένα με πολυδιάστατο τρόπο (βλ. και ενότητα 8.6). Η χρήση σχεσιακών ΣΔΒΔ εκμεταλλεύεται την ευελιξία και την ισχύ της τεχνολογίας των σύγχρονων συστημάτων. Κατά την αναλυτική επεξεργασία δεδομένων εκτελούνται πολύπλοκες ερωτήσεις που απαιτούν δυνατότητα διαχείρισης μεγάλου όγκου πληροφοριών. Τα πλεονεκτήματα των πολυδιάστατων συστημάτων βρίσκονται στη δυνατότητά τους να διαχειρίζονται δεδομένα, τα οποία είναι δομημένα με τρόπο που βρίσκεται πιο κοντά στις ανάγκες των εφαρμογών ανάλυσης (OLAP).

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

5.3.3 Βάση Μετα-Δεδομένων

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

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

Όπως φαίνεται και στη εικόνα 4, τα μετα-δεδομένα αποθηκεύονται σε ένα σύστημα, όπου υπάρχει πρόσβαση από κάθε δομικό στοιχείο της αρχιτεκτονικής. Το γεγονός αυτό δημιουργεί την ανάγκη ύπαρξης ενός σταθερού προτύπου για τα μετα-δεδομένα, καθώς, όπως προαναφέρθηκε, τα διάφορα δομικά στοιχεία που συμμετέχουν στην αρχιτεκτονική των Αποθηκών Δεδομένων είναι εφαρμογές ανεπτυγμένες ανεξάρτητα από τις πηγές και τις εφαρμογές ανάλυσης. Ένα τέτοιο πρότυπο έχει προταθεί από μια ομάδα εταιρειών του χώρου και ονομάζεται *Metadata Interchange Specification (MDIS).*

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

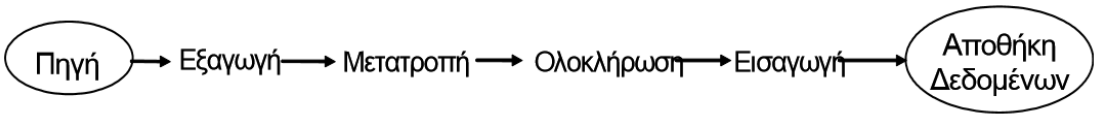
5.3.4 Σχεδίαση αρχιτεκτονικής Αποθηκών Δεδομένων

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

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

5.4 Μεταφορά Δεδομένων από τις πηγές στην Αποθηκών Δεδομένων

Βασικός παράγοντας για την επιτυχία των Αποθηκών Δεδομένων είναι η ορθή τροφοδοσία της Αποθήκης Δεδομένων από τις πηγές. Η διαδικασία μεταφοράς δεδομένων από τις πηγές στην Αποθήκη δεδομένων είναι αρκετά πολύπλοκη καθώς πολλά προβλήματα πρέπει να αντιμετωπισθούν. Τα βήματα που ακολουθούνται κατά τη μεταφορά των δεδομένων παρουσιάζονται στο εικόνα 5 (24) (25), (23).



Εικόνα 6. Διαδικασία μεταφοράς δεδομένων

5.4.1 Εξαγωγή και Μετατροπή Δεδομένων

Η Εξαγωγή και η Μετατροπή δεδομένων εκτελούνται από τους Μεταφορείς / Μετατροπείς του εικόνας 4. Για κάθε πηγή που χρησιμοποιούμε στο σύστημα εγκαθιστούμε λογισμικό που αντλεί τα δεδομένα από την πηγή, τα "καθαρίζει", κρατώντας μόνο αυτά που είναι πραγματικά χρήσιμα και τα μετασχηματίζει με βάση ένα καθορισμένο πρότυπο. Οι μετατροπές που γίνονται στα δεδομένα αφορούν τόσο τη δομή όσο και την τιμή τους. Για παράδειγμα, το πεδίο "Ημερομηνία" ενός πίνακα μπορεί να μετασχηματιστεί στα πεδία "Χρόνος", "Μήνας" και "Ημέρα", ενώ οι τιμές του πεδίου "Χαρακτηρισμός" είναι πιθανόν να μετατραπούν από "Α", "Β" κλπ σε "1", "2" κλπ αντίστοιχα. Αυτό το λογισμικό υλοποιείται με βάση τα ιδιαίτερα χαρακτηριστικά κάθε πηγής και εγκαθίσταται σε υπολογιστές με άμεση πρόσβαση στα δεδομένα της πηγής. Οι Αποθήκες Δεδομένων χρησιμοποιούν ποικίλα εργαλεία για εξαγωγή. Η εξαγωγή δεδομένων από τις απομακρυσμένες πηγές συχνά υλοποιείται μέσω πυλών (gateways) και καθιερωμένων προτύπων διασύνδεσης εφαρμογών (όπως ODBC, Oracle Open Connect, Information Builders EDA/SQL κλπ).

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

5.4.2 Ολοκλήρωση

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

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

5.4.3 Εισαγωγή δεδομένων

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

5.4.4 Ενημέρωση

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

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

Δεδομένου ότι οι Αποθήκες Δεδομένων συσσωρεύουν μεγάλη ποσότητα δεδομένων, οι εφαρμογή της παραπάνω μεθόδου ενημέρωσης καθίσταται απαγορευτική. Γιαυτό και είναι αναγκαία η διάγνωση των μεταβολών που συμβαίνουν στις πηγές (εισαγωγές, διαγραφές και

τροποποιήσεις εγγραφών), ώστε σε κάθε διαδικασία ενημέρωσης να μην γίνονται περιττές διαγραφές και εισαγωγές δεδομένων που στην πραγματικότητα παραμένουν αναλλοίωτα. Σύμφωνα με τις τεχνικές προοδευτικής ενημέρωσης και συντήρησης των δεδομένων, εισάγονται στην Αποθήκη νέες εγγραφές που προκύπτουν αποκλειστικά από αντίστοιχες εισαγωγές δεδομένων στις πηγές. Ομοίως, και οι διαγραφές και τροποποιήσεις εγγραφών προκύπτουν από αντίστοιχες πράξεις δεδομένων στις πηγές.

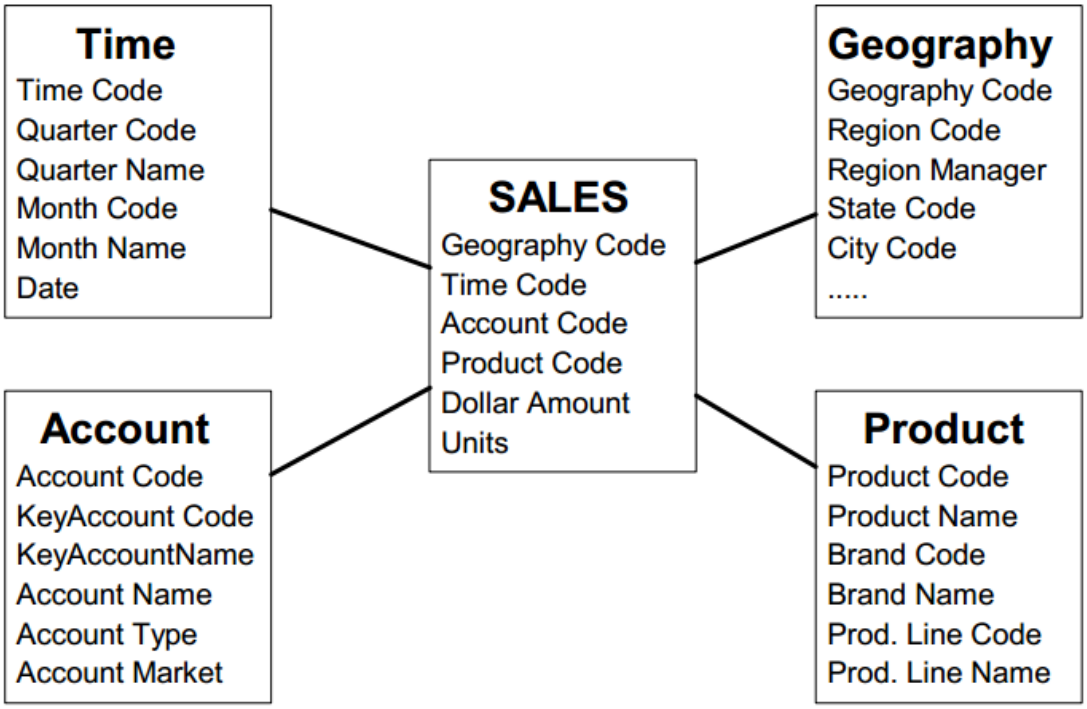
Για να γίνει εφικτή η εφαρμογή της προοδευτικής ενημέρωσης πρέπει οι πηγές να μας δίνουν τη δυνατότητα διάγνωσης των μεταβολών που συντελούνται στα δεδομένα τους. Στις περιπτώσεις που η πηγή είναι ένα σύγχρονο σύστημα βάσης δεδομένων, υπάρχουν τρεις βασικές τεχνικές με τις οποίες είναι εφικτή η διάγνωση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

- **Στιγμιότυπα**: Αρκετά συστήματα βάσεων δεδομένων είναι σε θέση να εξάγουν, όταν τους ζητηθεί, στιγμιότυπα (snapshots) από πίνακες της βάσης τους. Από τα στιγμιότυπα αυτά με διάφορες μεθόδους αποδοτικής σύγκρισης μπορούμε να διαγνώσουμε τις τροποποιήσεις που συνέβηκαν στην πηγή και να ενημερωθεί σχετικά η Αποθήκη.
- **Μηχανισμός καταγραφής (log)**: Τα περισσότερα σύγχρονα συστήματα βάσεων δεδομένων καταγράφουν όλες τις μεταβολές των δεδομένων τους και τις πράξεις που τις προκαλούν ώστε να μπορούν να παρέχουν ομαλή εκτέλεση των δοσοληψιών. Οι μεταφορείς που εξάγουν τα δεδομένα από τέτοιες πηγές μπορούν να έχουν άμεση πρόσβαση στις μεταβολές που συντελούνται, αν τους δοθεί η δυνατότητα πρόσβασης στο αρχείο (log file) που καταγράφονται αυτές οι μεταβολές.
- **Triggers**: Σε περίπτωση που μία πηγή είναι ένα μοντέρνο σύστημα που παρέχει τη δυνατότητα δημιουργίας triggers, μπορούμε για κάθε πίνακα της πηγής να δημιουργήσουμε έναν trigger που θα μας ενημερώνει για οποιαδήποτε μεταβολή συμβαίνει στον πίνακα αυτόν.

5.5 Σχεδίαση Αποθηκών Δεδομένων

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

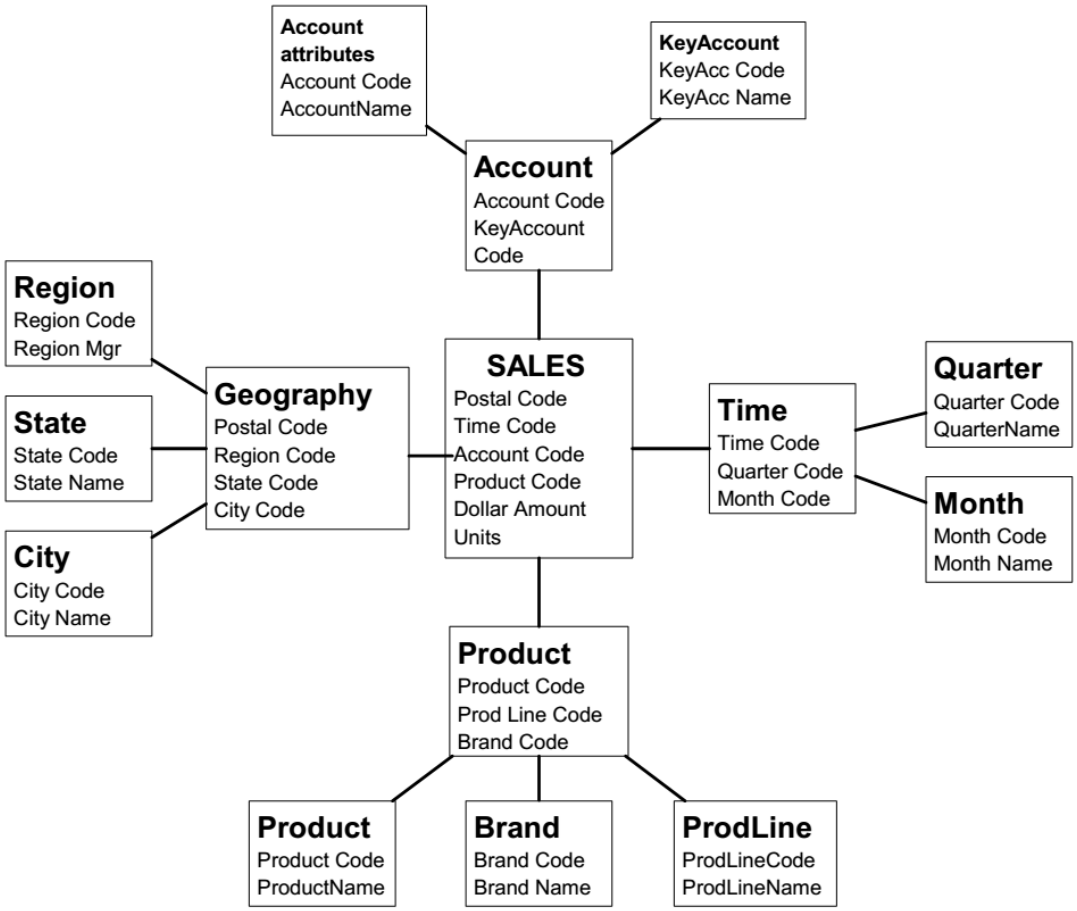
Οι πιο κατάλληλες τεχνικές για τη σχεδίαση των βάσεων των Αποθηκών Δεδομένων είναι τα *αστεροειδή σχήματα (star schemata)* και τα *σχήματα χιονονιφάδας (snowflake schemata)*. Το αστεροειδές σχήμα είναι πιο κοντά στο πολυδιάστατο χαρακτήρα των δεδομένων. Σε μία αστεροειδή βάση, υπάρχει ένας βασικός πίνακας που χαρακτηρίζεται *πίνακας συμβάντων (fact table)*. Υπάρχει επίσης ένας πίνακας για κάθε μία διάσταση. Κάθε εγγραφή του πίνακα συμβάντων αποτελείται από ένα δείκτη (ξένο κλειδί) σε μία εγγραφή κάθε ενός από τους πίνακες διαστάσεων. Κάθε *πίνακας διάστασης (dimension table)* περιλαμβάνει εγγραφές που αντιστοιχούν σε τιμές των διαστάσεων [21].



Εικόνα 7. Παράδειγμα αστεροειδούς σχήματος

Στην εικόνα 6, παρουσιάζεται ένα παράδειγμα αστεροειδούς σχήματος. Πρόκειται για το σχήμα Αποθήκης Δεδομένων που περιλαμβάνει δεδομένα σχετικά με τις πωλήσεις προϊόντων σε διάφορες πόλεις. Ο πίνακας των πωλήσεων (SALES) είναι στην προκειμένη περίπτωση ο πίνακας συμβάντων. Παρατηρούμε -για παράδειγμα- ότι μπορεί να εκτελεστεί οποιαδήποτε ερώτηση που συσχετίζει πωλήσεις, με τους λογαριασμούς (accounts) που παράγονται τα προϊόντα και με τις γεωγραφικές περιοχές που αυτά πωλούνται, εκτελώντας μόνο δύο πράξεις JOIN [21].

Η κύρια αδυναμία των αστεροειδών σχημάτων εντοπίζεται στον τρόπο με τον οποίο εκφράζουν τις ιεραρχίες των διαστάσεων. Για παράδειγμα, στη διάσταση χρόνος υπάρχει μία προφανής ιεραρχία μεταξύ ημερών, μηνών ετών κλπ. Μια εναλλακτική μοντελοποίηση των ιεραρχιών γίνεται από τα σχήματα χιονονιφάδας. Στην εικόνα 7παρουσιάζεται το παράδειγμα βάσης ίδιου περιεχομένου με τη βάση του εικόνας 6, αλλά οργανωμένη σύμφωνα με το σχήμα χιονονιφάδας. Πρόκειται για μία βελτίωση του αστεροειδούς σχήματος, όπου η ιεραρχία των διαστάσεων αναπαριστάται κανονικοποιώντας τους πίνακες των διαστάσεων.



Εικόνα 8. Παράδειγμα σχήματος χιονονιφάδας.

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

Εκτός από τους πίνακες συμβάντων και διαστάσεων, είναι πιθανόν να υπάρχουν και επιπρόσθετοι πίνακες με συγκεντρωτικά, προ-υπολογισμένα δεδομένα στην Αποθήκη Δεδομένων. Στην πιο απλή περίπτωση, τα συγκεντρωτικά δεδομένα αντιστοιχούν στην ομαδοποίηση των εγγραφών των πινάκων συμβάντων στη βάση συνδυασμού διαστάσεων. Στη βάση του παραδείγματος των εικόνων 6 και 7 μπορεί να προστεθούν πίνακες οι οποίοι να περιέχουν τις συνολικές πωλήσεις προϊόντων ανά γεωγραφική περιοχή και μονάδα χρόνου. Στην πραγματικότητα, δηλαδή, πρόκειται για πίνακες που περιέχουν πληροφορία που προκύπτει από τα δεδομένα του πίνακα συμβάντων. Η σκοπιμότητα ύπαρξης αυτών των πινάκων κρίνεται από την άφιξη σχετικών ερωτήσεων στη βάση από τις εφαρμογές ανάλυσης. Εναλλακτική της δημιουργίας τέτοιων πινάκων, είναι η εισαγωγή στους πίνακες συμβάντων εγγραφών που θα περιέχουν συγκεντρωτικές πληροφορίες για μερικές από τις διαστάσεις. Στην περίπτωση που θέλουμε να εισάγουμε στον πίνακα των πωλήσεων (SALES) συγκεντρωτικές τιμές ανά περιοχή και μονάδα χρόνου, θα εισάγουμε εγγραφές που

- τα πεδία που αντιστοιχούν στις διαστάσεις θα έχουν τις αντίστοιχες τιμές,
- τα πεδία που αντιστοιχούν στη τιμή που μετρά κάθε εγγραφή (units) θα υπάρχει η συγκεντρωτική τιμή, και
- στα πεδία των διαστάσεων που αθροίζονται (που δεν μας ενδιαφέρουν πλέον, δηλαδή) θα υπάρχουν τιμές NULL.

Η παρακάτω εγγραφή αντιστοιχεί στις συνολικές πωλήσεις του Ιουνίου στην περιοχή με κωδικό 16232 που παρέχει η αποθήκη δεδομένων των εικόνων 6 και 7.

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

Ευρετήρια Αποθηκών Δεδομένων

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

5.5.1 Δομές Ευρετηρίων και η χρήση τους

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

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

Εκτός από τα ευρετήρια τιμών σε ένα πίνακα, η δομή των αστεροειδών σχημάτων επιβάλει την χρήση των *ευρετηρίων σύνδεσης (join indices)*. Τα ευρετήρια αυτού του είδους παρέχουν τη συσχέτιση τη τιμής ενός ξένου κλειδιού ενός πίνακα με την αντίστοιχη τιμή του κλειδιού του πίνακα στον οποίο αναφέρεται. Σε μία βάση με αστεροειδές σχήμα μπορούμε να συσχετίσουμε τον πίνακα συμβάντων με τους πίνακες των διαστάσεων με τη χρήση των ευρετηρίων σύνδεσης. Για παράδειγμα, στη βάση της εικόνας 6 μπορεί να υπάρχει ένα ευρετήριο σύνδεσης στον κωδικό Postal Code που κρατά για κάθε διαφορετική πόλη τις εγγραφές στον πίνακα των συμβάντων που αντιστοιχούν στην πόλη αυτή.

5.5.2 Μετατροπή Πολύπλοκων Ερωτήσεων

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

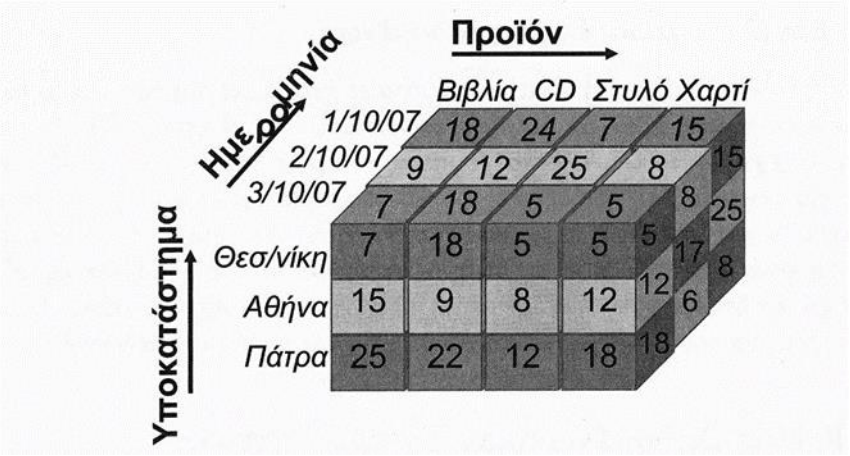
Η αποδοτική εκτέλεση των ερωτήσεων που περιλαμβάνουν πράξεις join μεταξύ πινάκων είναι επίσης αναγκαία. Ειδικές λύσεις προτείνονται για βάσεις με αστεροειδή σχήματα ή σχήματα χιονονιφάδας, καθώς η εκτέλεση ερωτήσεων στην Αποθήκη δεδομένων περιλαμβάνει, σε όλες σχεδόν τις περιπτώσεις, join μεταξύ του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

5.6 Κύβος Δεδομένων και Σχήμα Αστέρα

5.6.1 Κύβος δεδομένων

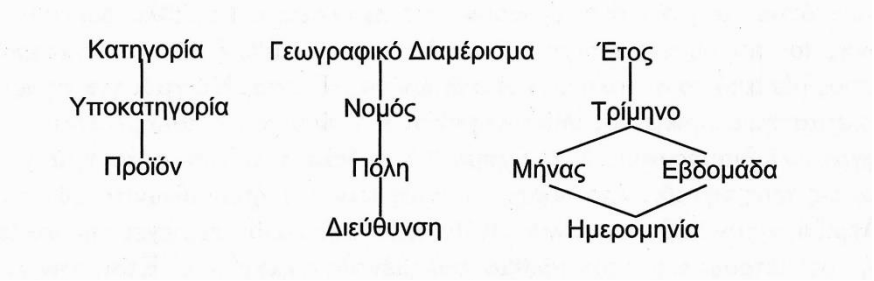
Κάθε αγορά δεδομένων είναι προσανατολισμένη σε μία μόνο διεργασία. Κατά τη διενέργεια μίας διεργασία συμβαίνουν *γεγονότα* (facts) που αφορούν σε αυτήν. Για παράδειγμα, σε μία αγορά δεδομένων που αφορά στις πωλήσεις, γεγονός αποτελεί η πώληση ενός συγκεκριμένου προϊόντος σε κάποιο υποκατάστημα, μία δεδομένη χρονική στιγμή. Μία αριθμητική ποσότητα που αφορά σε ένα γεγονός, ονομάζεται (αριθμητικό) *μέτρο* (measure). Στο γεγονός του προηγούμενου παραδείγματος, ένα μέτρο είναι ο αριθμός των τεμαχίων του προϊόντος που πωλήθηκαν σε πελάτες. Ένα γεγονός μπορεί να έχει περισσότερα μέτρα. Ένα άλλο παράδειγμα μέτρου είναι το συνολικό χρηματικό ποσό πώλησης όλων των τεμαχίων του προϊόντος. Οι πληροφορίες που περιγράφουν το γεγονός, ονομάζονται *διαστάσεις* (dimensions). Για ένα γεγονός πώλησης, διαστάσεις είναι, π.χ., το προϊόν που πωλήθηκε, το υποκατάστημα όπου έγινε η πώληση, η ημερομηνία πώλησης, κ.λπ. Κάθε διάσταση έχει ορισμένες ιδιότητες.

Σε εννοιολογικό επίπεδο, τα γεγονότα αναπαρίστανται ως πολυδιάστατοι κύβοι δεδομένων. Η εικόνα 8 απεικονίζει το παράδειγμα ενός κύβου δεδομένων, όπου τα γεγονότα αφορούν στις πωλήσεις ενός βιβλιοπωλείου. Κάθε άξονας του κύβου αντιστοιχεί σε μία διάσταση. Κάθε διάσταση αναπαρίσταται ως προς μία από το σύνολο των ιδιοτήτων της. Για παράδειγμα, για τη διάσταση υποκατάστημα, ορισμένες ιδιότητες είναι η διεύθυνση, η πόλη, ο νομός, και το γεωγραφικό διαμέρισμα. Στην Εικόνα 3.3, η διάσταση Υποκατάστημα απεικονίζεται ως προς την ιδιότητα πόλη. Οι τιμές των ιδιοτήτων όλων των διαστάσεων διαμερίζουν τον κύβο σε κελιά. Κάθε κελί του κύβου περιέχει την αντίστοιχη τιμή του μέτρου, π.χ., τον αριθμό πωληθέντων τεμαχίων. Έτσι, την 1/10/07 στο υποκατάστημα της Θεσσαλονίκης πωλήθηκαν συνολικά 7 βιβλία. Η αναπαράσταση με κύβο δεδομένων διευκολύνει πολύ την ανάλυση, επειδή η χρήση ενός κύβου μοιάζει με τη χρήση ενός φύλλου δεδομένων.



Εικόνα 9.Κύβος δεδομένων.

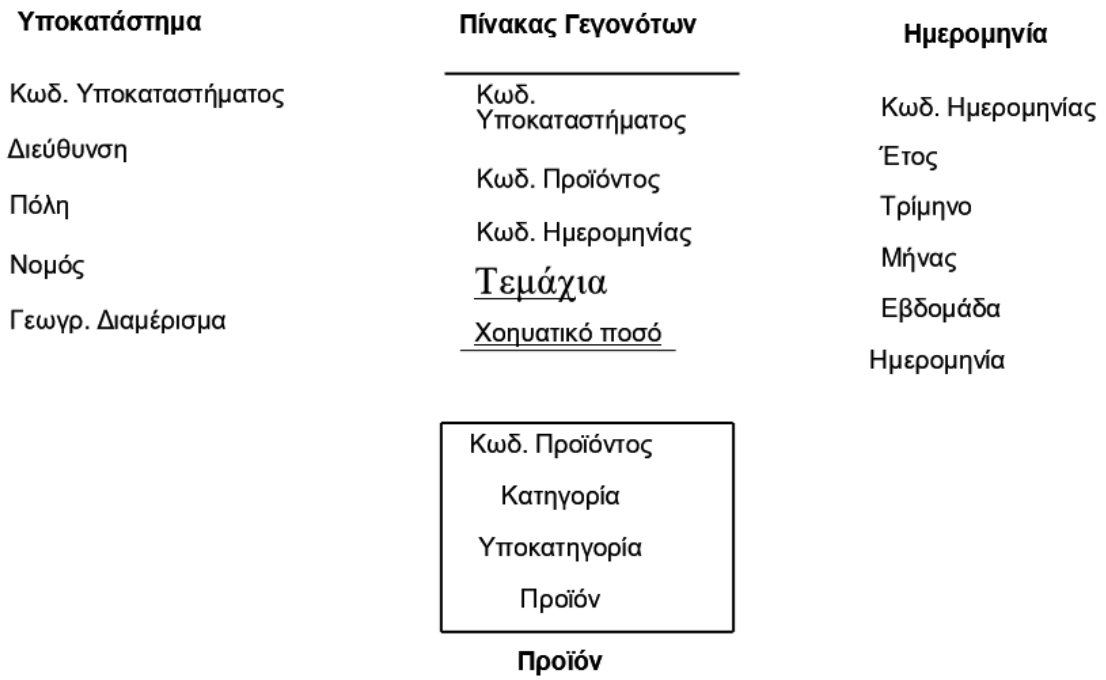
Σε μία διάσταση, κάποιες από τις ιδιότητές της μπορεί να απαρτίζουν ιεραρχία. Μία ιεραρχία αναπαριστά επίπεδα μεταξύ των ιδιοτήτων, από το ειδικότερο προς το γενικότερο, ως προς κάποια έννοια. Για παράδειγμα, για τη διάσταση υποκατάστημα, οι ιδιότητες διεύθυνση, η πόλη, ο νομός, και γεωγραφικό διαμέρισμα απαρτίζουν μία ιεραρχία ως προς τη γεωγραφική θέση του υποκαταστήματος. Έτσι, η διεύθυνση είναι ειδικότερη από την πόλη, ενώ η πόλη είναι ειδικότερη από το νομό, κ.ο.κ. Μία ιεραρχία απεικονίζεται με ένα πλέγμα, όπου κάθε κόμβος αντιστοιχεί σε μία ιδιότητα. Συνηθέστερη μορφή είναι ένα εκφυλισμένο πλέγμα, δηλαδή, μία γραμμική λίστα. Για τις διαστάσεις προϊόν, υποκατάστημα, και ημερομηνία, παραδείγματα ιεραρχιών απεικονίζονται στην Εικόνα 3.4. Για να διευκολύνουμε την ανάλυση, μπορούμε να αλλάζουμε το επίπεδο της ιεραρχίας σε κάθε ιδιότητα, παράγοντας έτσι διαφορετική όψη του κύβου δεδομένων. Για παράδειγμα, μπορούμε να εξετάσουμε τις πωλήσεις, ανά μήνα αντί ανά ημέρα, και να διακρίνουμε, π.χ., ότι τους μήνες Αύγουστο και Νοέμβριο είχαμε μειωμένες πωλήσεις σε σχέση με άλλους μήνες. Αυτές οι ενέργειες εκτελούνται με τις πράξεις OLAP, οι οποίες θα αναλυθούν στη συνέχεια.



Εικόνα 10. Ιεραρχίες Διαστάσεων

5.6.2 **Σχήμα αστέρα**

Ο σχεδιασμός ενός κύβου δεδομένων γίνεται με *σχήμα αστέρα* (star schema). Σε ένα σχήμα αστέρα, τα γεγονότα αναπαρίστανται στον *πίνακα γεγονότων* (fact table), ενώ κάθε διάσταση αναπαρίσταται με ξεχωριστό *πίνακα διαστάσεων* (dimension table). Η εικόνα 5 απεικονίζει το σχήμα αστέρα που αντιστοιχεί στον κύβο της εικόνας 3 [10] .



Εικόνα 11. Σχήμα αστέρα

Οι στήλες σε έναν πίνακα διάστασης αντιστοιχούν στο κύριο κλειδί (στην εικόνα 5 κλειδιά είναι οι κωδικοί) και στις υπόλοιπες ιδιότητες της διάστασης. Στον πίνακα γεγονότων, οι στήλες αντιστοιχούν σε ξένα κλειδιά, ένα προς κάθε πίνακα διάστασης. Επίσης, για κάθε αριθμητικό μέτρο, υπάρχει και μία ξεχωριστή στήλη (στην εικόνα 3.5 τα αριθμητικά μέτρα είναι τα Τεμάχια και το Χρηματικό ποσό).

Το σχήμα αστέρα παίρνει την ονομασία του από τη δομή που προκύπτει, με τον πίνακα γεγονότων στο κέντρο και τις διαστάσεις τοποθετημένες ακτινωτά γύρω του. Σε αντίθεση με ένα Διάγραμμα Οντοτήτων-Συσχετίσεων (ΔΟΣ), το διάγραμμα ενός σχήματος αστέρα γίνεται πολύ ευκολότερα αντιληπτό. Η

απλότητά του πηγάζει από το γεγονός ότι ο σχεδιασμός λαμβάνει υπόψη μόνο την ανάγκη σχεδιασμού ενός κύβου και όχι μιας σχεσιακής βάσης δεδομένων. Επειδή σε έναν κύβο δεν υπάρχει η ανάγκη εξυπηρέτησης συναλλαγών, δεν χρειάζονται οι σχεδιαστικές αρχές ενός ΔΟΣ. Το σημαντικότερο μειονέκτημα ενός σχήματος αστέρα είναι ότι δεν παρέχουν υποστήριξη για ρητή αναπαράσταση των ιεραρχιών που σχηματίζουν οι ιδιότητες των διαστάσεων. Ωστόσο, ο σχεδιασμός με σχήμα αστέρα αποτελεί τη δημοφιλέστερη μεθοδολογία.

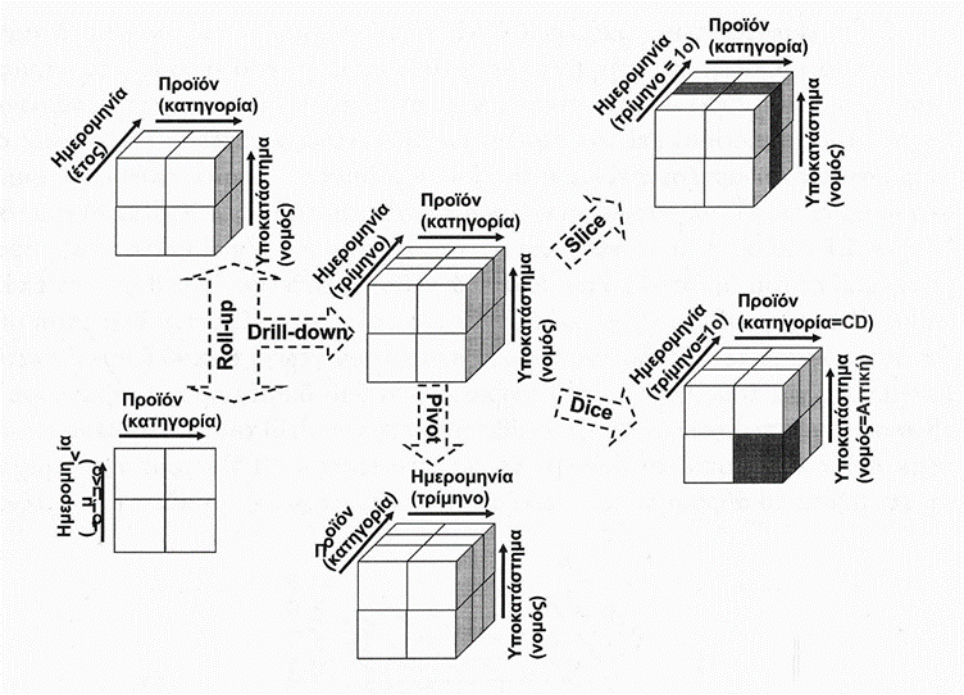
Εκτός του σχήματος αστέρα, υπάρχει και το *σχήμα χιονονιφάδας* (snowflake), όπου εφαρμόζεται κανονικοποίηση στους πίνακες διαστάσεων. Όπως θα αναφερθεί στη συνέχεια, η κανονικοποίηση των πινάκων διαστάσεων δεν είναι καλή πρακτική, κάτι που καθιστά το σχήμα χιονονιφάδας όχι αναγκαίο. Τέλος, η συνένωση πολλών σχημάτων αστέρα οδηγεί στο *όχημα αστερισμού* (constellation), τα οποία έχουν κοινό πίνακα γεγονότων, κάποιους κοινούς πίνακες διαστάσεων αλλά και κάποιους διαφορετικούς πίνακες διαστάσεων. Η συνένωση σχημάτων αστέρα, θα αναλυθεί περισσότερο στη συνέχεια.

5.7 Πράξεις OLAP

Η τεχνολογία OLAP βοηθά στην εύκολη διατύπωση αναλυτικών ερωτήσεων επί κύβων δεδομένων, καθώς και στη γρήγορη εκτέλεσή τους. Αντίθετα, η διατύπωση αντίστοιχων αναλυτικών ερωτημάτων σε σχεσιακά ΣΔΒΔ απαιτεί τη σύνταξη περίπλοκων ερωτημάτων στη γλώσσα SQL, τα οποία εκτελούνται στη συνέχεια με μεγάλο υπολογιστικό κόστος (10), (22).

Όλες οι πράξεις OLAP μπορούν να διατυπωθούν εύκολα με γραφικό τρόπο, αν και έχει αναπτυχθεί η γλώσσα *Multidimansional Expression Language* (MDX langauge), η οποία χρησιμοποιείται ως πρότυπο για την ανάπτυξη εργαλείων OLAP. Οι συνηθέστερα χρησιμοποιούμενες πράξεις OLAP παρουσιάζονται στην εικόνα 12 και αναλύονται στη συνέχεια (10);

Roll-up. Παράγει κύβο δεδομένων με μειωμένο επίπεδο λεπτομέρειας. Η πράξη roll-up γίνεται όταν: (α) σε κάποιες διαστάσεις επιλέγουμε ανώτερο επίπεδο στην ιεραρχία τους ή (β) αφαιρούμε κάποιες διαστάσεις. Στην εικόνα 6, από τον κεντρικό κύβο παράγουμε τον επάνω-αριστερά ανεβαίνοντας στην ιεραρχία της διάστασης Ημερομηνία (από το επίπεδο τρίμηνο στο επίπεδο έτος). Ο κάτω-αριστερά κύβος παράγεται αφαιρώντας τη διάσταση Υποκατάστημα.



Εικόνα 12. **Πράξεις OLAP**

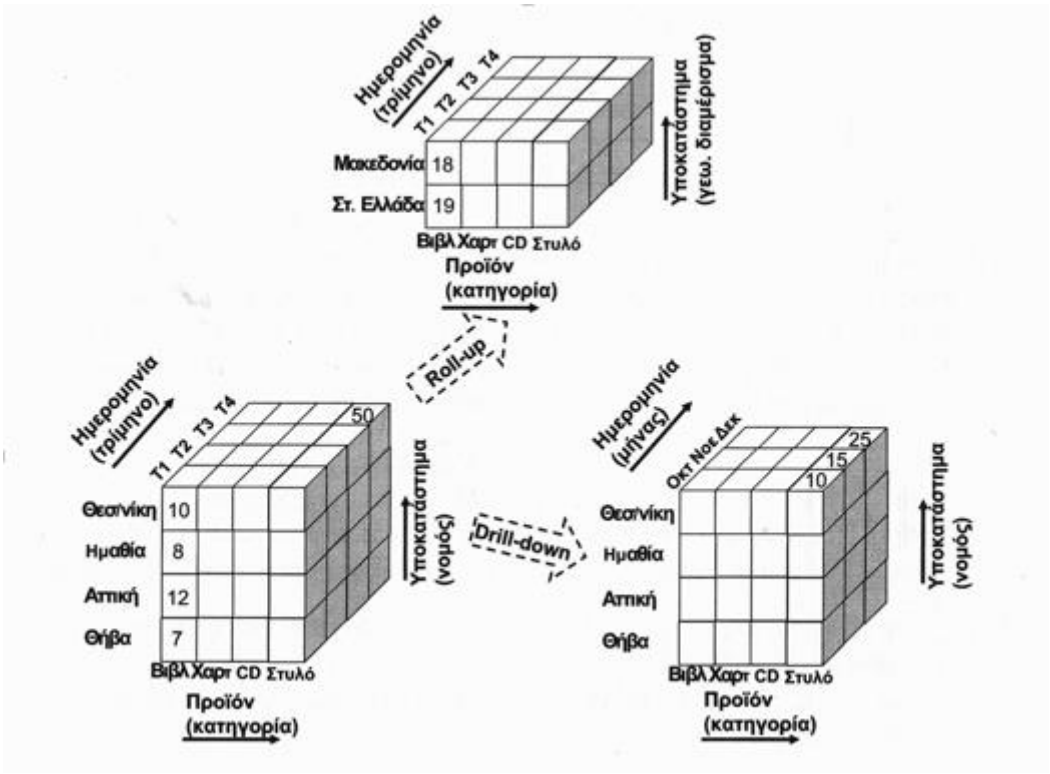
Drill-down. Παράγει κύβο δεδομένων με αυξημένο επίπεδο λεπτομέρειας. Η πράξη drill-down γίνεται όταν: (α) σε κάποιες διαστάσεις επιλέγουμε κατώτερο επίπεδο στην ιεραρχία τους ή (β) προσθέτουμε κάποιες διαστάσεις. Στην Εικόνα 6, από τους δύο αριστερούς κύβους, παράγουμε τον κεντρικό κύβο με αντίστροφες πράξεις από ότι στην περίπτωση του roll-up.

Slice. Η πράξη αυτή παράγει κύβο εφαρμόζοντας επιλογή σε μία μόνο διάσταση (αντιστοιχεί στις σχεσιακές πράξεις επιλογής και προβολής). Στην Εικόνα 6, από τον κεντρικό κύβο παράγουμε το σκιασμένο τμήμα του επάνω-δεξιά κύβου, θέτοντας κριτήρια επιλογής στη διάσταση Ημερομηνία.

Dice. Π πράξη αυτή παράγει κύβο εφαρμόζοντας επιλογή σε μία ή περισσότερες διαστάσεις. Στην εικόνα 6, από τον κεντρικό κύβο παράγουμε το σκιασμένο τμήμα του κάτω-δεξιά κύβου, θέτοντας κριτήρια επιλογή σε όλες τις διαστάσεις.

Pivot. Παράγει κύβο με άλλη διάταξη των διαστάσεων. Στην εικόνα 6, από τον κεντρικό κύβο η πράξη pivot παράγει τον κάτω κύβο, αντιμεταθέτοντας τις διαστάσεις Προϊόν και Ημερομηνία.

Με τη βοήθεια των πράξεων OLAP είναι εύκολη η ανάλυση με διαισθητικό τρόπο. Όμως, για την εφαρμογή τους απαιτείται ο ορισμός του τρόπου παραγωγής των κύβων-αποτελεσμάτων, μέσω μίας *συναθροιστικής συνάρτησης* (aggregation function) επί των τιμών των αριθμητικών μέτρων. Οι βασικές συναθροιστικές συναρτήσεις είναι αυτές του αθροίσματος (sum), πλήθους (count),μέσου όρου (avg), μεγίστου (max), και ελαχίστου (min). Για παράδειγμα, στην εικόνα 7 ορίζουμε τη συναθροιστική συνάρτηση του αθροίσματος. Με εφαρμογή πράξης roll-up στον κάτω-αριστερά κύβο πωλήσεων, παίρνουμε τον επάνω κύβο. Η πράξη roll-up έγινε αλλάζοντας επίπεδο ιεραρχίας στη διάσταση υποκατάστημα, από το επίπεδο του νομού σε αυτό του γεωγραφικού διαμερίσματος. Επομένως, για τους νομούς που ανήκουν στο ίδιο διαμέρισμα, λαμβάνουμε το άθροισμα των πωλήσεων. Στην αντίθετη πράξη του drill-down, λαμβάνουμε τον κάτω δεξιά κύβο, όπου αναλύουμε το τέταρτο τρίμηνο (Τ4) στους αντίστοιχους μήνες, οπότε το άθροισμα των πωλήσεων του τριμήνου επιμερίζεται ανά μήνα.



Εικόνα 13. Πράξεις OLAP με εφαρμογή συνθροιστικής συνάρτησης.

Για την εφαρμογή μίας συναθροιστικής συνάρτησης κατά την εκτέλεση μίας πράξης OLAP, πρέπει οι εμπλεκόμενες διαστάσεις να υποστηρίζουν τη συνάρτηση. Αυτό δεν είναι πάντοτε εφικτό.

Μέρος ΙΙ: Συμβολή διατριβής

Κεφάλαιο 5ο

6 Υλοποίηση του Ανιχνευτή Επιθέσεων

Σε αυτό το κεφάλαιο περιγράφεται η υλοποίηση του Ανιχνευτή Επιθέσεων μέσα από μία οκτώ σειρά βημάτων, τα οποία αφορούν την συλλογή, προετοιμασία, προεπεξεργασία των δεδομένων, την κατασκευή της Αποθήκης Δεδομένων και τέλος την εξόρυξη γνώσης από αυτήν. Η υλοποίηση του Ανιχνευτή Επιθέσεων αποτελεί και την συμβολή της διατριβής (1), (23) (24).

Σύνολο δεδομένων (Data Set)

Το σύνολο δεδομένων που θα χρησιμοποιηθεί για την εξόρυξη δεδομένων είναι το KDD CUP 1999 από το UCI ML Repository, το οποίο αποτελείται από 41 χαρακτηριστικά (34 αριθμητικά – συνεχόμενων τιμών και 7 κατηγοριακά) και αντιστοιχεί στο 10% ενός αρχικού συνόλου δεδομένων περιλαμβάνοντας συνολικά 494,021 εγγραφές δεδομένων. Επιπλέον στο σύνολο δεδομένων περιλαμβάνεται και η μεταβλητής κλάσης, με όνομα label, η οποία αναφέρεται στη μορφή πιθανής επίθεσης (22 διαφορετικές μορφές επιθέσεων). Όταν η τιμή του χαρακτηριστικού label είναι διαφορετικό της τιμής “normal”, η συγκεκριμένη σύνδεση αντιστοιχεί σε μιας μορφής επίθεσης (1).

Το αρχείο δεδομένων KDD CUP 99 περιλαμβάνει πληροφορίες συνδέσεων δεδομένων, τις οποίες και ταξινομεί ως (i) φυσιολογικές ή τις θεωρεί ως (ii) προσπάθειες επίθεσης. Όσες εγγραφές θεωρούνται πως αντιστοιχούν σε επιθέσεις (22 διαφορετικές μορφές επιθέσεων), ανήκουν στα εξής τέσσερις (4) γενικές μορφές επιθέσεων:

1. DoS (Deny of service)
2. Probe (Information fathering)
3. U2R (User to Root)
4. U2L (Remote to Local)

Στον παρακάτω πίνακα , παρουσιάζεται σε ποια από τις 4 βασικές μορφές επιθέσεων αντιστοιχεί, κάθε πιθανή τιμή της μεταβλητής κλάσης:[1]

Κατηγορία επιθέσεων	**Ονομασία σχετικών επιθέσεων**
DoS	back, land, neptune, pod, smurf, teardrop
Probe	satan, ipsweep, nmap, portsweep
U2R	buffer_overflow, loadmodule, perl, rootkit
U2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

Πίνακας 3. Ονομασία επιθέσεων που αντιστοιχούν σε κάθε βασική κατηγορία επιθέσεων

Τα χαρακτηριστικά του συνόλου δεδομένων, μπορούν να κατηγοριοποιηθούν σε τρεις (3) βασικές κατηγορίες δεδομένων:

- **Βασικά χαρακτηριστικά** (Basic features): αναφέρεται σε χαρακτηριστικά που σχετίζονται με τα χαρακτηριστικά μιας σύνδεσης με χρήση του πρωτοκόλλου TCP/IP

- **Χαρακτηριστικά κυκλοφορίας – κίνησης** (Traffic features): χωρίζονται σε δυο μεγάλες κατηγορίες: (i) χαρακτηριστικά σχετικά με συνδέσεις προς τον ίδιο κόμβο τα τελευταία 2 δευτερόλεπτα και (ii) χαρακτηριστικά σχετικά με συνδέσεις στην ίδια υπηρεσία τα τελευταία 2 δευτερόλεπτα.

- **Χαρακτηριστικά περιεχομένου** (Content features): αναφέρονται σε χαρακτηριστικά που περιγράφουν μια προσπάθεια επίθεσης, όπως ο αριθμός αποτυχημένων προηγούμενων προσπαθειών σύνδεσης κτλ

Αναλυτικά, το σύνολο δεδομένων περιλαμβάνει τα εξής 42 χαρακτηριστικά:

ΑΑ	**Χαρακτηριστικό**	**Περιγραφή**	**Μορφή**
1	Duration	αριθμός δευτερολέπτων της σύνδεσης	Αριθμητική
2	protocol_type	τύπος του πρωτοκόλλου, πχ tcp, udp	Κατηγοριακή
3	Service	υπηρεσιών προορισμού, πχ http, telnet	Κατηγοριακή

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/training_attack_types

ΑΑ	Χαρακτηριστικό	Περιγραφή	Μορφή
4	Flag	κανονική ή λάθος στη σύνδεση	Κατηγοριακή
5	src_bytes	αριθμό δεδομένων bytes από την πηγή στον προορισμό	Αριθμητική
6	dst_bytes	τον αριθμό των bytes δεδομένων από τον προορισμό στην πηγή	Αριθμητική
7	Land	1: εάν η σύνδεση είναι από / προς το ίδιο port/host 0: διαφορετικά	Κατηγοριακή
8	wrong_fragment	αριθμός των λανθασμένων fragments	Αριθμητική
9	urgent	αριθμός επειγόντων πακέτων	Αριθμητική
10	Hot	αριθμός hot indicators	Αριθμητική
11	num_failed_logins	αριθμός των αποτυχημένες προσπαθειών σύνδεσης	Αριθμητική
12	logged_in	1: αν έχετε συνδεθεί με επιτυχία, 0: διαφορετικά	Κατηγοριακή
13	num_compromised	αριθμός επικίνδυνων συνθηκών	Αριθμητική
14	root_shell	1: εάν ληφθεί root shell, 0: διαφορετικά	Κατηγοριακή
15	su_attempted	1: εάν υπάρχει “su root” εντολή απόπειρα, 0: διαφορετικά	Κατηγοριακή
16	num_root	αριθμός των root προσβάσεων	Αριθμητική
17	num_file_creations	αριθμός των πράξεων δημιουργίας αρχείου	Αριθμητική
18	num_shells	Αριθμός αιτημάτων κελύφους	Αριθμητική

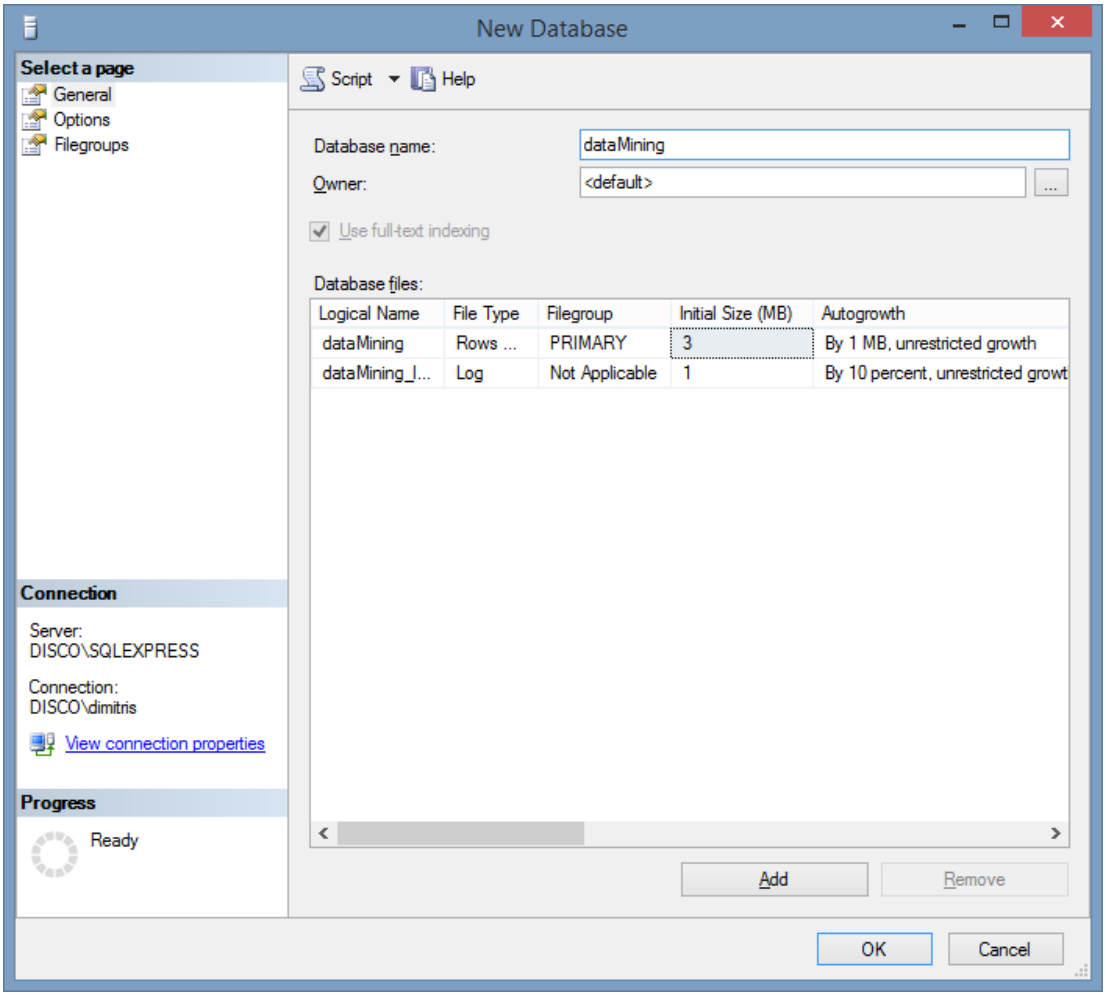
ΑΑ	Χαρακτηριστικό	Περιγραφή	Μορφή
19	num_access_files	αριθμό πράξεων για τον έλεγχο αρχείων πρόσβασης	Αριθμητική
20	num_outbound_cmds	αριθμός των εξερχόμενων εντολών σε ένα FTP session	Αριθμητική
21	is_host_login	1: αν η είσοδος ανήκει στη host list 0: διαφορετικά	Κατηγοριακή
22	is_guest_login	1: εάν η σύνδεση είναι guest, 0: διαφορετικά	Κατηγοριακή
23	Count	αριθμός των συνδέσεων στον ίδιο host με την τρέχουσα σύνδεση κατά τα τελευταία δύο δευτερόλεπτα	Αριθμητική
24	srv_count	αριθμός των συνδέσεων στην ίδια υπηρεσία με την τρέχουσα σύνδεση κατά τα τελευταία δύο δευτερόλεπτα	Αριθμητική
25	serror_rate	% συνδέσεων με SYN error	Αριθμητική
26	srv_serror_rate	% συνδέσεων (υπηρεσία) με SYN error	Αριθμητική
27	rerror_rate	% συνδέσεων με REJ error	Αριθμητική
28	srv_rerror_rate	% συνδέσεων (υπηρεσία) με REJ error	Αριθμητική
29	same_srv_rate	% συνδέσεων στην ίδια υπηρεσία	Αριθμητική
30	diff_srv_rate	% συνδέσεων σε διαφορετική υπηρεσία	Αριθμητική
31	srv_diff_host_rate	% συνδέσεων με διαφορετικούς host	Αριθμητική
32	dst_host_count	αριθμός των συνδέσεων στον ίδιο host με την τρέχουσα σύνδεση κατά τα τελευταία δύο δευτερόλεπτα	Αριθμητική

ΑΑ	Χαρακτηριστικό	Περιγραφή	Μορφή
33	dst_host_srv_count	αριθμός των συνδέσεων στην ίδια υπηρεσία με την τρέχουσα σύνδεση κατά τα τελευταία δύο δευτερόλεπτα	Αριθμητική
34	dst_host_same_srv_rate	% συνδέσεων στην ίδια υπηρεσία	Αριθμητική
35	dst_host_diff_srv_rate	% συνδέσεων σε διαφορετική υπηρεσία	Αριθμητική
36	dst_host_same_src_port_rate	% συνδέσεων με ίδιο source port	Αριθμητική
37	dst_host_srv_diff_host_rate	% συνδέσεων σε άλλο host	Αριθμητική
38	dst_host_serror_rate	% συνδέσεων με SYN error	Αριθμητική
39	dst_host_srv_serror_rate	% συνδέσεων (υπηρεσία) με SYN error	Αριθμητική
40	dst_host_rerror_rate	% συνδέσεων με REJ error	Αριθμητική
41	dst_host_srv_rerror_rate	% συνδέσεων (υπηρεσία) με REJ error	Αριθμητική
42	label	Χαρακτηρισμός σύνδεσης (**μεταβλητή κλάσης**)	Κατηγοριακή

Πίνακας 4: Περιγραφή χαρακτηριστικών συνόλου δεδομένων

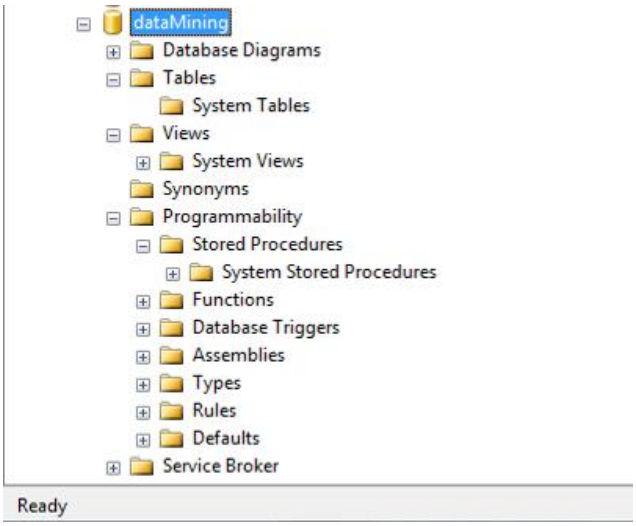
Αρχική μεταφορά δεδομένων στον SQL Server

Αρχικά θα πρέπει να δημιουργηθεί μια καινούργια βάση δεδομένων στον SQL Server. Το όνομα που θα δοθεί στην καινούργια βάση δεδομένων είναι dataMining:



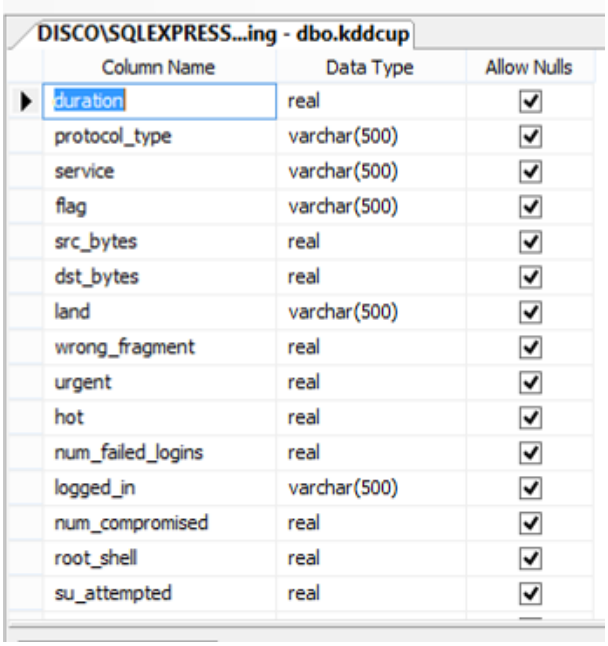
Εικόνα 8: Δημιουργία καινούργιας κενής βάσης δεδομένων

Η βάση δεδομένων που δημιουργήθηκε είναι κενή αρχικά, χωρίς δεδομένα:



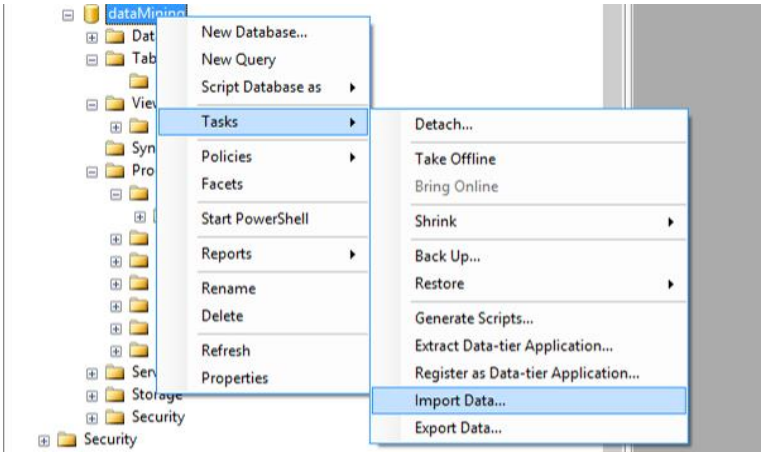
Εικόνα 9. Αρχικά περιεχόμενα βάσης δεδομένων

Στην συνέχεια, θα δημιουργηθεί πίνακας στη βάση δεδομένων dataMinning με το όνομα kddcup, με βάση την περιγραφή των χαρακτηριστικών του συνόλου δεδομένων που αναφέρονται στον Πίνακα 4:



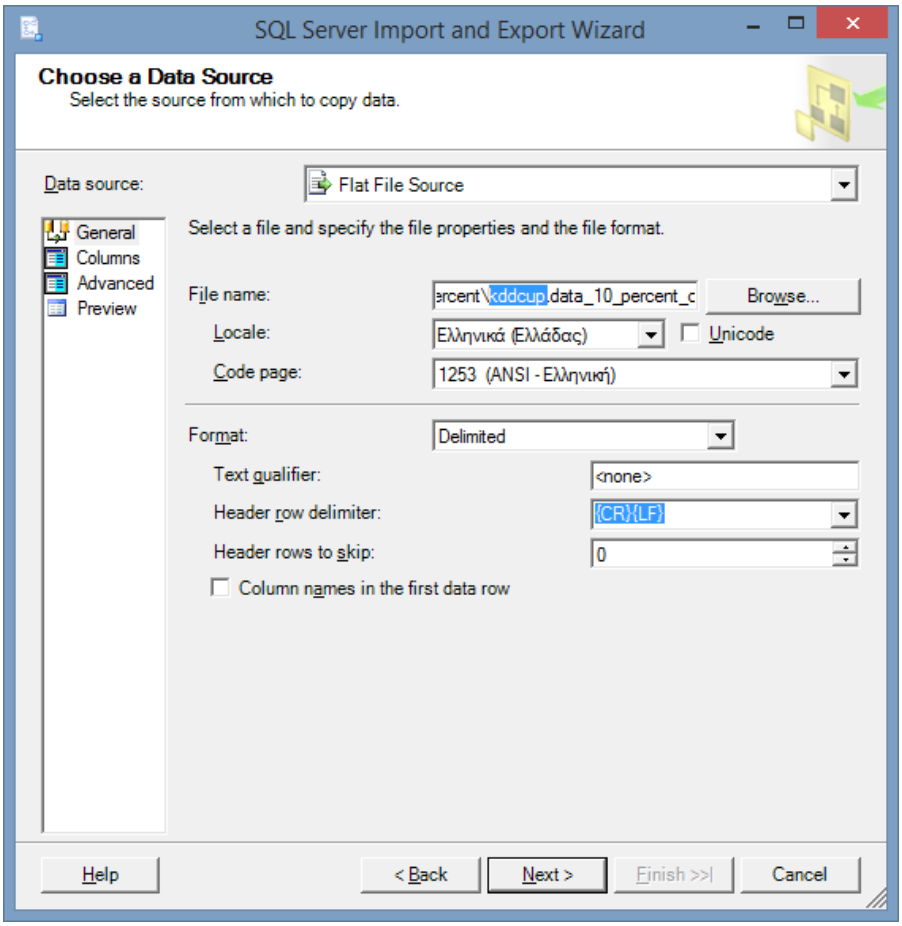
Εικόνα 10. Δημιουργία πίνακα με όνομα kddcup στη βάση δεδομένων dataMinning

Στην συνέχεια , θα εισαχθούν τα δεδομένα του data set στην βάση δεδομένων που μόλις δημιουργήθηκε. Για τον λόγο αυτόν, αρχικά επιλέγεται εισαγωγή δεδομένων για τη βάση δεδομένων dataMining:



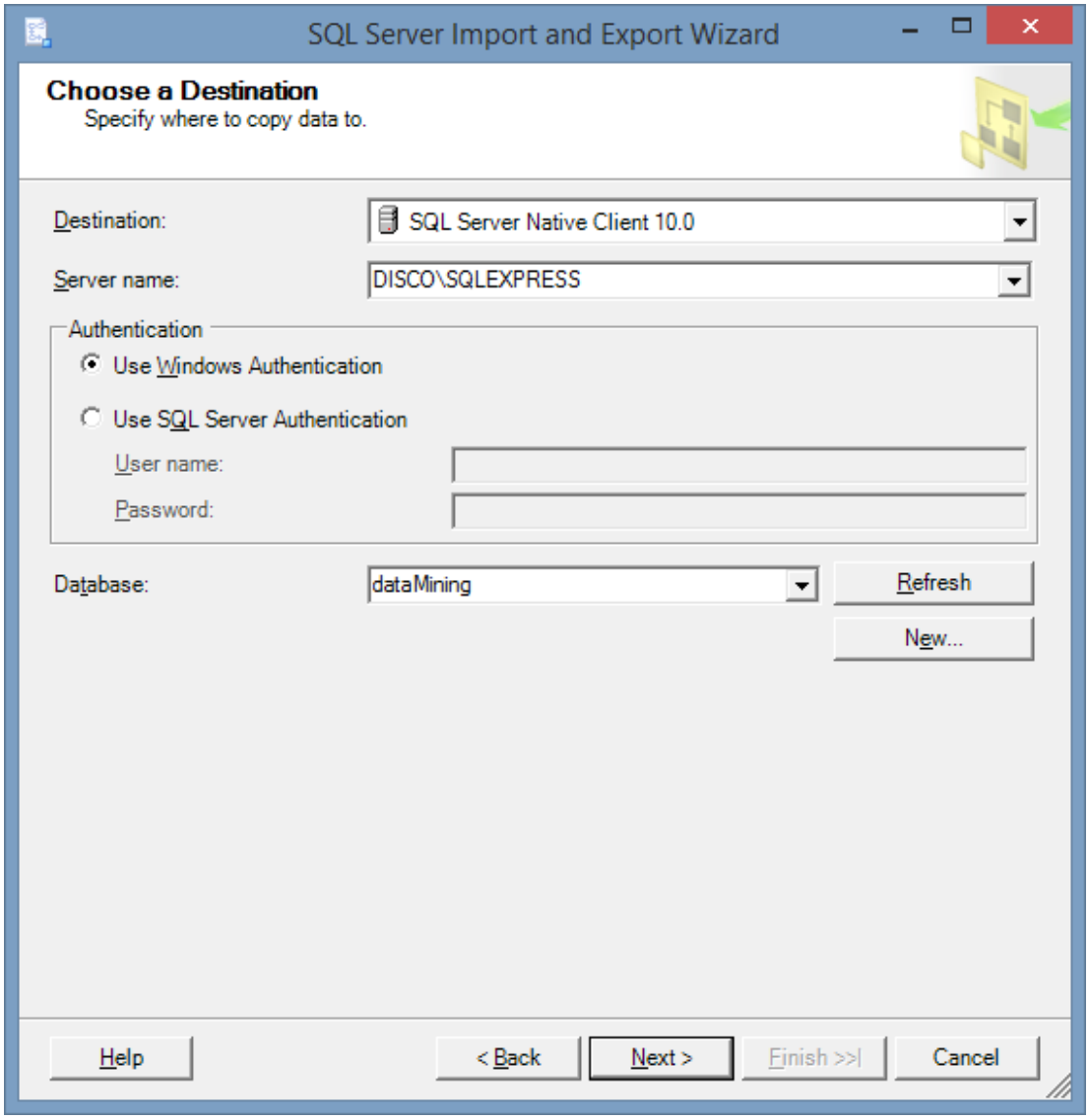
Εικόνα 11. Επιλογή εισαγωγής δεδομένων σε βάση δεδομένων του SQL Server

Στην φόρμα που εμφανίζεται επιλέγεται ως μορφή του αρχείου εισαγωγής δεδομένων: αρχείο κειμένου (Flat File Source) και επιλέγεται το αρχείο “kddcup.data_10_percent_corrected” που περιλαμβάνει τα αρχεία της παρούσας άσκησης. Τις υπόλοιπες τιμές των πεδίων, στην φόρμα εισαγωγής δεδομένων, τις αφήνουμε με τις προεπιλεγμένες τους τιμές:



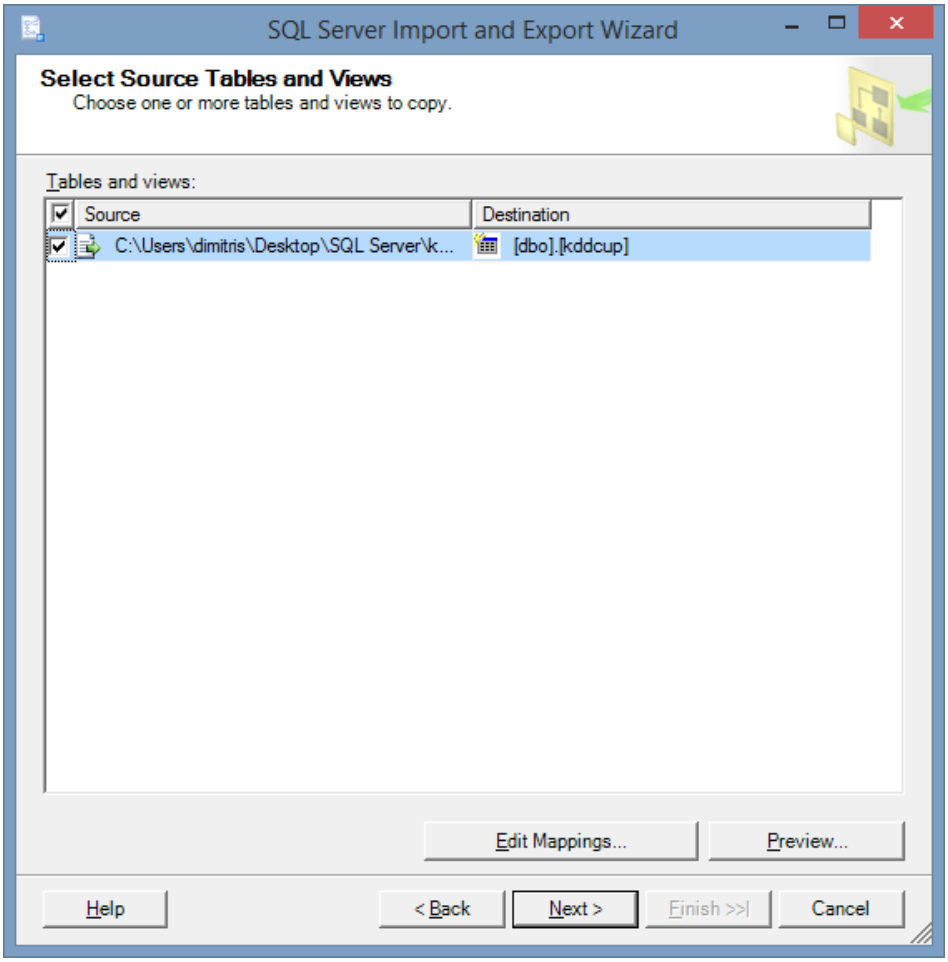
Εικόνα 12. Επιλογή αρχείου εισαγωγής δεδομένων στον SQL Server

Στην επόμενη φόρμα της διαδικασίας εισαγωγής δεδομένων, επιλέγεται η τοποθεσία εισαγωγής των δεδομένων, όπου επιλέγεται η κενή βάση dataMining, που μόλις δημιουργήθηκε:



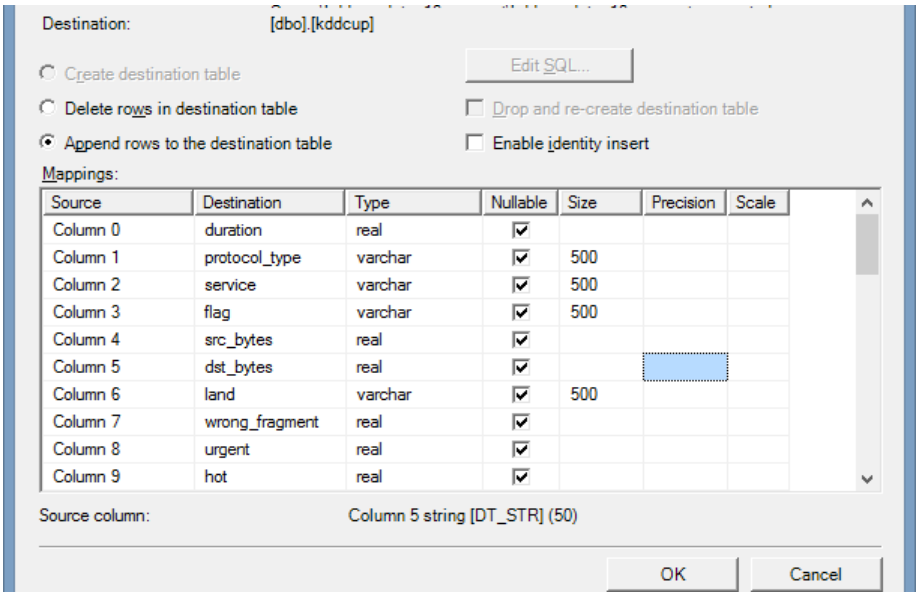
Εικόνα 13. Επιλογή προορισμούς εισαγωγής δεδομένων

Πατώντας το κουμπί επόμενο, στη φόρμα που εμφανίζεται, μπορεί να γίνει προεπισκόπηση των αποτελεσμάτων εισαγωγής των δεδομένων στη βάση δεδομένων dataMining:



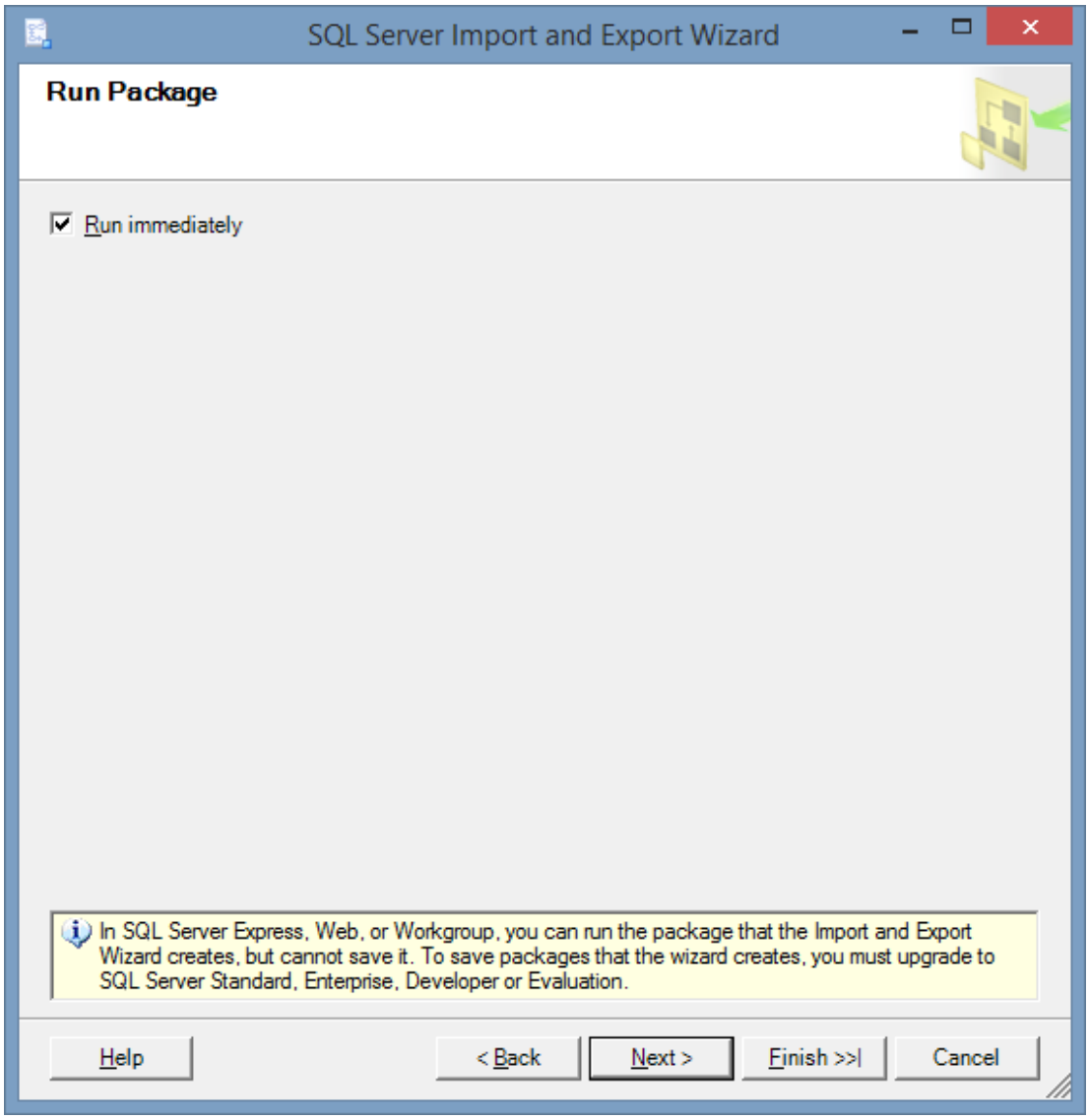
Εικόνα 14. Αντιστοίχηση μεταφοράς δεδομένων μεταξύ πηγών δεδομένων

Πατώντας το κουμπί Edit Mappings, εμφανίζεται ποιο πεδίο του αρχείου δεδομένων, αντιστοιχεί σε ποιο πεδίο του πίνακα kddcup:



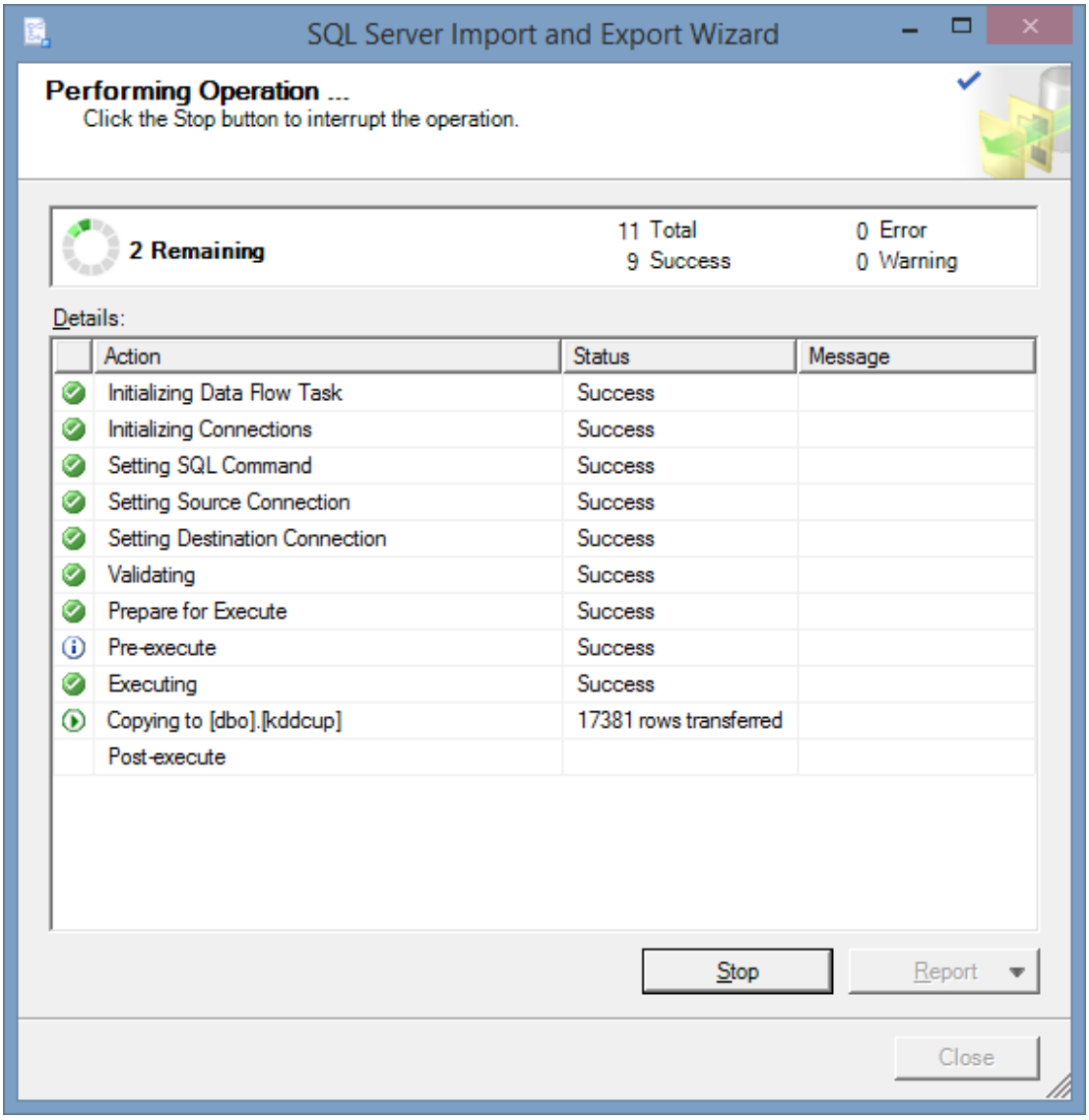
Εικόνα 15. Αντιστοίχηση πεδίων αρχείου κειμένου με πεδία του πίνακα kddcup

Στην συνέχεια, πατώντας το κουμπί επόμενο, επιλέγεται η άμεση εκτέλεσης της μεταφοράς δεδομένων:



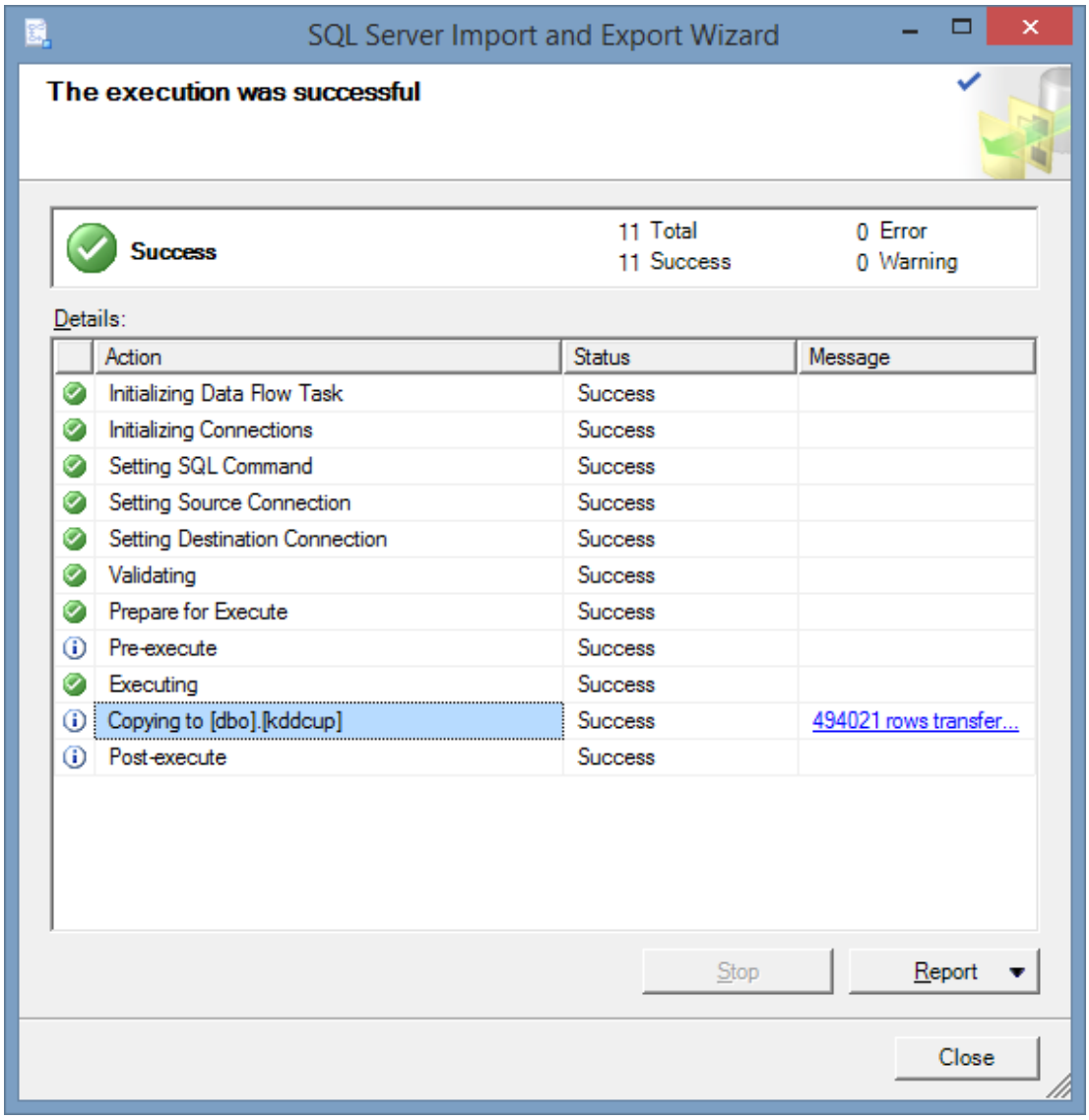
Εικόνα 16. Επιλογή άμεσης εκτέλεσης μεταφοράς δεδομένων

Στην επόμενη φόρμα – σελίδα που εμφανίζεται, πατώντας το κουμπί Finish, ξεκινάει άμεσα η μεταφορά των δεδομένων, με δυναμική προβολή της πορείας μεταφοράς των δεδομένων:



Εικόνα 14: Ενημέρωση πορείας μεταφοράς δεδομένων

Και με το πέρας της ολοκλήρωσης μεταφοράς των δεδομένων, η εφαρμογή μας ενημερώνει πως μεταφέρθηκαν 494,021 εγγραφές:



Εικόνα 18. Ενημέρωση επιτυχίας μεταφοράς δεδομένων και αριθμού σχετικών εγγραφών

Πλέον έχει ολοκληρωθεί η μεταφορά των δεδομένων στη βάση δεδομένων *dataMinning* και πιο συγκεκριμένα στον πίνακα *kddcup*.

Στη συνέχεια θα προσθέσουμε πεδίο με όνομα attackType, το οποίο θα αντιστοιχεί στη κατηγορία της επίθεσης κάθε εγγραφής και θα γεμίσει με τις αντίστοιχες κατηγορίες με την εκτέλεση του παρακάτω ερωτήματος:

```
update kddcup set attackType='dos'

where label ='back' or label='land' or label='neptune' or label='pod' or label='smurf' or
label='teardrop'



update kddcup set attackType='probe'

where label ='satan' or label='ipsweep' or label='nmap' or label='portsweep '



update kddcup set attackType='u2r'

where label ='buffer_overflow' or label='loadmodule' or label='perl' or label='rootkit'



update kddcup set attackType='u2l'

where label ='ftp_write' or label='guess_passwd' or label='imap' or label='multihop' or
label='phf' or label='spy'

or label='warezclient' or label='warezmaster'



update kddcup set attackType='normal'

where label ='normal'
```

6.1 **Προπαρασκευή δεδομένων**

Στη συνέχεια πρέπει να γίνει η απαραίτητη προπαρασκευή δεδομένων που περιλαμβάνει για το σύνολο δεδομένων της παρούσας εργασίας την αφαίρεση χαρακτηριστικών που δεν επηρεάζουν την τιμή της εξαρτημένης μεταβλητής, την αφαίρεση εγγραφών με λανθασμένες τιμές και τη μετατροπή λοιπών χαρακτηριστικών από αριθμητική μορφή σε κατηγοριακή μορφή, προκειμένου να είναι ευκολότερη η κατασκευή σε επόμενο χρόνο του κύβου OLAP[12], [13],[14].

Αρχικά, εξετάζονται ποια χαρακτηριστικά έχουν μικρή επίδραση στην αξιολόγηση μια σύνδεσης αν αντιστοιχεί σε επίθεση ή όχι, προκειμένου να μικρύνει το σύνολο δεδομένων. Πιο συγκεκριμένα εξετάζεται η συχνότητα εμφάνισης των διαφόρων τιμών στα χαρακτηριστικά του δείγματος, εκτελώντας για όλα τα πεδία ερωτήματα της μορφής:

```
select distinct land, count(*) as sinolo

from kddcup

group by land
```

Μετά την παραπάνω διαδικασία, διαπιστώνετε πως για τα παρακάτω πεδία, η επίδρασή του στην αξιολόγηση μιας σύνδεσης ως επίθεση ή όχι, είναι μικρή, εξαιτίας της πλειονότητας ίδιων τιμών στις εγγραφές του συνόλου δεδομένων:

Πεδίο land

Τιμή	Συχνότητα
1	22
0	493999

Πεδίο wrong_fragment

Τιμή	Συχνότητα
0	492783
1	970
3	268

Πεδίο urgent

Τιμή	Συχνότητα
0	494017
3	1
2	2
1	1

Πεδίο hot

Τιμή	Συχνότητα
0	490829
Υπόλοιπες τιμές	3192

Πεδίο num_failed_logins

Τιμή	Συχνότητα
0	493958
3	1
5	1
4	1
1	57
2	3

Πεδίο num_compromised

Τιμή	Συχνότητα
0	491797
Υπόλοιπες τιμές	2224

Πεδίο root_shell

Τιμή	Συχνότητα
1	493966
0	55

Πεδίο su_attempted

Τιμή	Συχνότητα
0	494009
1	6
2	6

Πεδίο num_root

Τιμή	Συχνότητα
0	493436
Υπόλοιπες τιμές	585

Πεδίο num_file_creations

Τιμή	Συχνότητα
0	493756
Υπόλοιπες τιμές	265

Πεδίο num_shells

Τιμή	Συχνότητα
0	493970
1	48
2	3

Πεδίο num_access_files

Τιμή	Συχνότητα
0	493567
Υπόλοιπες τιμές	454

Πεδίο num_outbound_cmds

Τιμή	Συχνότητα
0	494021

Πεδίο is_host_login

Τιμή	Συχνότητα
0	494021

Πεδίο is_guest_login

Τιμή	Συχνότητα
0	493336
1	685

Τα παραπάνω πεδία δε θα ληφθούν υπόψη κατά την κατασκευή του κύβου OLAP για τη μελέτη του συνόλου δεδομένων. Πιο συγκεκριμένα δε θα συμπεριληφθούν στην ανάλυση δεκαπέντε (15) χαρακτηριστικά με αύξοντα αριθμό: 7, 8 , 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 καθώς η πλειονοτήτων των τιμών για τα αντίστοιχα χαρακτηριστικά, με πολύ μεγάλο ποσοστό είναι η ίδια για όλες τις εγγραφές.

Τα χαρακτηριστικά που απομένουν πλέον είναι εικοσιέξι (26) για το σύνολο δεδομένων της παρούσας έρευνας.

Επίσης, από την παραπάνω ανάλυση, διαπιστώνετε πως υπάρχουν χαρακτηριστικά, όπου μια τιμή ενός χαρακτηριστικού, εμφανίζεται σε περισσότερες από το 90% του συνόλου των εγγραφών, με την εκτέλεση ερωτημάτων της μορφής:

```
declare @sinolo as real
```

```
select @sinolo=count(*)

from kddcup



select distinct dst_host_srv_rerror_rate, 100 * count(*)/@sinolo

from kddcup

group by dst_host_srv_rerror_rate

order by dst_host_srv_rerror_rate
```

Ομοίως, τα αντίστοιχα χαρακτηριστικά θα αφαιρεθούν από το σύνολο δεδομένων που θα χρησιμοποιηθεί για την υλοποίηση του κύβου (1,27, 28,31,40,41):

Πεδίο duration

Τιμή	Συχνότητα
0	97.5 %
Υπόλοιπες τιμές	2.5 %

Πεδίο rerror_rate

Τιμή	Συχνότητα
0	94.1 %
Υπόλοιπες τιμές	5.9 %

Πεδίο srv_rerror_rate

Τιμή	Συχνότητα
0	94 %
Υπόλοιπες τιμές	6 %

Πεδίο srv_diff_host_rate

Τιμή	Συχνότητα
0	93 %
Υπόλοιπες τιμές	7 %

Πεδίο dst_host_rerror_rate

Τιμή	Συχνότητα
0	92.9 %
Υπόλοιπες τιμές	7.1 %

Πεδίο dst_host_srv_rerror_rate

Τιμή	Συχνότητα
0	93.1 %
Υπόλοιπες τιμές	6.9 %

Μετά τις παραπάνω αφαιρέσεις χαρακτηριστικών από το σύνολο δεδομένων, τα υποψήφια χαρακτηριστικά για συμπερίληψη στην αποθήκη δεδομένων που θα υλοποιηθεί είναι είκοσι (20) χαρακτηριστικά:

- 4 κατηγοριακά χαρακτηριστικά
- 16 αριθμητικά (με συνεχής τιμές) χαρακτηριστικά

Στη συνέχεια , τα δεκαέξι (16) αριθμητικά χαρακτηριστικά θα εκφραστούν σε κατηγοριακή μορφή με τρεις (min, other, max) ή πέντε (very low, low, medium, high, very high) με βάση τη διακύμανση τιμών που παρατηρείται σε κάθε ένα από αυτά. Αν το σύνολο των εγγραφών (min+max) που ισούνται ή με την ελάχιστη τιμή για το αντίστοιχο χαρακτηριστικό ή με τη μέγιστη τιμή για το χαρακτηριστικό, είναι μεγαλύτερο από το 66.66 % θα χρησιμοποιηθούν τρεις τιμές για την περιγραφή τους (min, other, max), διαφορετικά πέντε πιθανές ισοζυγισμένες κατηγορίες τιμών (very low, low, medium, high, very high). Για τον προσδιορισμό των αντίστοιχων ποσοστών θα χρησιμοποιηθούν ερωτήματα της μορφής:

```
declare @sinolo as real

select @sinolo=count(*) from kddcup



declare @min as real

select @min=min(dst_host_srv_serror_rate) from kddcup



declare @max as real

select @max=max(dst_host_srv_serror_rate) from kddcup



select @min,100*count(*)/@sinolo

from kddcup

where dst_host_srv_serror_rate=@min



select @max, 100*count(*)/@sinolo

from kddcup

where dst_host_srv_serror_rate=@max
```

Πεδίο src_bytes (5)

5 τιμές (very low, low, medium, high, very high), min+max=23%

Πεδίο dst_bytes (6)

3 τιμές (min, other, max), min+max=83%

Πεδίο count (23)

5 τιμές (very low, low, medium, high, very high), min+max=46%

Πεδίο srv_count (24)

5 τιμές (very low, low, medium, high, very high), min+max=46%

Πεδίο serror_rate (25)

3 τιμές (min, other, max), min+max=99%

Πεδίο srv_serror_rate (26)

3 τιμές (min, other, max), min+max=99%

Πεδίο same_srv_rate (29)

3 τιμές (min, other, max), min+max=78%

Πεδίο diff_srv_rate (30)

3 τιμές (min, other, max), min+max=78%

Πεδίο dst_host_count (32)

3 τιμές (min, other, max), min+max=88%

Πεδίο dst_host_srv_count (33)

3 τιμές (min, other, max), min+max=68%

Πεδίο dst_host_same_srv_rate (34)

3 τιμές (min, other, max), min+max=73%

Πεδίο dst_host_diff_srv_rate (35)

3 τιμές (min, other, max), min+max=70%

Πεδίο dst_host_same_src_port_rate (36)

3 τιμές (min, other, max), min+max=87%

Πεδίο dst_host_srv_diff_host_rate (37)

3 τιμές (min, other, max), min+max=90%

Πεδίο dst_host_serror_rate (38)

3 τιμές (min, other, max), min+max=99%

Πεδίο dst_host_srv_serror_rate (39)

3 τιμές (min, other, max), min+max=99%

Εκ των οποίων, τα χαρακτηριστικά : 6, 25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39 θα εκφραστούν με τρεις (3) πιθανές κατηγοριακές τιμές (min, other, max) και τα χαρακτηριστικά 5, 23, 24 με πέντε (5) πιθανές κατηγοριακές τιμές (very low, low, medium, high, very high).

Για τον προσδιορισμό τον ορίων μεταξύ των πέντε κατηγοριακών εύρων τιμών, θα χρησιμοποιηθούν ερωτήματα της μορφής:

```
/*

https://msdn.microsoft.com/en-us/library/ms180169.aspx

*/

declare @sinolo as real

select @sinolo=count(*)
```

```
from kddcup


declare @min as real
select @min=min(src_bytes) from kddcup


declare @max as real
select @max=max(src_bytes) from kddcup


declare @split1 as real
declare @split2 as real
declare @split3 as real
declare @split4 as real
set @split1=0
set @split2=0
set @split3=0
set @split4=0


declare @count1 as real
declare @count2 as real
declare @count3 as real
declare @count4 as real
set @count1=0
set @count2=0
set @count3=0
```

```
set @count4=0


declare @timi as real

declare @pososto as real


DECLARE vendor_cursor CURSOR FOR

select distinct src_bytes , 100 * count(*)/@sinolo

from kddcup

group by src_bytes

order by src_bytes


OPEN vendor_cursor


FETCH NEXT FROM vendor_cursor INTO @timi, @pososto


WHILE @@FETCH_STATUS = 0

BEGIN


set @count1=@count1+@pososto

set @count2=@count2+@pososto

set @count3=@count3+@pososto

set @count4=@count4+@pososto
```

```
if (@count4>=80 and @split3>0 and @split4=0)

begin

set @split4=@timi

print @timi

end


if (@count3>=60 and @split2>0 and @split3=0)

begin

set @split3=@timi

print @timi

end


if (@count2>=40 and @split1>0 and @split2=0)

begin

set @split2=@timi

print @timi

end


if (@count1>=20 and @split1=0)

begin

set @split1=@timi

print @timi

end
```

```
FETCH NEXT FROM vendor_cursor INTO @timi, @pososto

end


CLOSE vendor_cursor;

DEALLOCATE vendor_cursor;


select @min,@split1,@split2,@split3,@split4,@max
```

Με την χρήση ερωτημάτων της παραπάνω μορφής, διαπιστώνονται τα όρια τιμών μεταξύ των κατηγορικών τιμών ως εξής:

Πεδίο src_bytes (5)

- very low [0-1]
- low (1- 481]
- medium (481-1032]
- high (1032-1033]
- very high(1033 - 6,933756E+08]

Για τα δυο τελευταία πεδία (count, srv_count), διαπιστώνεται πως μεγάλο ποσοστό των εγγραφών (>45%) αντιστοιχεί στη μέγιστη τιμή που παρουσιάζουν οι τιμές για τα αντίστοιχα πεδία και συνεπώς, ταξινομούνται σε δυο πιθανές κατηγοριακές τιμές:

Πεδίο count (23)

- other [0- 511)
- max [511]

Πεδίο srv_count (24)

- other [0- 511)

- max [511]

Τέλος, για να υπάρχει ομοιομορφία στην έκφραση των κατηγοριακών **src_bytes** και **dst_bytes,** θα εκφράσουμε και το 2ο πεδίο σε κατηγοριακή μορφή με πέντε πιθανά εύρη τιμών. Εκτελώντας το παραπάνω ερώτημα SQL για το πεδίο dst_bytes, προκύπτουν τα εξής εύρη τιμών:

Πεδίο dst_bytes (6)

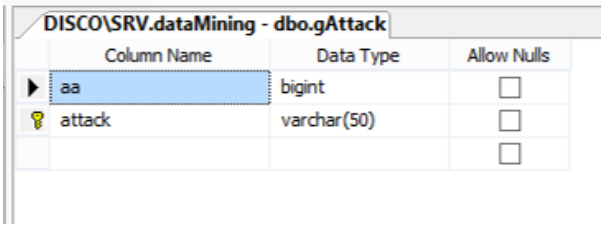
- very low [0-1]
- low (1-4]
- medium (4-5]
- high (5-6]
- very high(6 - 5155468]

Μετά τα παραπάνω έχουν προσδιοριστεί τα χαρακτηριστικά (20 συνολικά) που θα χρησιμοποιηθούν για την υλοποίηση αποθήκης δεδομένων, καθώς επίσης και τα εύρη τιμών καθενός από εκείνα.

6.2 Υλοποίηση Κανονικοποιημένης Βάσης Δεδομένων

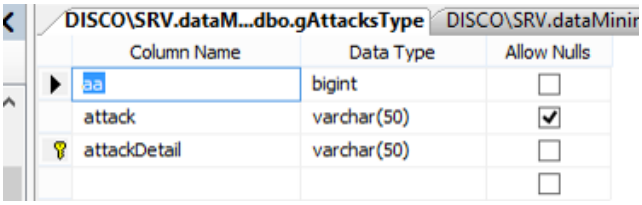
Μετά όσων αναφέρθηκαν στην προηγούμενη ενότητα, απαιτείται μια ριζική επανασχεδίαση της βάσης δεδομένων, προκειμένου να περιλαμβάνει παραπάνω πίνακες (όχι όλες τις πληροφορίες σε έναν πίνακα) και η διασύνδεση των πινάκων με σχέσεις, προκειμένου στη συνέχεια να υλοποιηθεί σχετική αποθήκη δεδομένων[12], [13],[14].

Καταρχάς , θα πρέπει να προστεθούν ένας πίνακας (με όνομα gAttack) με ένα μόνο πεδίο (ως κύριο κλειδί) που θα περιλαμβάνει τη κατηγορία επίθεσης (dos, probe, u2r, u2l, normal) ως εξής:



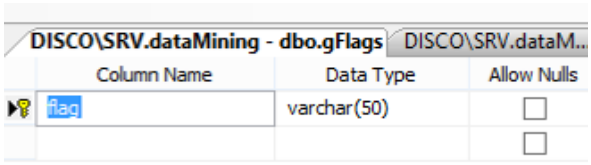
Εικόνα 15 Σχεδίαση πίνακα gAttack

Καθώς και ένας πίνακα (με όνομα gAttacksType) με τα 22 είδη επιθέσεων (πεδίο attackDetail, ως κύριο κλειδί) και δευτερεύον κλειδί το πεδίο attack που θα συνδέεται με τον πίνακα gAttack (κατηγορίες επιθέσεων):



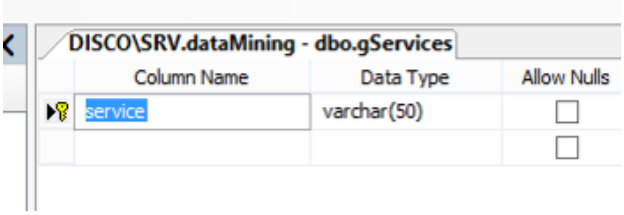
Εικόνα 16: Σχεδίαση πίνακα gAttacksType

Στη συνέχεια , θα πρέπει να δημιουργηθούν τέσσερις (4) πίνακες με τις πιθανές τιμές των τεσσάρων κατηγορικών μεταβλητών που θα χρησιμοποιηθούν για την κατασκευή της αποθήκης δεδομένων. Πιο συγκεκριμένα θα προστεθεί πίνακας με όνομα *gFlags*, που θα περιλαμβάνει τις πιθανές τιμές του πεδίου *flag* στον πίνακα *kddcup*:



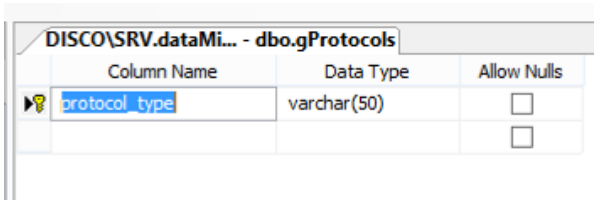
Εικόνα 17: Σχεδίαση πίνακα gFlags

Επίσης, θα προστεθεί πίνακας με όνομα *gServices*, που θα περιλαμβάνει τις πιθανές τιμές του πεδίου *service* στον πίνακα *kddcup*:



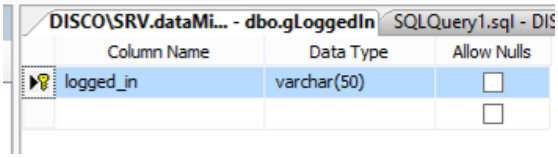
Εικόνα 18: Σχεδίαση πίνακα gServices

Ακόμα, θα πρέπει προστεθεί πίνακας με όνομα *gProtocols*, που θα περιλαμβάνει τις πιθανές τιμές του πεδίου *protocol_type* στον πίνακα *kddcup*:



Εικόνα 19: Σχεδίαση πίνακα gProtocols

Τέλος, θα πρέπει προστεθεί πίνακας με όνομα *gLoggedIn*, που θα περιλαμβάνει τις πιθανές τιμές του πεδίου *logged_in* στον πίνακα *kddcup*:



Εικόνα 20: Σχεδίαση πίνακα gLoggedIn

Η ενημέρωση των πινάκων που δημιουργήθηκαν με τις απαραίτητες τιμές μπορεί να γίνει εύκολα, με την εκτέλεση ερωτημάτων της μορφής:

```
insert into gLoggedIn (logged_in)

        select distinct logged_in from kddcup
```

Κατόπιν, θα πρέπει αρχικά να ενημερωθούν οι τιμές των δυο χαρακτηριστικών (23, 24) που θα εκφραστούν με δυο κατηγορικές τιμές (max, other). Αρχικά θα πρέπει να αλλάξει ο τύπος των αντίστοιχων πεδίων από αριθμητικά σε αλφαριθμητικά (varchar), εκτελώντας ερωτήματα της μορφής:

```
declare @max as real

select @max=MAX(COUNT) from kddcup
```

```
ALTER TABLE kddcup ALTER COLUMN count varchar(50)



update kddcup set count='other' where count<>CAST(@max as varchar(50))
```

Στη συνέχεια θα πρέπει να ενημερωθούν οι τιμές για τα χαρακτηριστικά (25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39), που θα εκφραστούν με τρείς (3) κατηγορικές τιμές (min, other, max). Η αντίστοιχη μετατροπή θα γίνει με την εκτέλεση ερωτημάτων της μορφής:

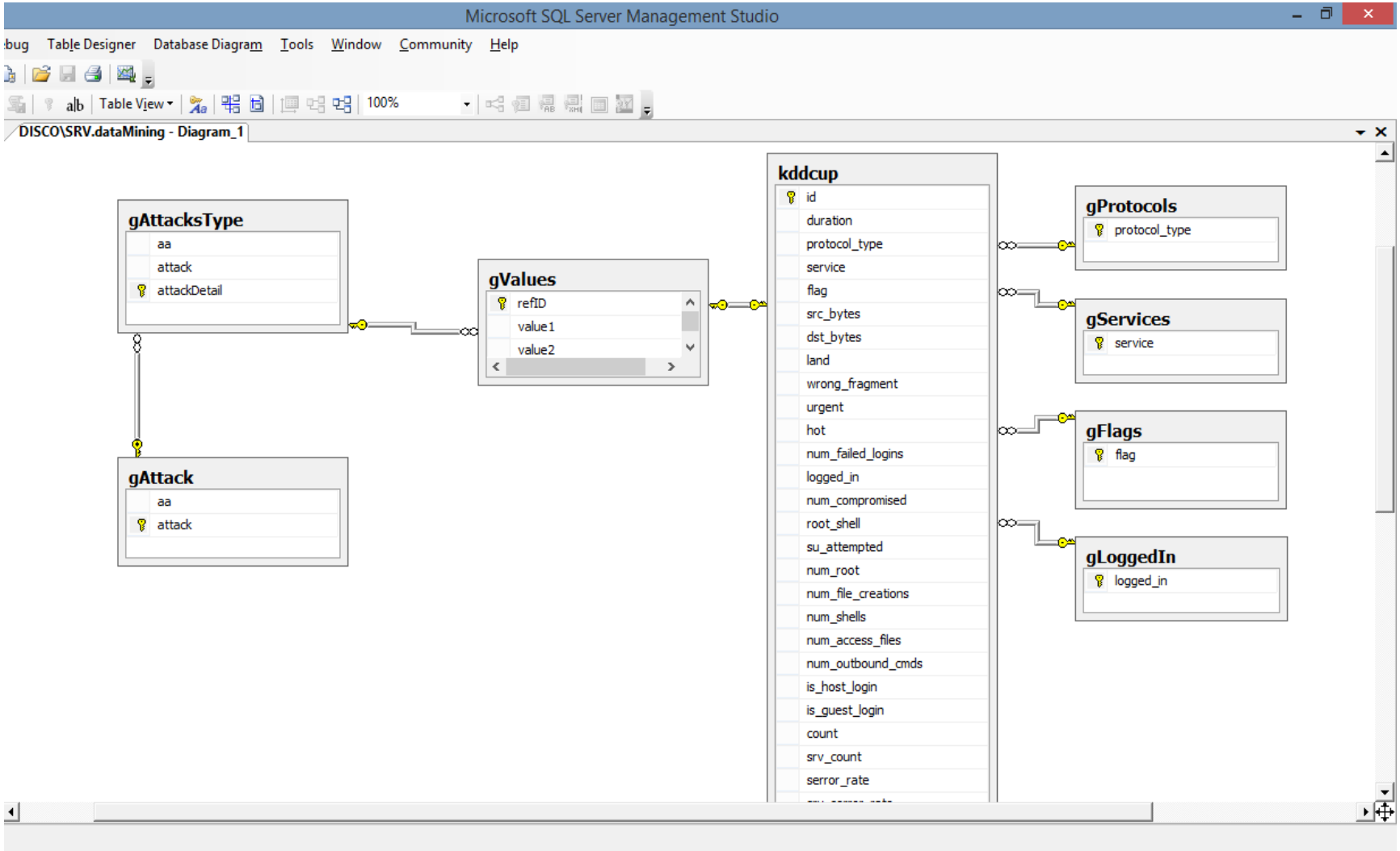
```
declare @max as real

select @max=MAX(srv_count) from kddcup



declare @min as real

select @min=MIN(srv_count) from kddcup



ALTER TABLE kddcup ALTER COLUMN srv_count varchar(50)



update kddcup set srv_count='other'

where srv_count<>CAST(@max as varchar(50)) and srv_count<>CAST(@min as
varchar(50))
```

Τέλος, θα πρέπει να ενημερωθούν και οι κατηγορικές τιμές για τα χαρακτηριστικά 5, 6 που θα αντιστοιχούν σε πέντε πιθανές κατηγορικές τιμές (very low, low, medium, high, very high). Η μετατροπή των τιμών μπορεί να γίνει με την εκτέλεση ερωτημάτων της μορφής:

```
declare @border1 as real

declare @border2 as real

declare @border3 as real
```

```
declare @border4 as real


set @border1=1

set @border2=4

set @border3=5

set @border4=6


update kddcup set dst_bytes=0 where dst_bytes <=@border1

update kddcup set dst_bytes=1 where dst_bytes >@border1 and dst_bytes <=@border2

update kddcup set dst_bytes=2 where dst_bytes >@border2 and dst_bytes <=@border3

update kddcup set dst_bytes=3 where dst_bytes >@border3 and dst_bytes <=@border4

update kddcup set dst_bytes=4 where dst_bytes >@border4


ALTER TABLE kddcup ALTER COLUMN dst_bytes varchar(50)


update kddcup set dst_bytes='very low' where dst_bytes='0'

update kddcup set dst_bytes='low' where dst_bytes='1'

update kddcup set dst_bytes='medium' where dst_bytes='2'

update kddcup set dst_bytes='high' where dst_bytes='3'

update kddcup set dst_bytes='very high' where dst_bytes='4'
```

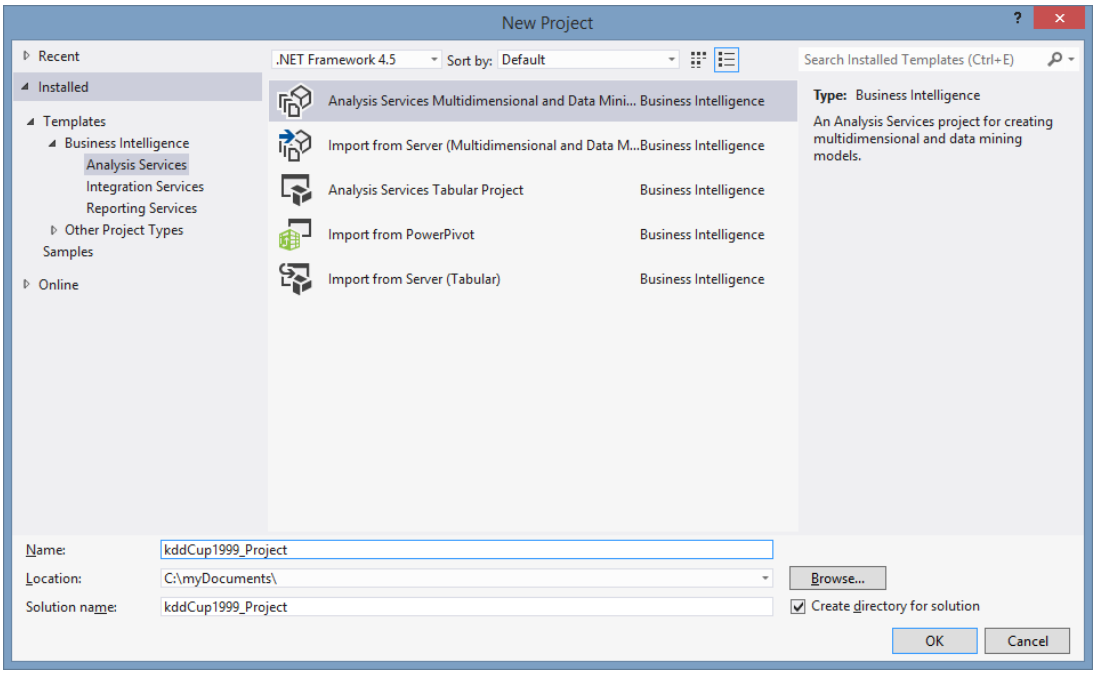
Έχοντας μετατρέψει όλα τα αριθμητικά χαρακτηριστικά που θα χρησιμοποιηθούν για την υλοποίηση αποθήκης δεδομένων σε κατηγοριακά, στη συνέχεια θα οριστούν οι απαραίτητες συσχετίσεις μεταξύ των πινάκων που δημιουργήθηκαν:


Microsoft SQL Server Management Studio
bug Table Designer Database Diagram Tools Window Community Help
Table View
100%
DISCO\SRV.dataMining - Diagram_1
gAttacksType
aa
attack
attackDetail
gValues
refID
value1
value2
gAttack
aa
attack
kddcup
id
duration
protocol_type
service
flag
src_bytes
dst_bytes
land
wrong_fragment
urgent
hot
num_failed_logins
logged_in
num_compromised
root_shell
su_attempted
num_root
num_file_creations
num_shells
num_access_files
num_outbound_cmds
is_host_login
is_guest_login
count
srv_count
serror_rate
gProtocols
protocol_type
gServices
service
gFlags
flag
gLoggedIn
logged_in

Εικόνα 21: Συσχετίσεις μεταξύ των πινάκων στη βάση dataMining

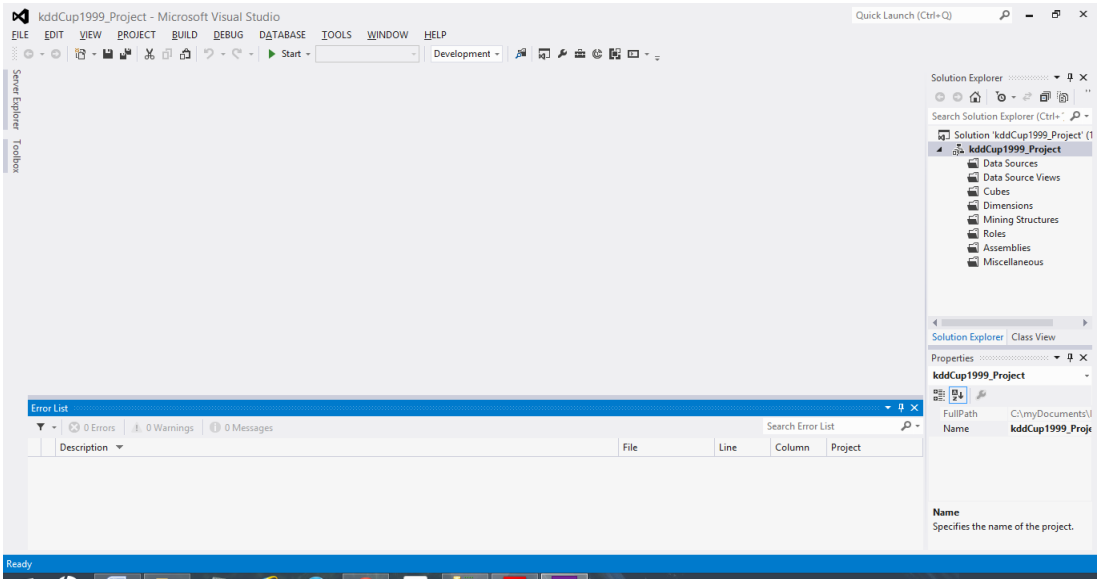
6.3 Υλοποίηση αποθήκης δεδομένων

Αρχικά φορτώνουμε την εφαρμογή Visual Studio 2012 και επιλέγουμε την δημιουργία καινούργιο Project, με το όνομα *kddCup1999_Project* με βάση το template: Analysis Services *Multidimensional and Data Mining Project* (21):



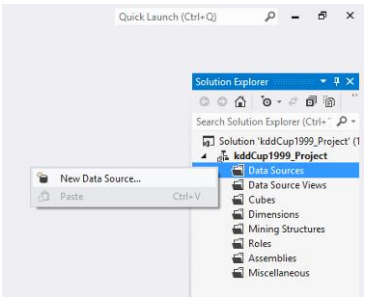
Εικόνα 22: Δημιουργία καινούργιου Analysis Services Multidimensional project στο Visual Studio 2012

Πατώντας το κουμπί OK, δημιουργείται ένα κενό *project* της επιλεγμένης μορφής:



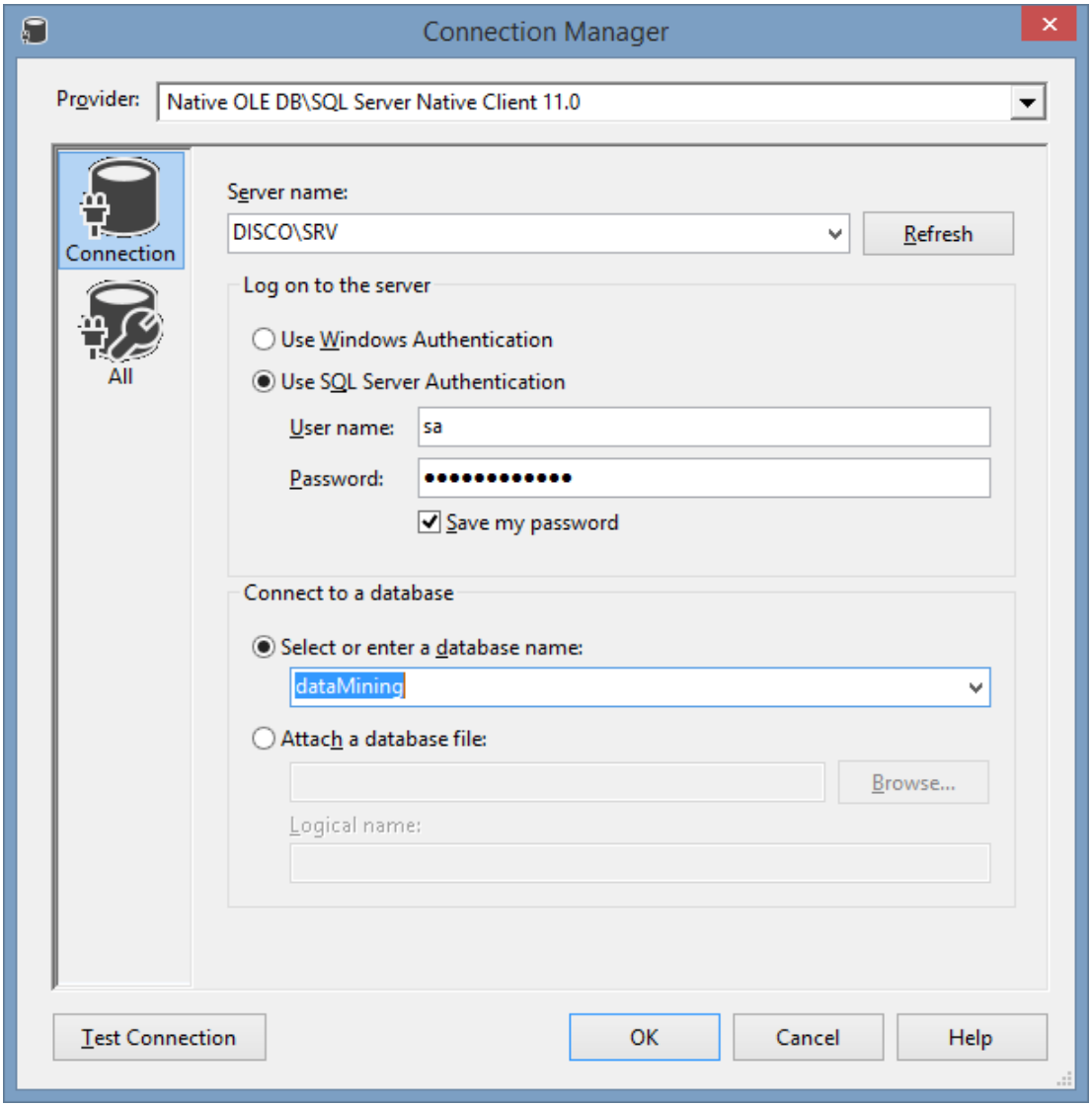
Εικόνα 23: Κενό Analysis Services Multidimensional and Data Mining project

Με πρώτο μετέπειτα βήμα, τη δημιουργία ενός καινούργιου *Data Source*:



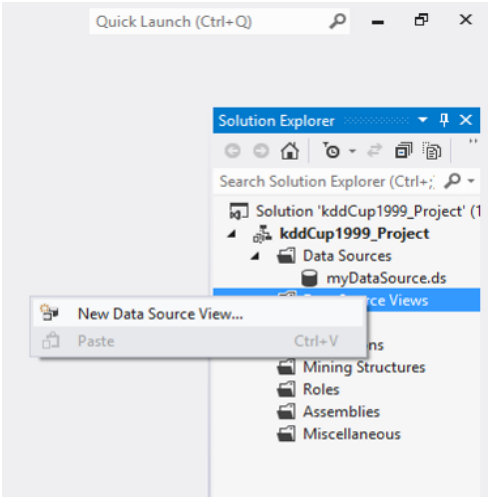
Εικόνα 24: Επιλογή δημιουργίας καινούργιου Data Source

Όπου επιλέγεται η δημιουργία *Data Source*, για τη βάση που μόλις δημιουργήσαμε με όνομα *dataMining* στον τοπικό server: *DISCO/SRV*:



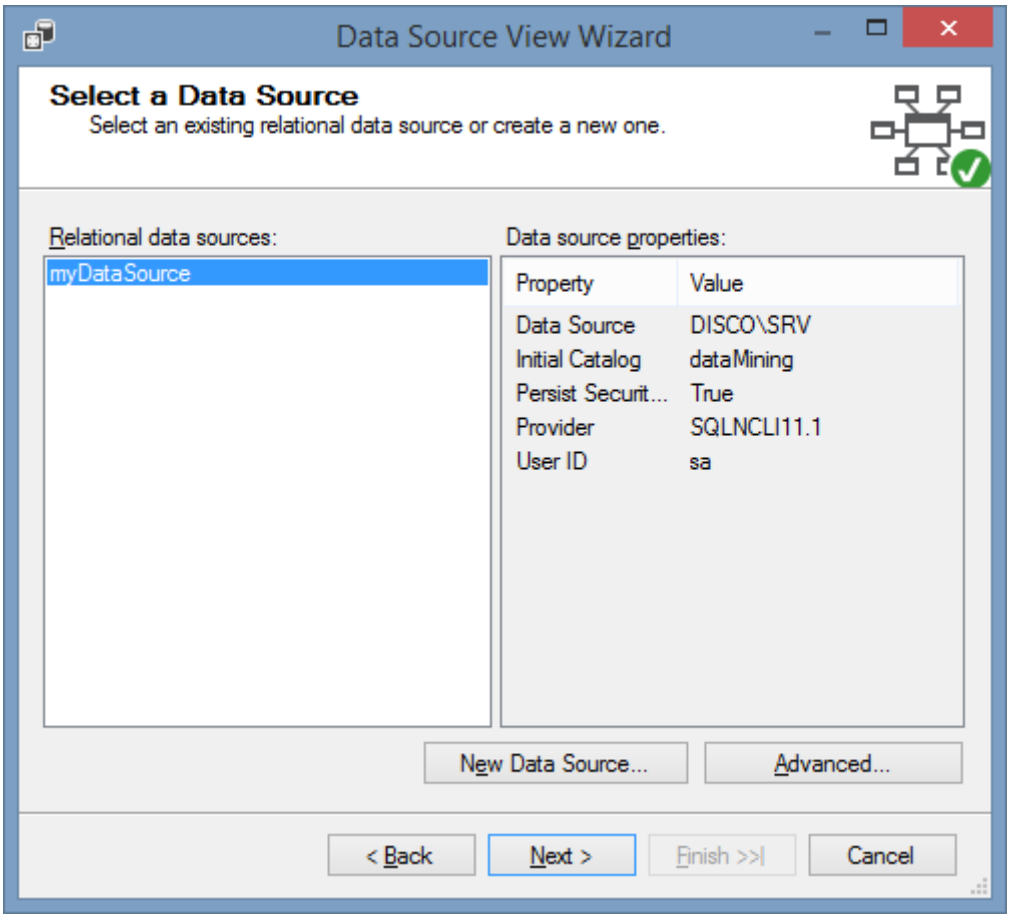
Εικόνα 25: Επιλογή βάσης δεδομένων για το Data Source

Το οποίο ονομάζουμε ως *myDataSource.* Στη συνέχεια επιλέγουμε να δημιουργούμε ένα καινούργιο *Data Source View*:



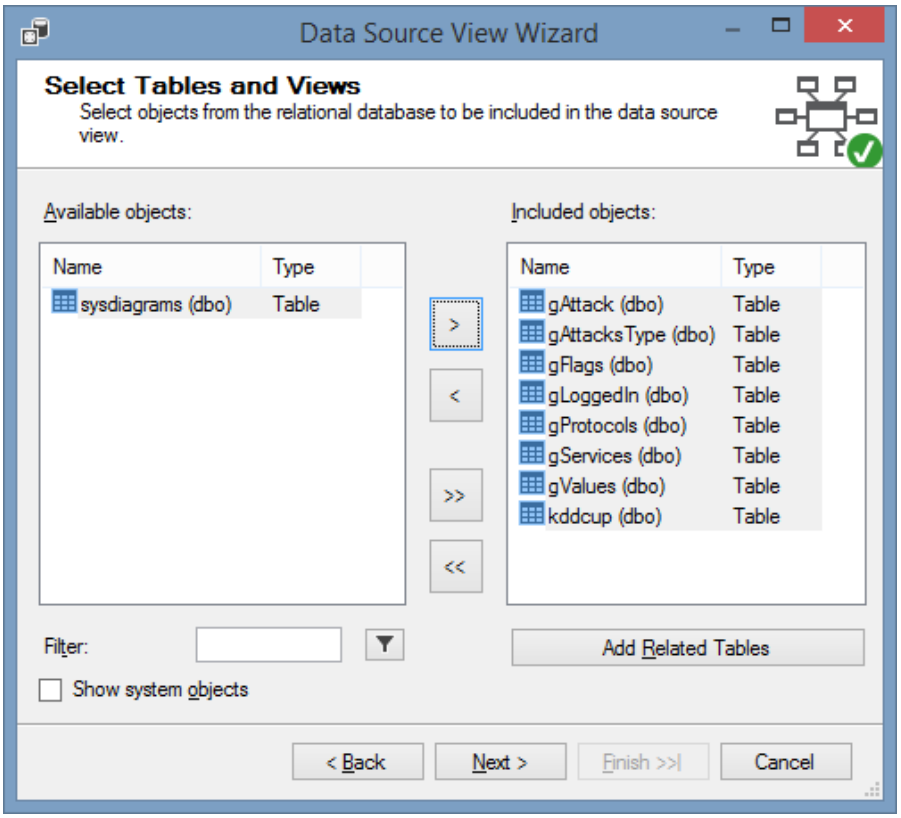
Εικόνα 26: Επιλογή δημιουργίας καινούργιου Data Source View

Όπου αρχικά επιλέγουμε να χρησιμοποιήσουμε το *Data Source* που μόλις δημιουργήσαμε:



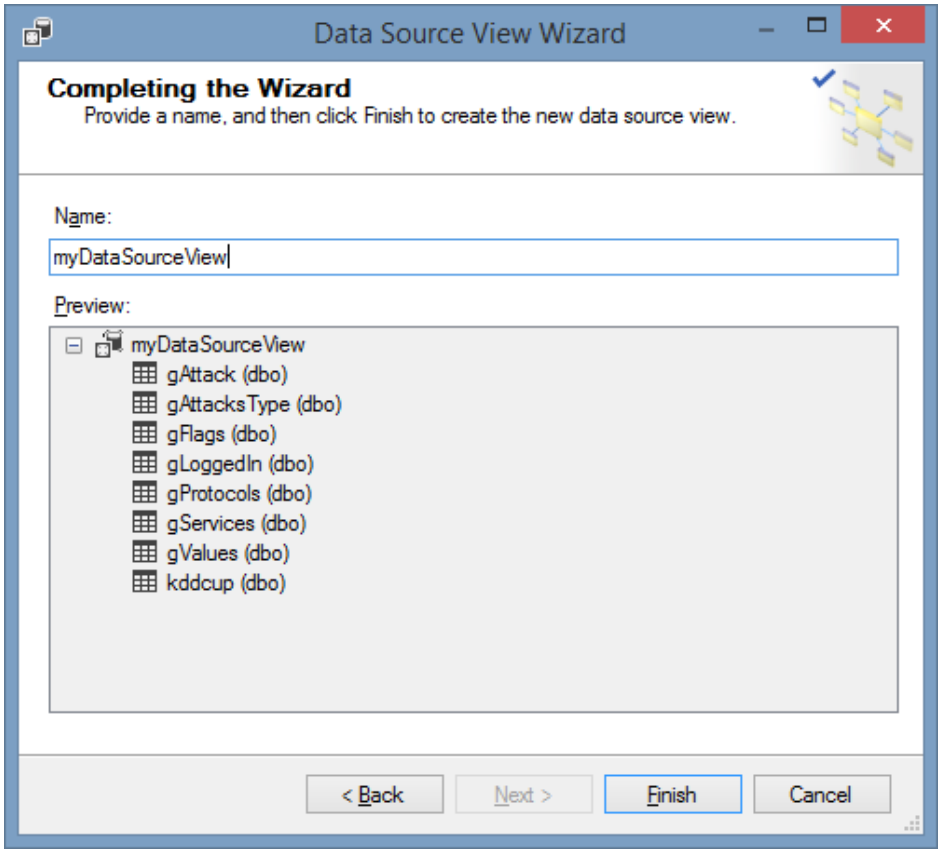
Εικόνα 27: Επιλογή data source για το καινούργιο data source view

Κατόπιν, επιλέγουμε να συμπεριληφθούν στο data source view όλοι οι πίνακες, εκτός από τον πίνακα *sysdiagrams*:



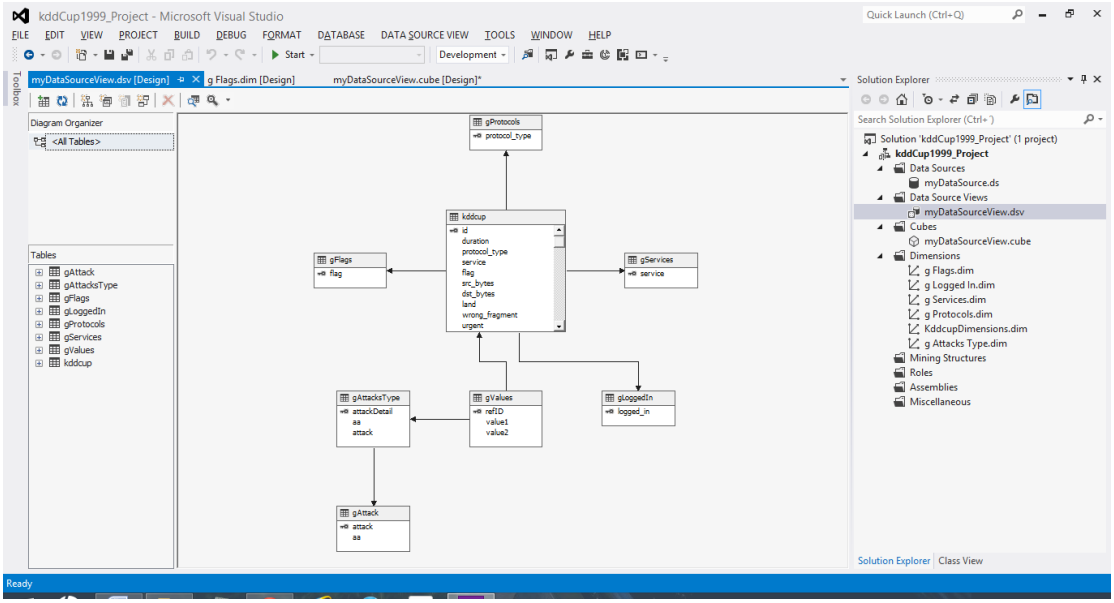
Εικόνα 28: Επιλογή πινάκων που θα συμπεριληφθούν στο data source view

Τέλος, αποθηκεύουμε το data source view με το όνομα: *myDataSourceView*



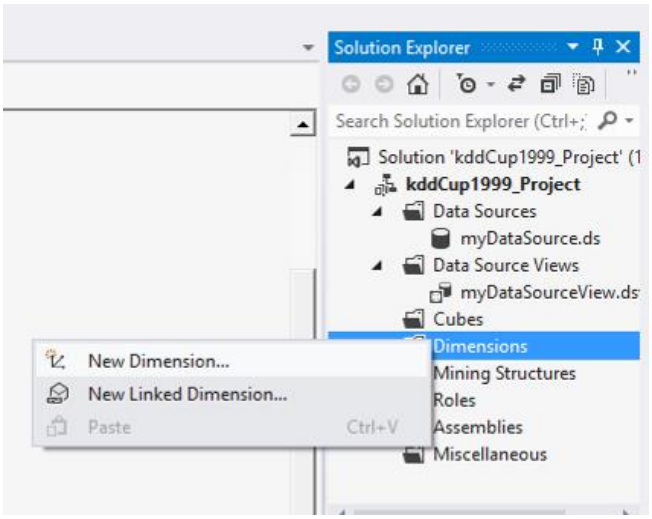
Εικόνα 29: Ορισμό ονόματος νέου data source view

Και στη συνέχεια εμφανίζονται οι συσχετίσεις πινάκων για το καινούργιο *data source view*:



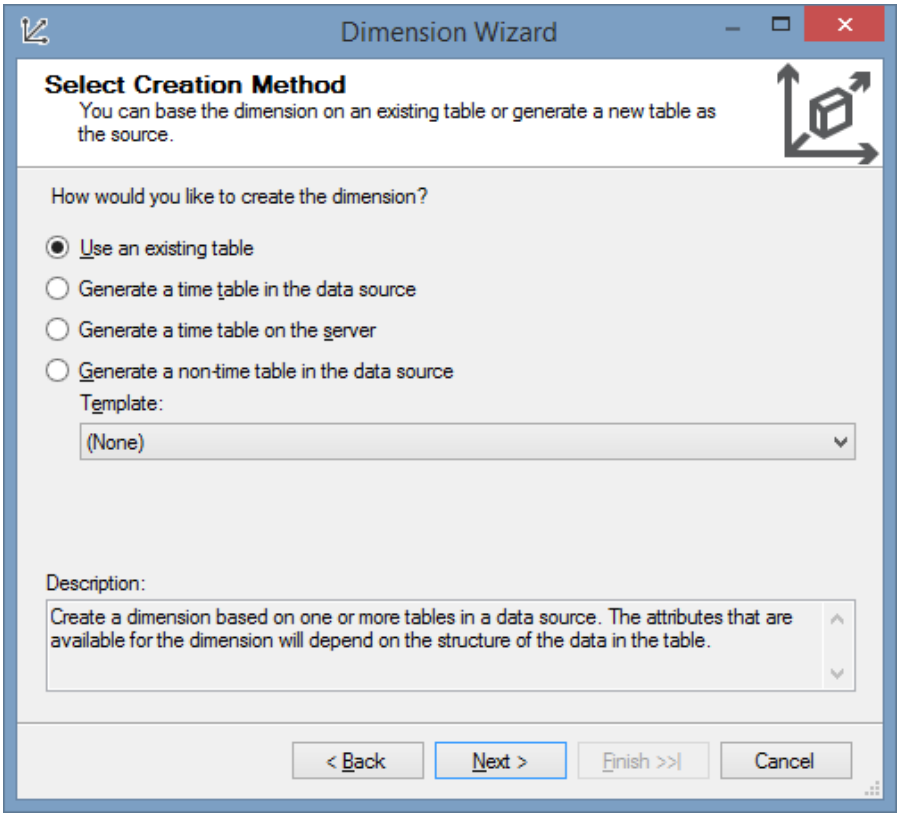
Εικόνα 30: Εμφάνιση πληροφοριών για το data source view: myDataSourceView

Στη συνέχεια θα οριστούν οι διαστάσεις (*dimensions*) της αποθήκης δεδομένων:



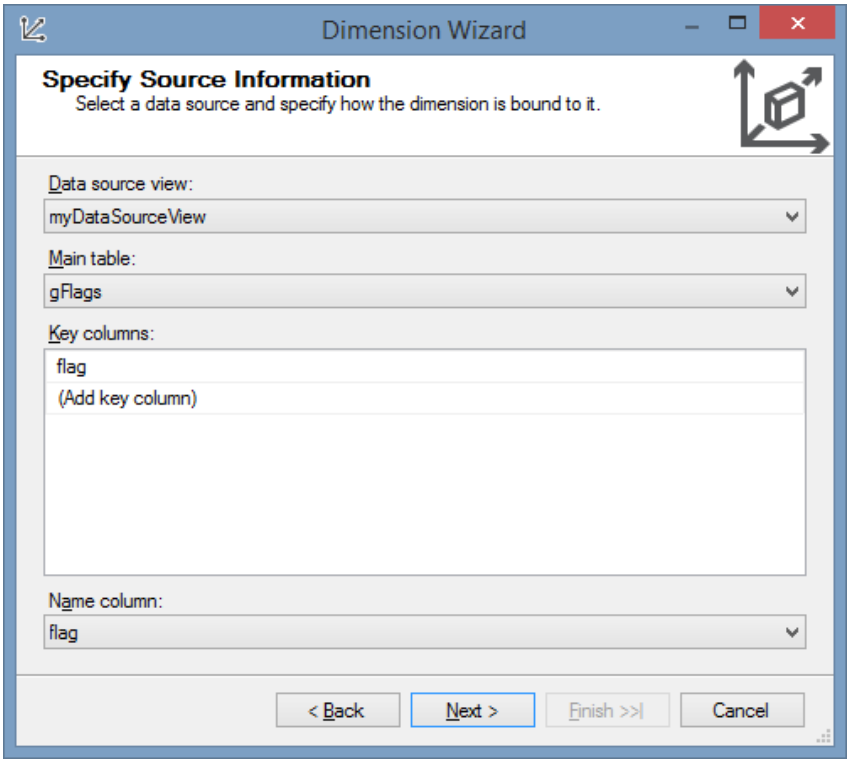
Εικόνα 31: Επιλογή δημιουργίας καινούργιας διάστασης (dimension)

Για τα χαρακτηριστικά (*protocol_type, service, flag, logged_in*) του συνόλου δεδομένων που αντιστοιχούν αρχικά σε κατηγοριακές τιμές, επιλέγεται η δημιουργία με χρήση υπάρχοντα πίνακα στη βάση δεδομένων:



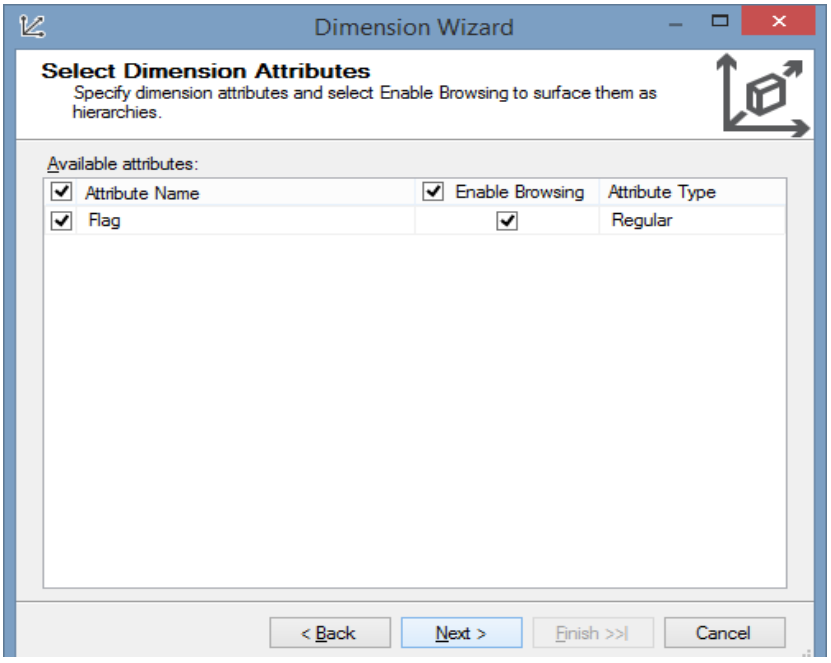
Εικόνα 32: Επιλογή χρήσης υπάρχοντος πίνακα για τον ορισμό διάστασης σε αποθήκη δεδομένων

Και στην επόμενη φόρμα, επιλέγεται το αντίστοιχο πεδίο που θα χρησιμοποιηθεί (η εικόνα αναφέρεται στο πεδίο *flag* του πινάκα *gFlags*):



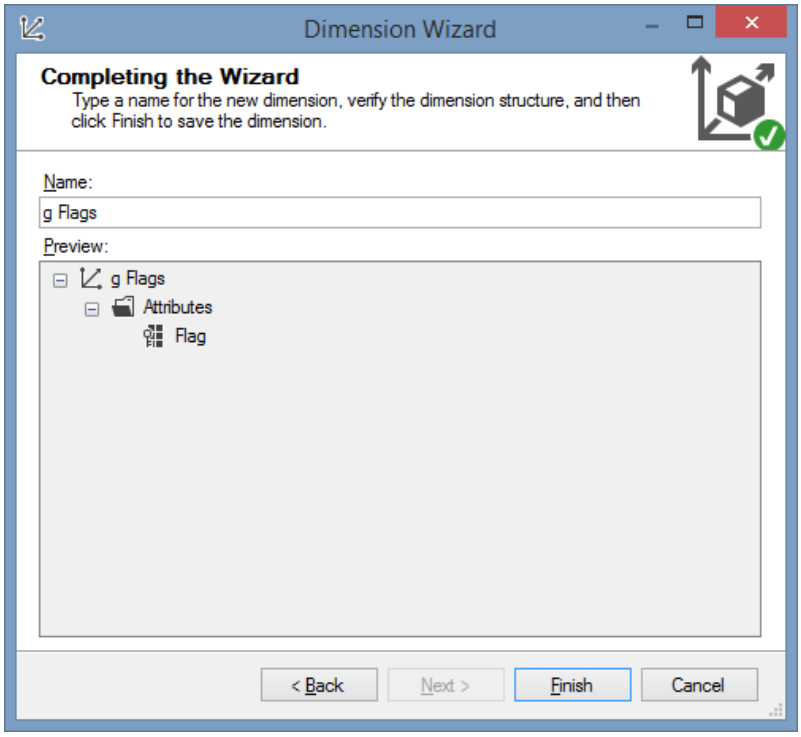
Εικόνα 33: Καθορισμό πίνακα - πεδίου διάστασης αποθήκης δεδομένων

Κατόπιν, δηλώνεται να μπορεί να χρησιμοποιηθεί το επιλεγμένο πεδίο για προσπέλαση δεδομένων στην αποθήκη δεδομένων που θα δημιουργηθεί:



Εικόνα 34: Δήλωσε πεδίων χρήσης για την προσπέλαση δεδομένων στην αποθήκη δεδομένων

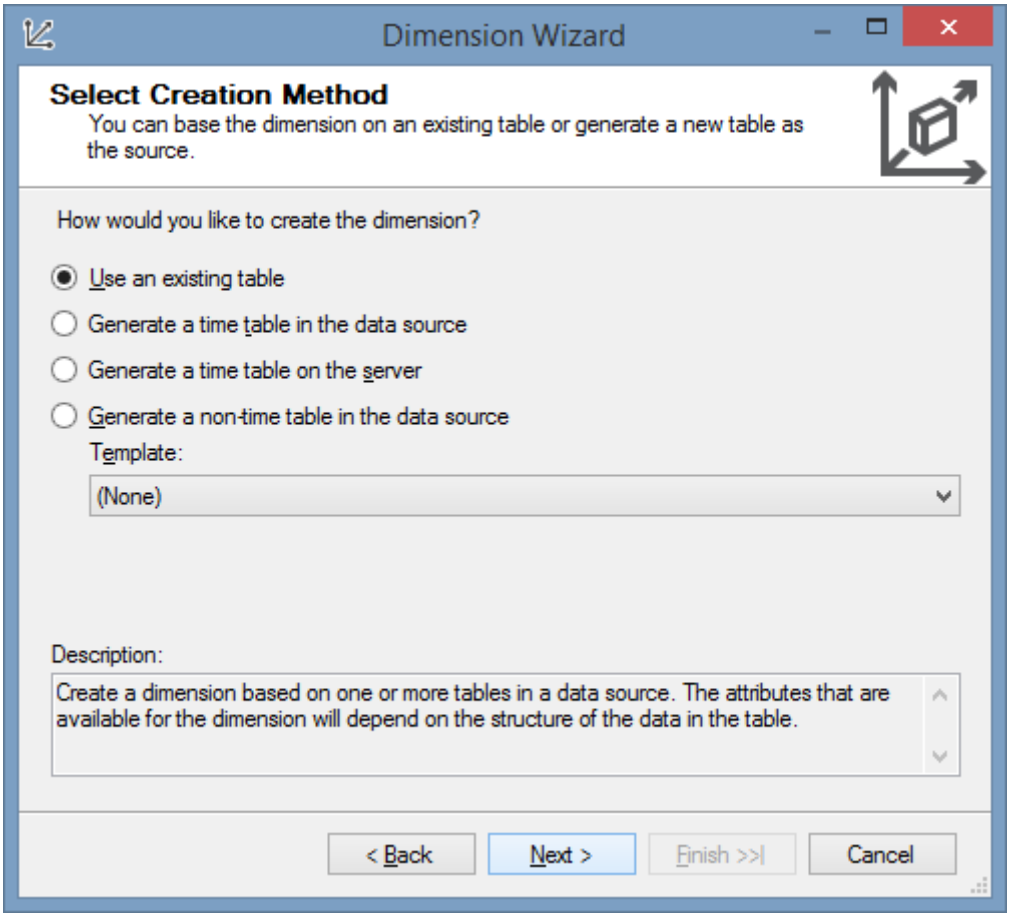
Στη συνέχεια, επιλέγουμε να ονοματίσουμε την καινούργια διάσταση με το προτεινόμενο όνομα και πατάμε το κουμπί *Finish*, για την οριστική αποθήκευση της νέας διάστασης:



Εικόνα 35: Οριστικοποίηση αποθήκευσης νέας διάστασης

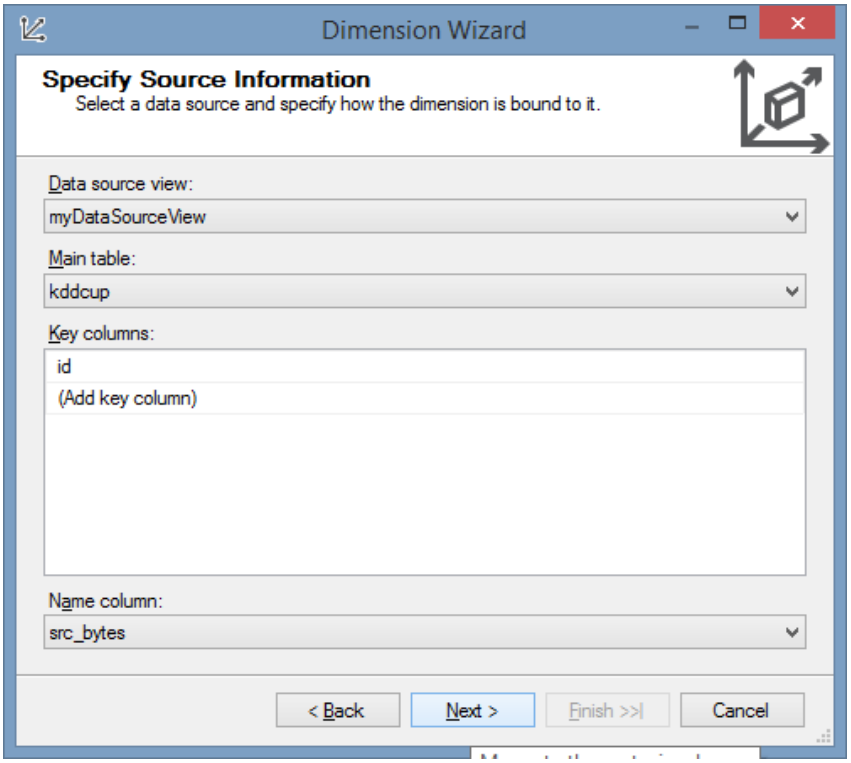
Με τον ίδιο ακριβώς τρόπο θα δημιουργηθούν οι διαστάσεις (*dimensions*) για τα πεδία *protocol_type*, *service*, *logged_id*.

Κατόπιν θα οριστούν οι διαστάσεις για τα 14 επιπλέον χαρακτηριστικά (5, 6, 23, 24, 25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39) του συνόλου δεδομένων που μετατράπηκε ο τύπος τους από αριθμητικά σε κατηγοριακά. Αρχικά για το πεδίο, επιλέγεται να δημιουργηθεί χρησιμοποιηθεί ο πίνακας *kddcup* της βάσης δεδομένων:



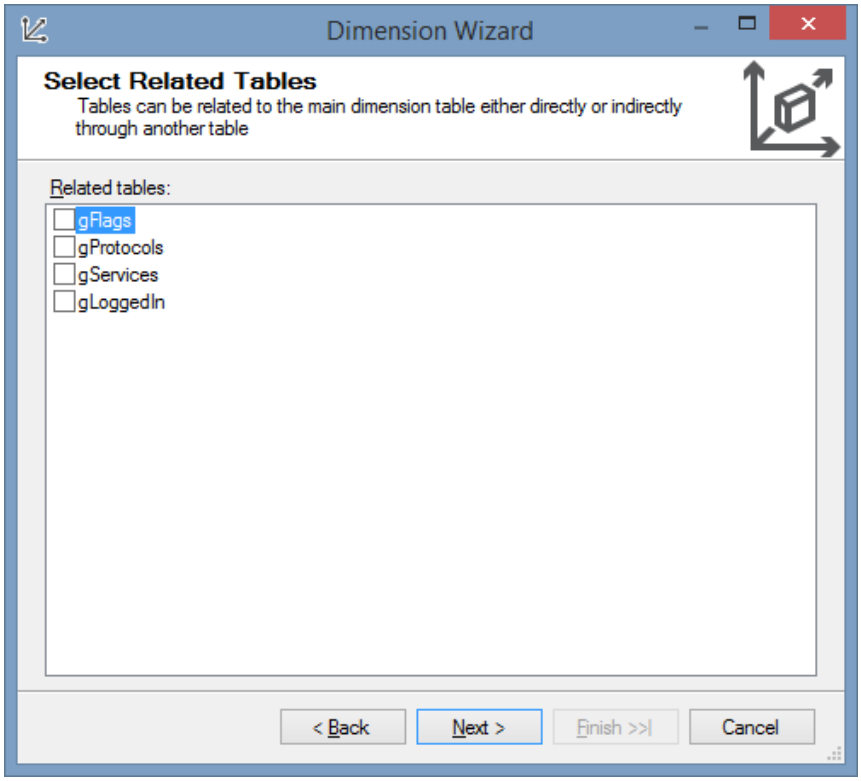
Εικόνα 36: Επιλογή χρήσης υπάρχοντα πίνακα για την προσθήκη διάστασης αποθήκης δεδομένων

Στη συνέχεια δηλώνεται να χρησιμοποιηθεί το πεδίο κύριου κλειδιού (*id*) στον πίνακα *kddcup*:



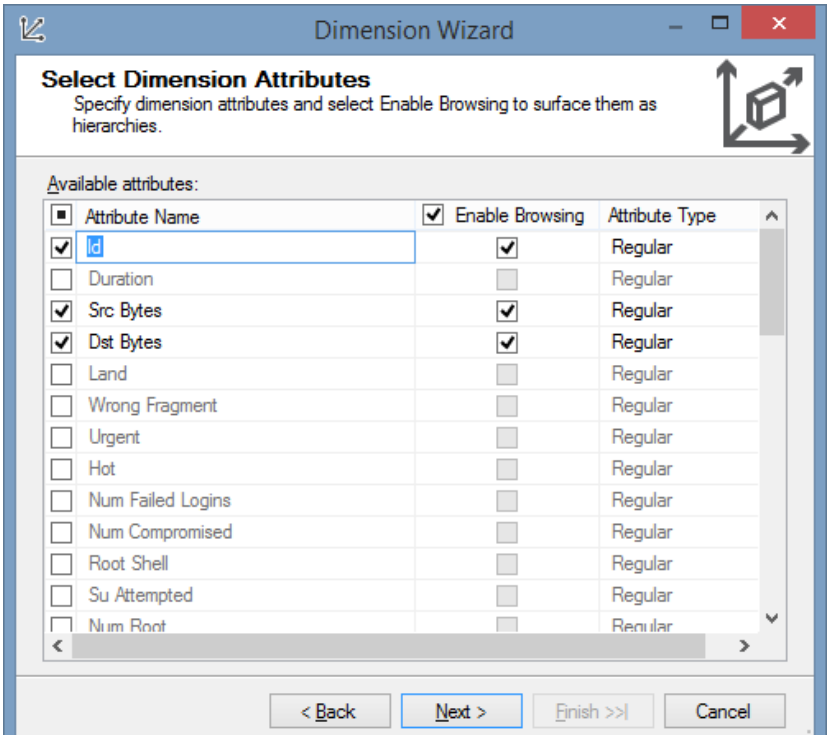
Εικόνα 37: Επιλογή πεδίου αντιστοίχισης σε διάσταση αποθήκης δεδομένων

Δεν δηλώνονται κάποιες συσχετίσεις με τους συνδεόμενους πίνακες με τον πίνακα *kddcup*, καθώς, θεωρείται πως δεν υπάρχει σχετική συσχέτιση δομή μεταξύ των σχετικών πεδίων:



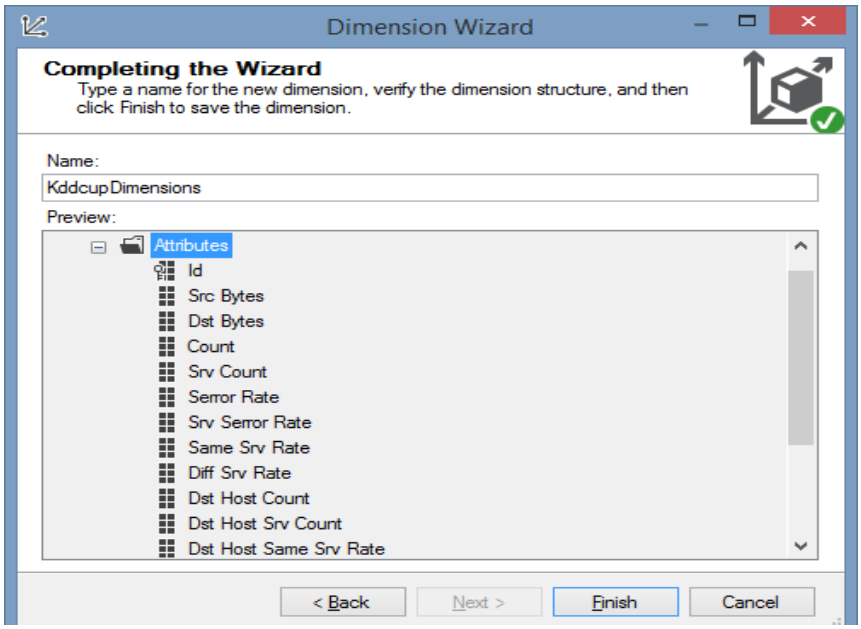
Εικόνα 38: Δήλωση συνδέσμων διάστασης με άλλους πίνακες

Και δηλώνονται να χρησιμοποιηθούν ως διαστάσεις τα επιπλέον δεκαέξι (16) χαρακτηριστικά 5, 6, 23, 24, 25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39 που θα χρησιμοποιηθούν ως διαστάσεις της αποθήκης δεδομένων:



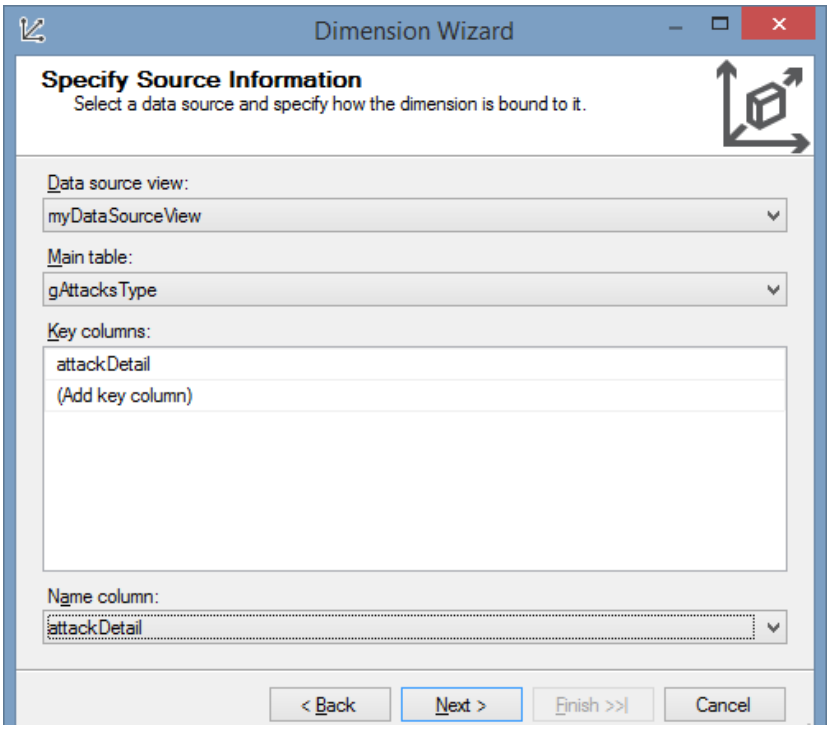
Εικόνα 39: Επιλογή χαρακτηριστικών πίνακα που θα χρησιμοποιηθούν ως διαστάσεις της αποθήκης δεδομένων

Την αντίστοιχη ομάδα χαρακτηριστικών διάστασης την ονομάζουμε *KddcupDimensions* και πατάμε το κουμπί *Finish* για την οριστική της αποθήκευση:



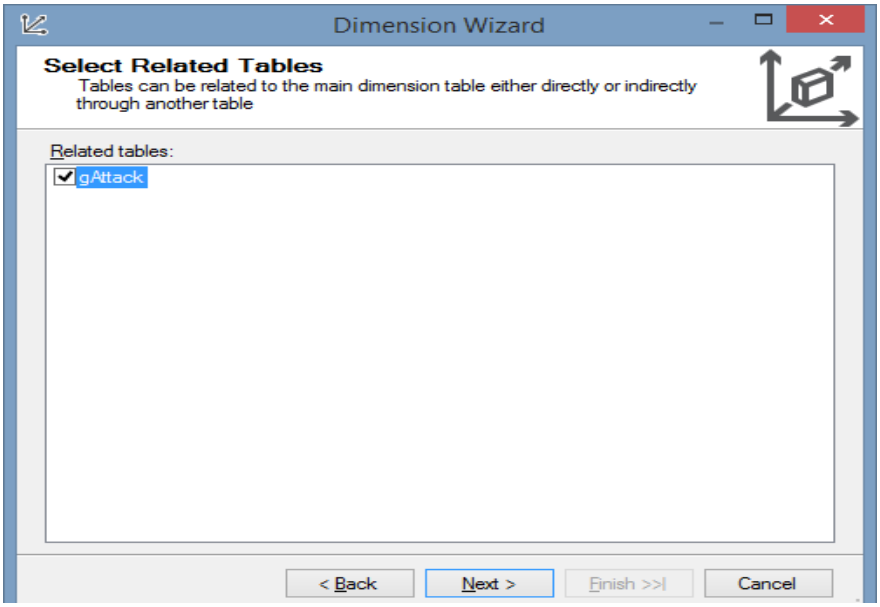
Εικόνα 40: Αποθήκευση ομάδας χαρακτηριστικών διάστασης αποθήκης δεδομένων

Τέλος, θα προστεθεί μια επιπλέον διάσταση σχετικά με την κατηγορία της επίθεσης, στην οποία και θα δημιουργηθεί ιεραρχία της μορφής Κατηγορία επίθεσης -> Τίτλος μορφής επίθεσης. Αρχικά ορίζουμε πως θα χρησιμοποιηθεί υπάρχοντας πίνακας (*gAttacksType*) της βάσης δεδομένων *dataMining*:



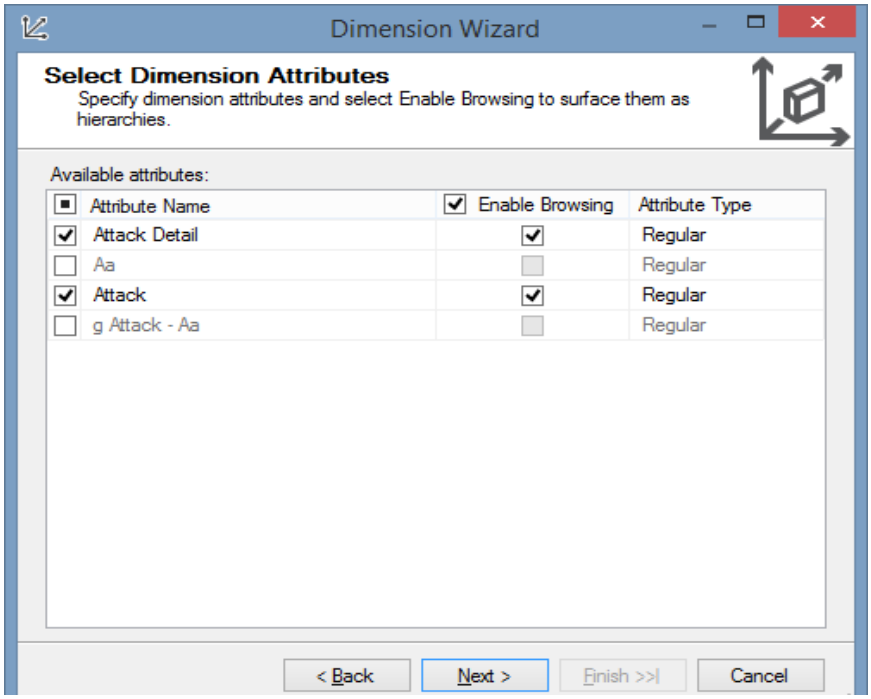
Εικόνα 41: Επιλογή πίνακα - πεδίου για τον ορισμό διάστασης σε αποθήκη δεδομένων

Δηλώνουμε πως θα χρησιμοποιηθεί η συσχέτιση με τον πίνακα *gAttack*:



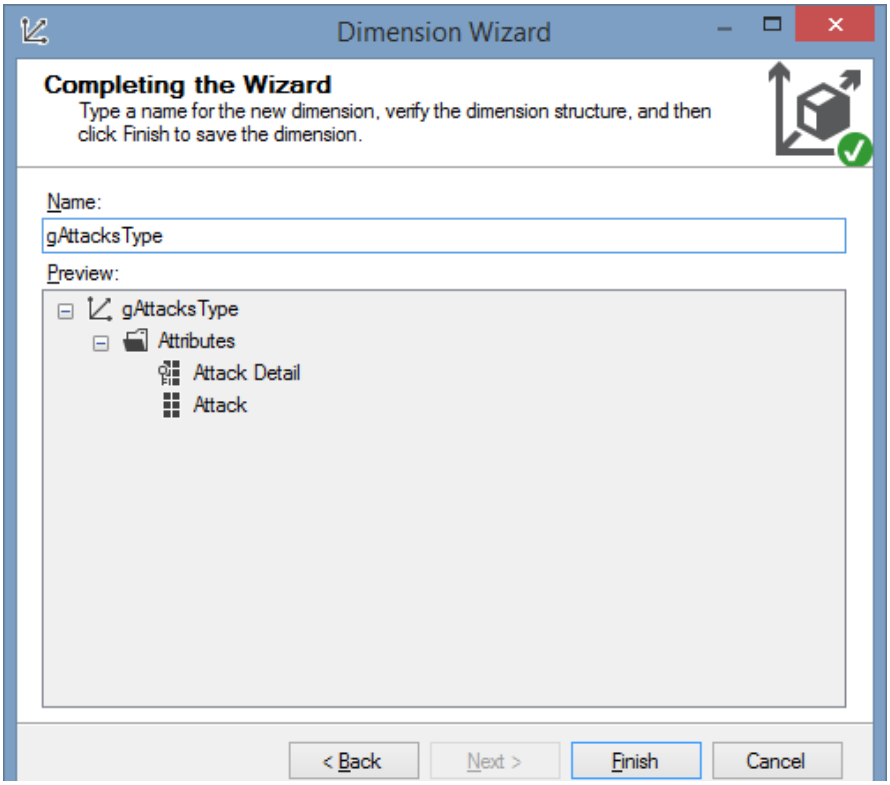
Εικόνα 42: Δήλωση συσχετιζόμενων πινάκων

Καθώς επίσης, πως για την προσπέλαση τιμών θα χρησιμοποιηθούν τα πεδία attack και *attackDetail*. Το πεδίο *attackDetail* αποτελεί το κύριο κλειδί του πίνακα *gAttacksType*, ενώ το πεδίο attack δευτερεύων κλειδί σύνδεσης με τον πίνακα *gAttack*:



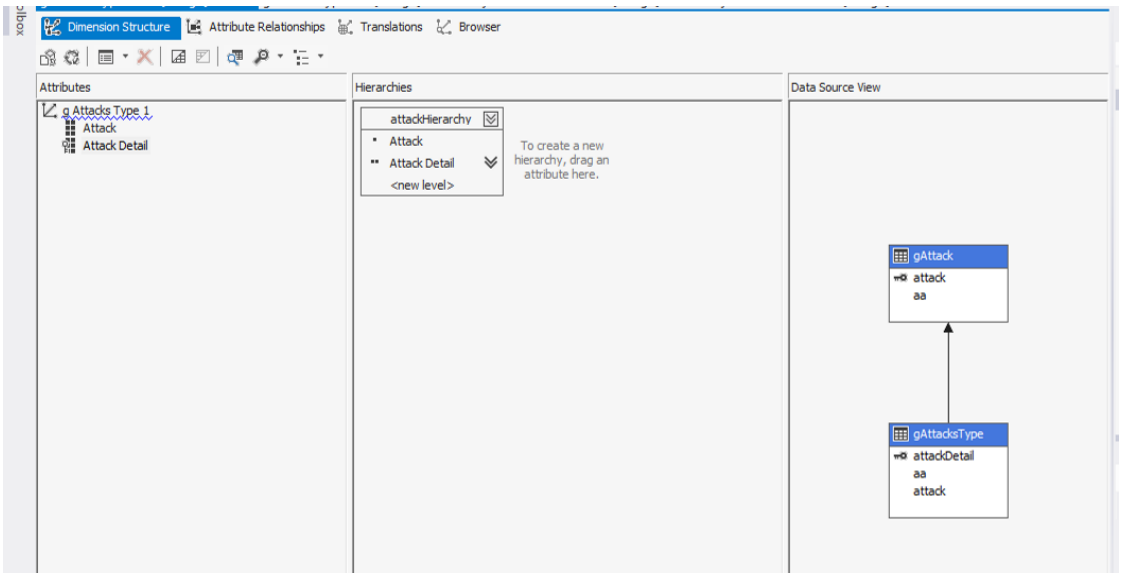
Εικόνα 43: Δήλωση πεδίων διάστασης αποθήκης δεδομένων

Τέλος, ορίζουμε το όνομα της αντίστοιχης διάστασης ως *gAttacksType*:



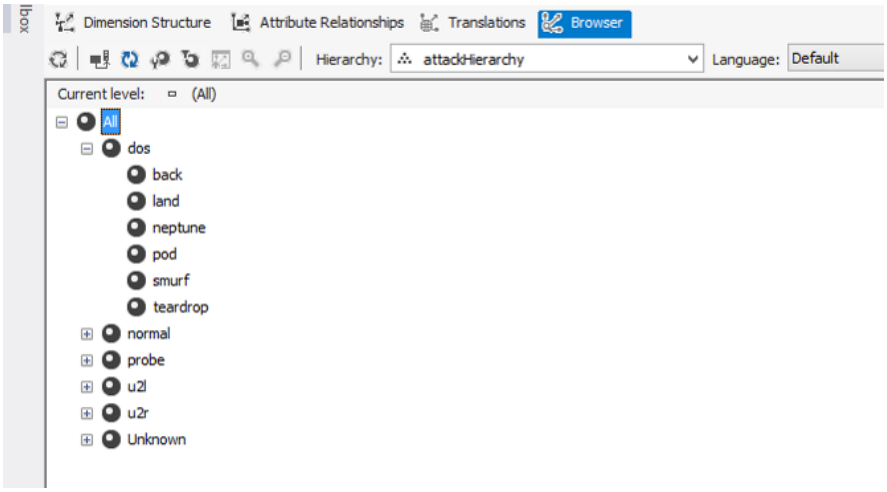
Εικόνα 44: Επιλογή ονόματος διάστασης αποθήκης δεδομένων

Κατόπιν επιλέγονται τα πεδία – ιεραρχία που θα χρησιμοποιηθούν για την αντίστοιχη διάσταση και το όνομα (*attackHierarchy*) του αντίστοιχου πεδίου ιεράρχησης:



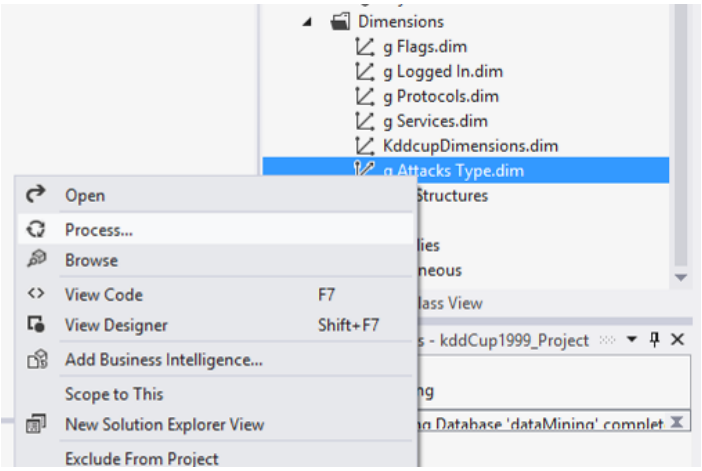
Εικόνα 45: Ορισμός ιεραρχίας διάστασης αποθήκης δεδομένων

Επιπροσθέτως, είναι δυνατόν με την επιλογή της καρτέλας *Browse* να εμφανιστούν οι αντίστοιχες κατηγορίες επιθέσεων -> τίτλος επίθεσης που περιλαμβάνει η σχετική διάσταση της αποθήκης δεδομένων:



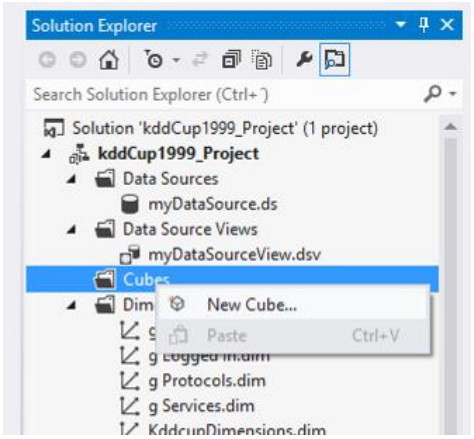
Εικόνα 46: Προβολή τιμών διάστασης αποθήκης δεδομένων

Πριν τη δυνατότητα προεπισκόπησης τιμών μιας διάστασης αποθήκης δεδομένων θα πρέπει εκείνη να γίνει *Processed*:



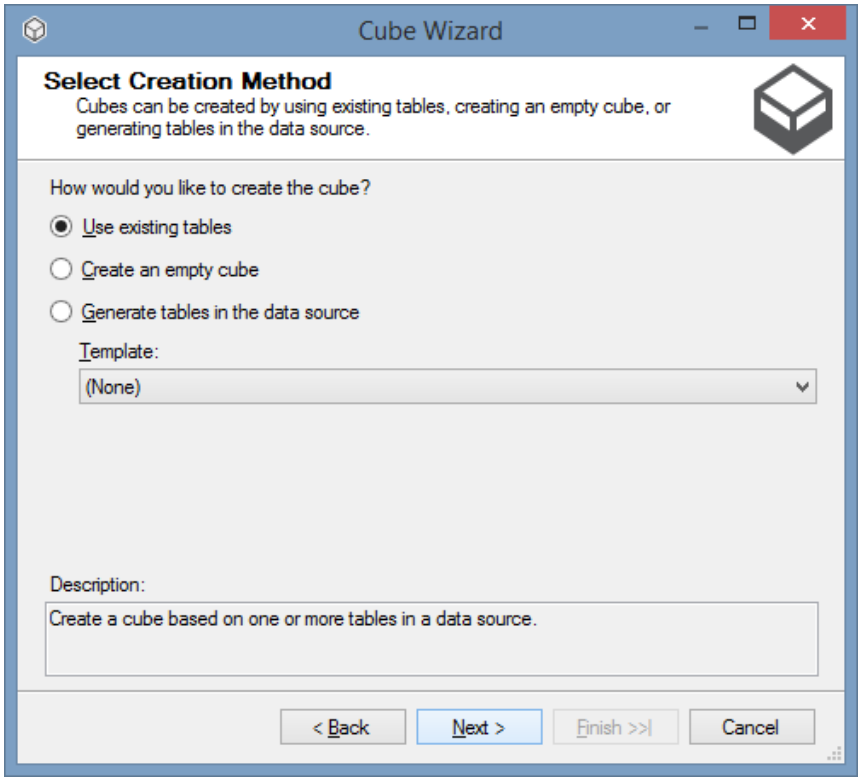
Εικόνα 47: Απαραίτητα Process διάστασης κύβου πριν τη δυνατότητα προσπέλασης των τιμών διάστασης

Έχοντας κάνει process όλες τις διαστάσεις κύβου που δημιουργήθηκαν, μπορεί πλέον να ξεκινήσει η υλοποίηση της αποθήκης δεδομένων. Αρχικά επιλέγεται η δημιουργία καινούργιου κύβου:



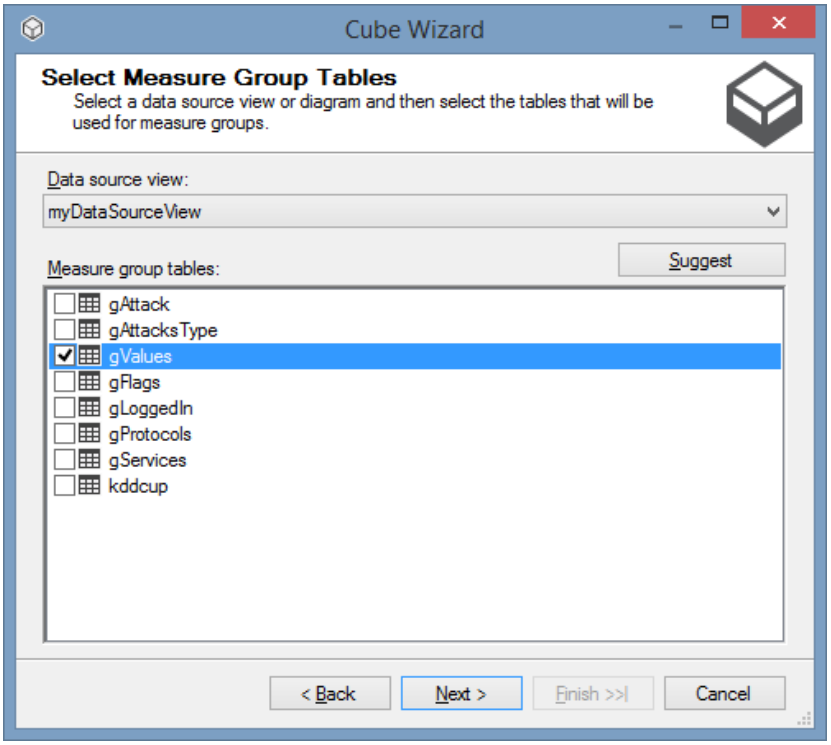
Εικόνα 48: Επιλογή δημιουργίας καινούργιου κύβου

Δηλώνεται πως θα χρησιμοποιηθούν δεδομένα από υπάρχοντα πίνακα στη βάση δεδομένων:



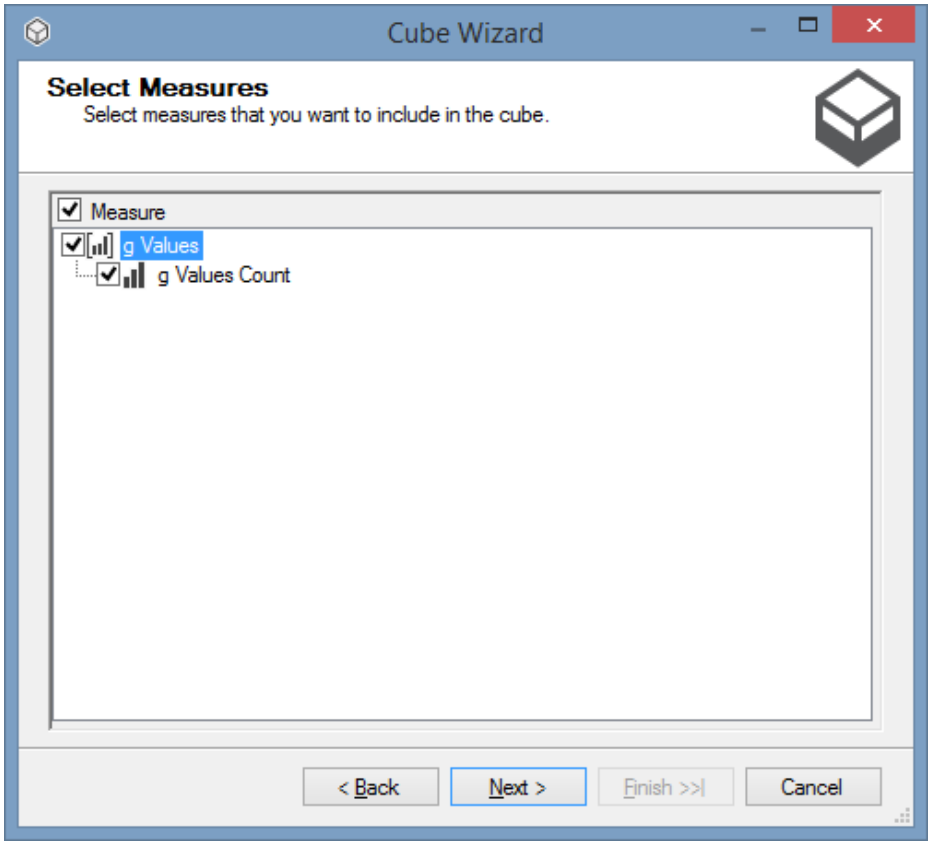
Εικόνα 49: Δήλωση μορφής δεδομένων προέλευσης κύβου

Στη συνέχεια δηλώνεται το μέτρο του κύβου:



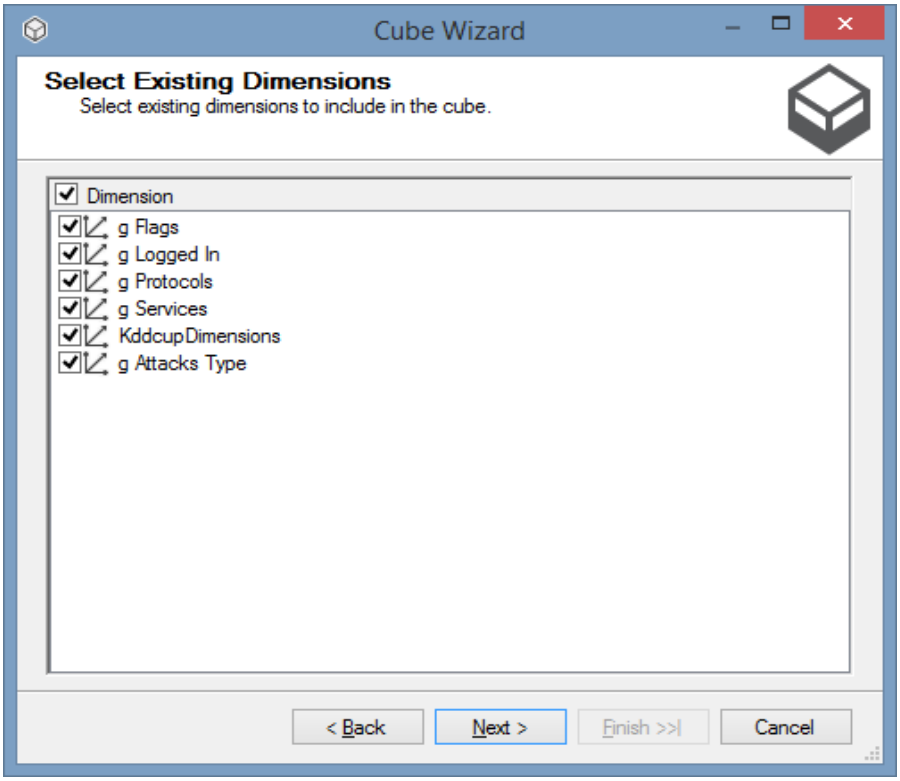
Εικόνα 50: Δήλωση πίνακα βάσης δεδομένων που θα χρησιμοποιηθεί ως μέτρο του κύβου

Κατόπιν ορίζεται το πεδίο του πίνακα *gValues* που θα χρησιμοποιηθεί ως αριθμητικό μέτρο (αριθμός εγγραφών):



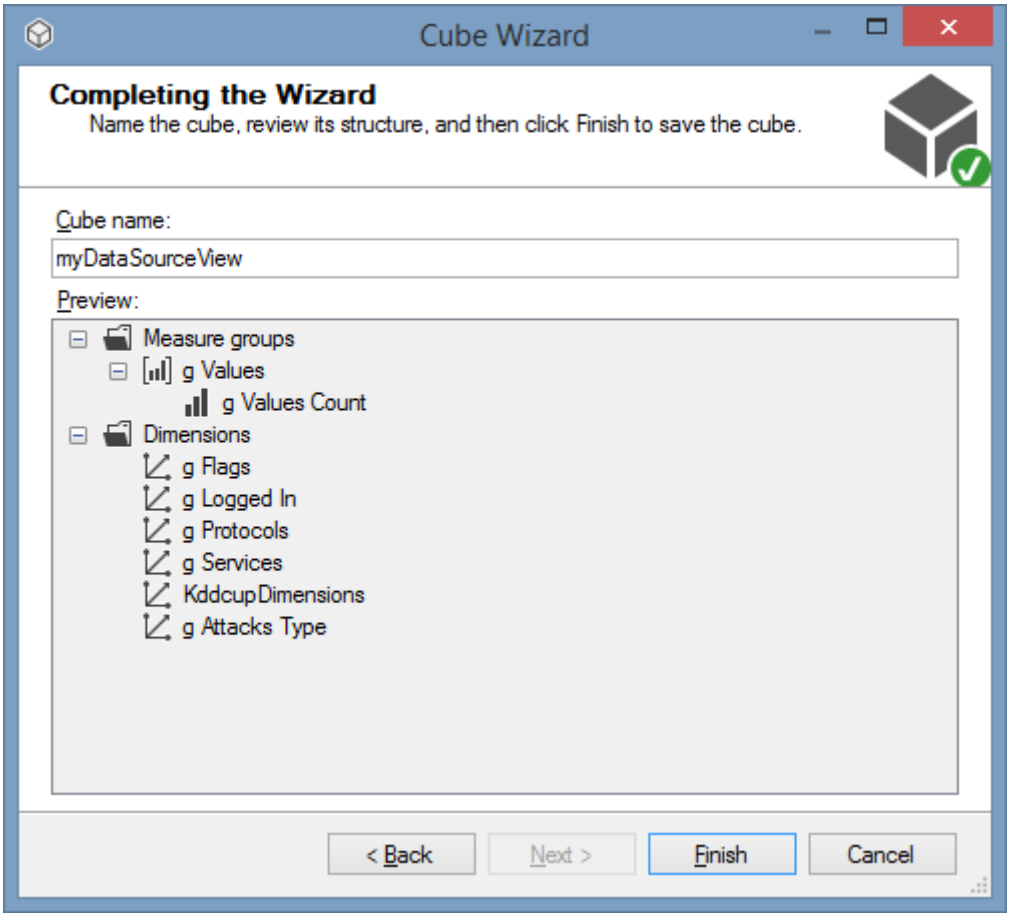
Εικόνα 51: Δήλωση πεδίου μέτρου του κύβου

Κατόπιν ορίζονται οι διαστάσεις (*dimensions*) του κύβου:



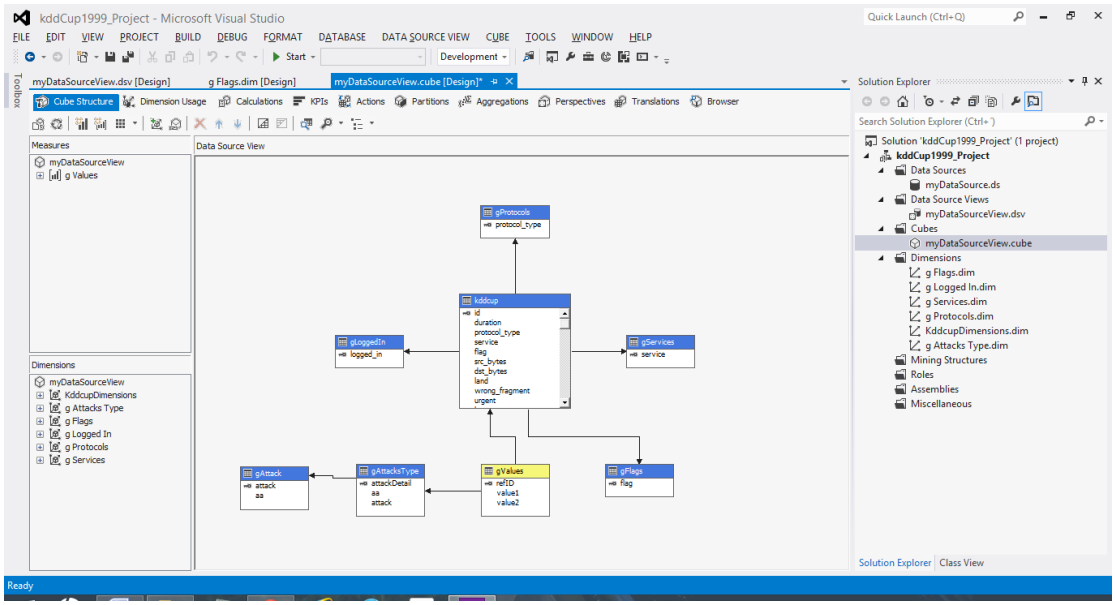
Εικόνα 52: Δήλωση διαστάσεων κύβου

Ο κύβος που μόλις δημιουργήθηκε ονομάζεται επίσης ως *myDataSourceView*:



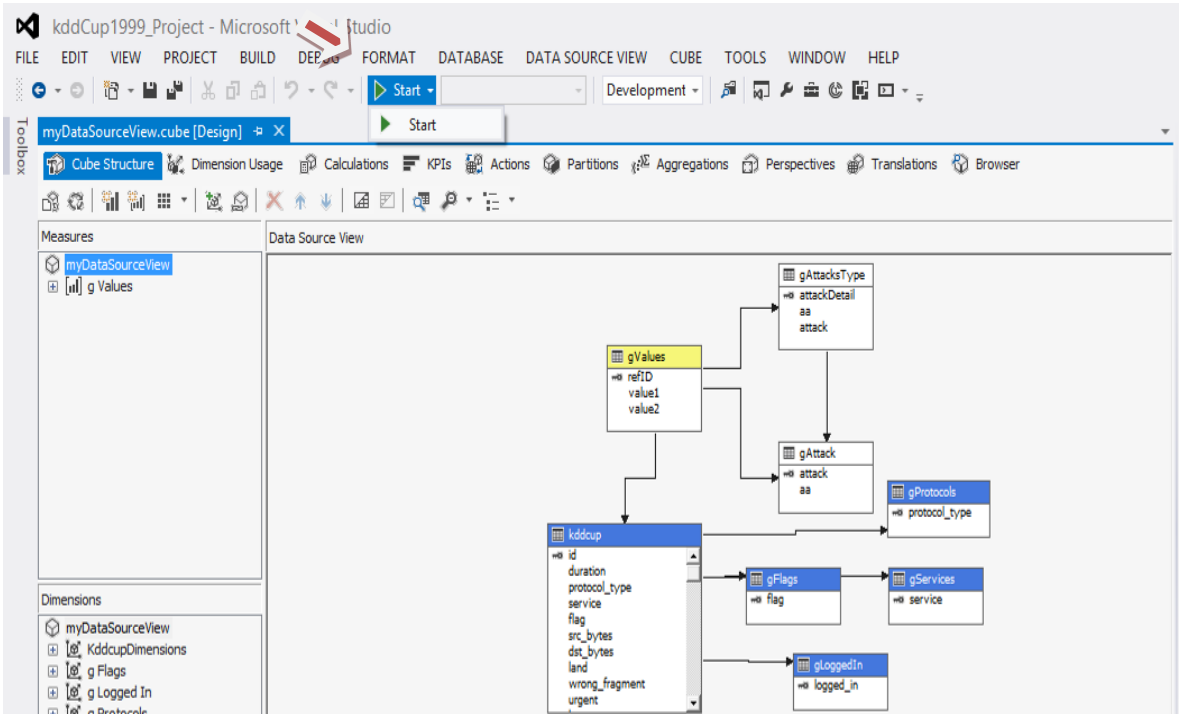
Εικόνα 53: Ονομασία κύβου αποθήκης δεδομένων

Και πλέον η αποθήκη δεδομένων είναι έτοιμη προς χρήση:



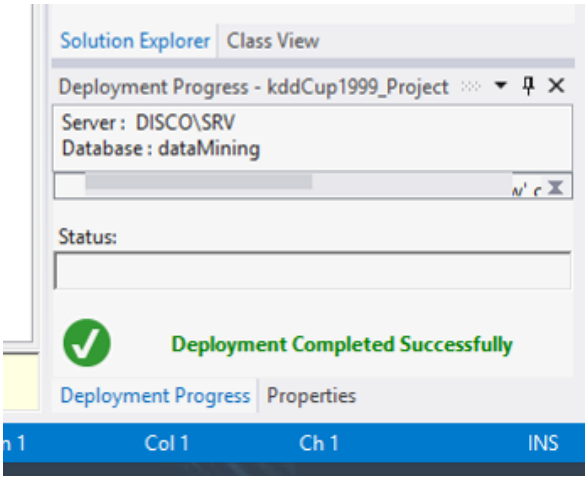
Εικόνα 54: Δημιουργία κύβου

Πριν τη προβολή των στοιχείων αποθήκης δεδομένων, θα πρέπει ο κύβος να γίνει deployed, πατώντας το κουμπί Start στην περιοχή μενού της εφαρμογής:



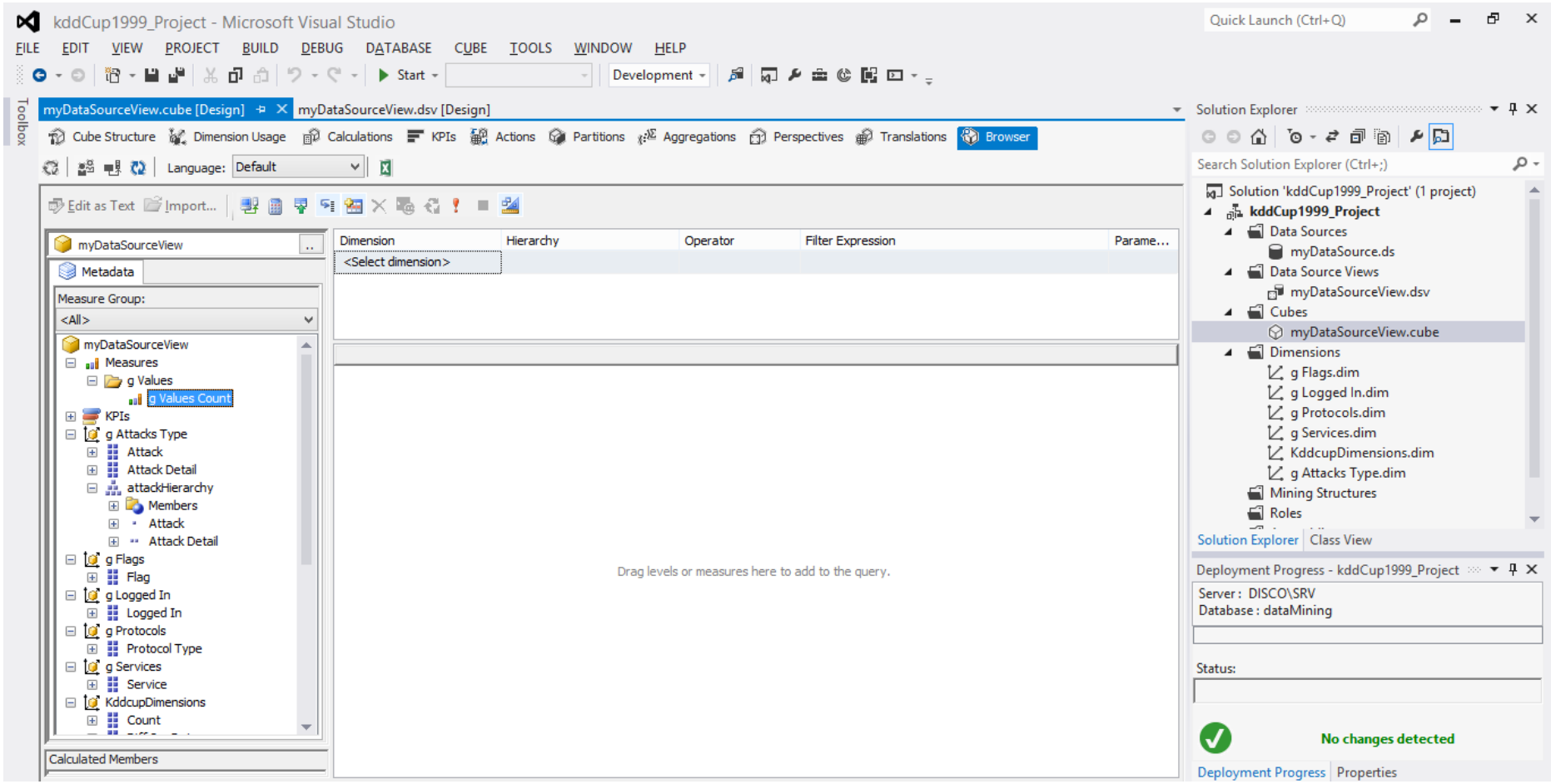
Εικόνα 55: Επιλογή έναρξης deployment αποθήκης δεδομένων

Αν το deployment του κύβου είναι επιτυχές, εμφανίζεται σχετικό μήνυμα στο κάτω – δεξιό μέρος του παραθύρου της εφαρμογής:



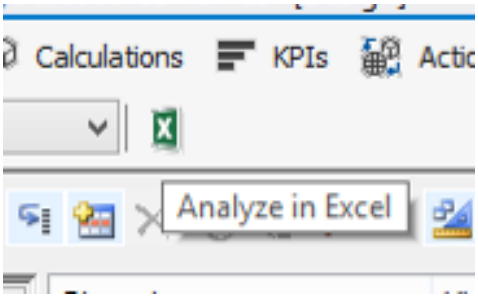
Εικόνα 56: Ενημέρωση επιτυχούς deployment του κύβου

Για τη προσπέλαση των δεδομένων του κύβου, επιλέγεται τη καρτέλα Browse, οπότε και εμφανίζονται στα αριστερά του παραθύρου οι διαστάσεις κύβου και τα μέτρα:



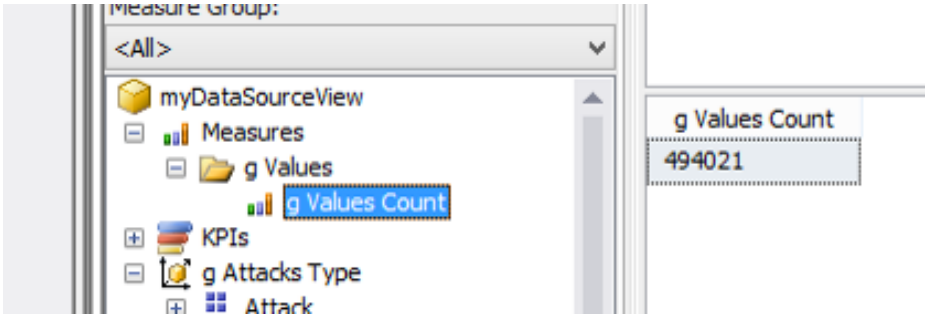
Εικόνα 57: Εμφάνιση διαστάσεων, ιεραρχιών και μέτρων αποθήκης δεδομένων

Θα πρέπει να αναφερθεί πως από την έκδοση Visual Studio 2010 (στα πλαίσια της παρούσας εργασίας χρησιμοποιήθηκε το Visual Studio 2012), έχει αφαιρεθεί η δυνατότητα προσπέλασης των δεδομένων με τη μορφή πίνακα (δηλώνοντας στήλες και γραμμές), παρά παρέχεται μόνο η δυνατότητα δήλωσης πεδίων γραμμών. Εναλλακτικά μπορεί να σχεδιαστεί κύβος δεδομένων (όπως το Visual Studio 2008), πατώντας το κουμπί Analyze in Excel, όπως θα δούμε παρακάτω:



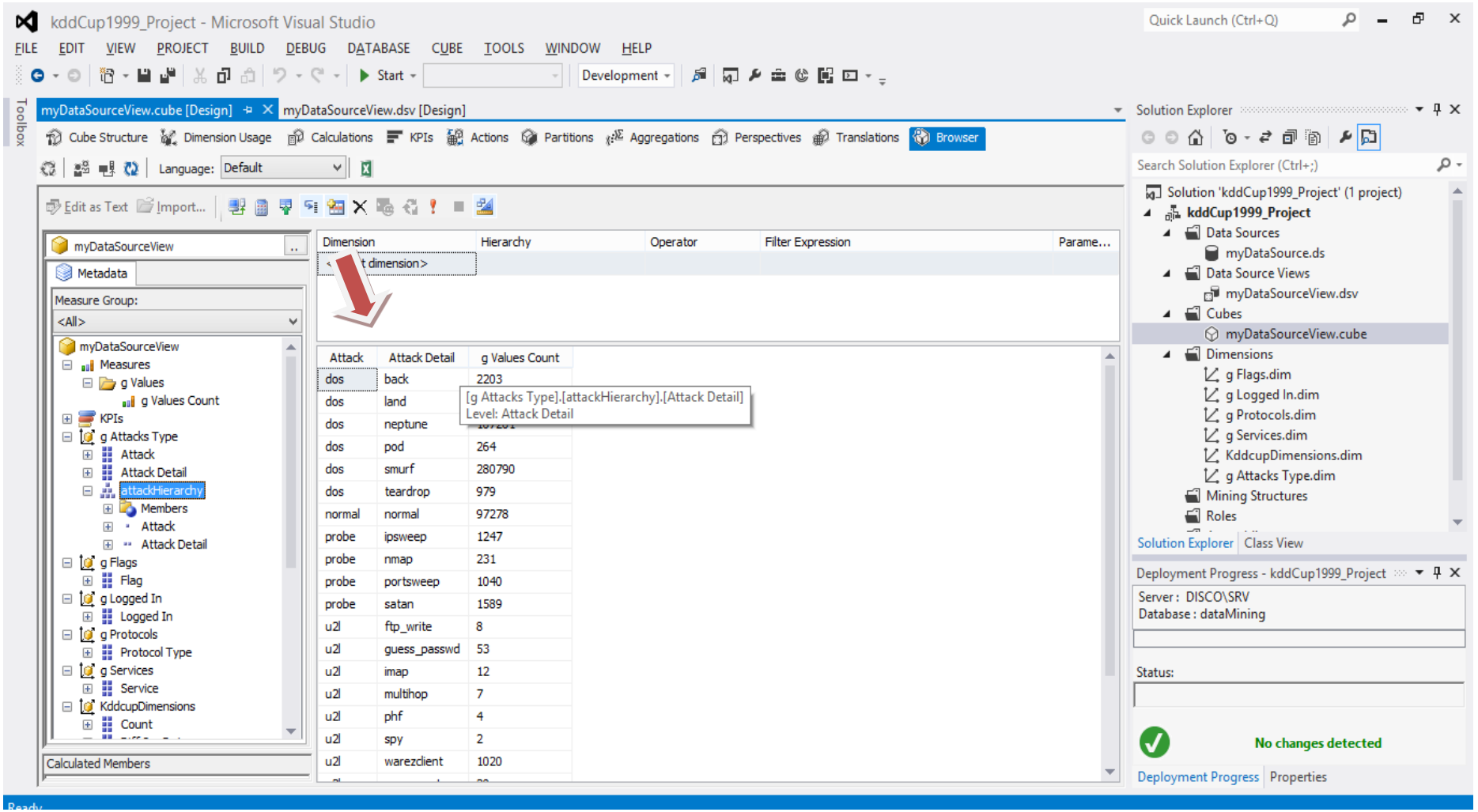
Εικόνα 58: Κουμπί μεταφοράς αποθήκης δεδομένων στο Excel

Αρχικά μεταφέρουμε στο πλαίσια σχεδίασης της αποθήκης δεδομένων (στο Visual Studio 2012), το πεδίο μέτρου της αποθήκης δεδομένων (αριθμός μετρήσεων συνδέσεων):



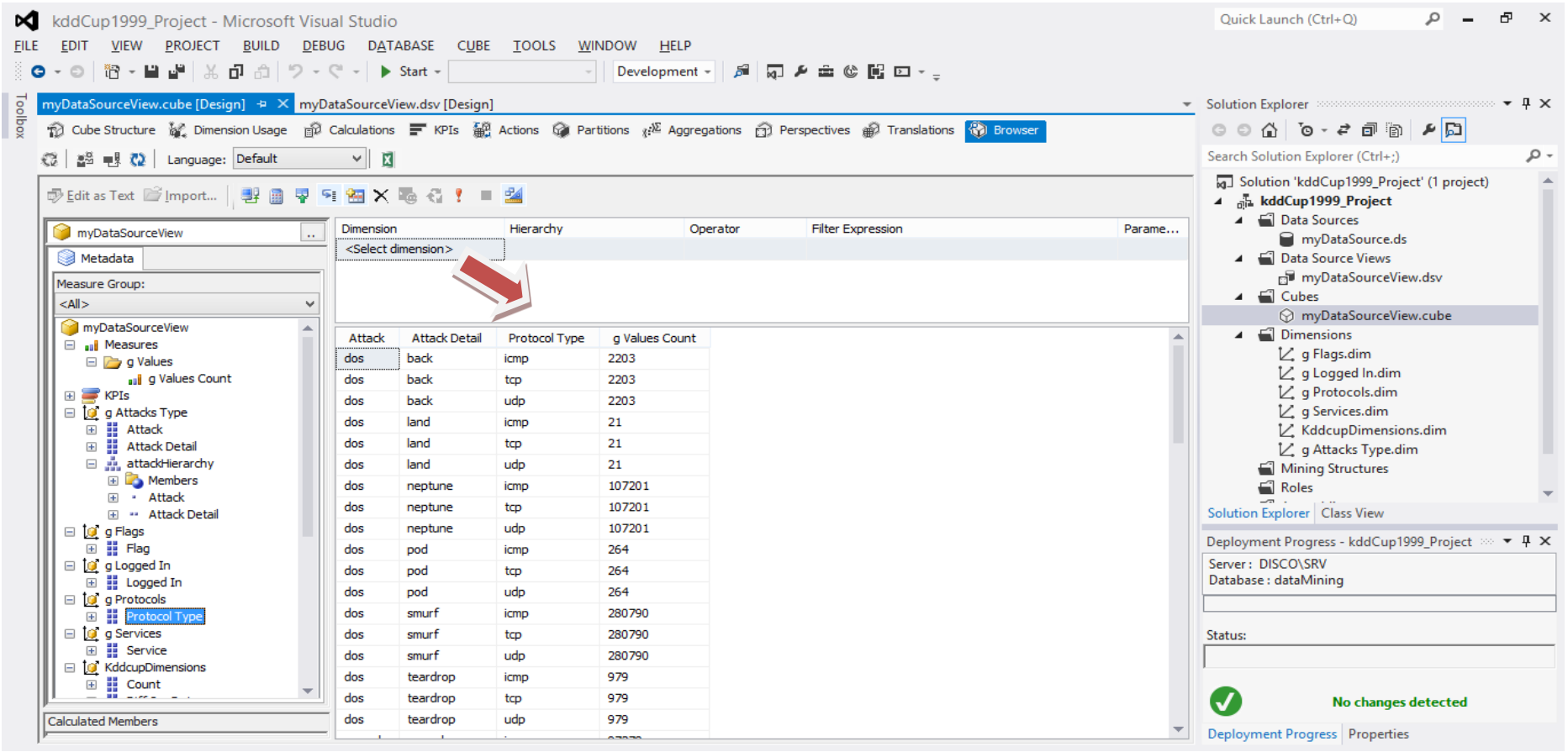
Εικόνα 59: Επιλογή - εμφάνιση μέτρου αποθήκης δεδομένων

Στη συνέχεια για να προβληθεί μέρος της αποθήκης δεδομένων επιλέγονται αντίστοιχες διαστάσεις. Έστω πως επιθυμείται η προβολή πληροφοριών σχετικά με το είδος επιθέσεων. Σε αυτή την περίπτωση μεταφέρεται (drag and drop) στο πλαίσιο σχεδίασης της αποθήκης δεδομένων, η διάσταση *gAttacksType* και χαρακτηριστικό ιεραρχίας *attackHierarchy,* οπότε εμφανίζεται ο αριθμός επιθέσεων ανά κατηγορία επίθεσης -> τίτλος μορφής επίθεσης:


kddCup1999_Project - Microsoft Visual Studio
Quick Launch (Ctrl+Q)
FILE EDIT VIEW PROJECT BUILD DEBUG DATABASE CUBE TOOLS WINDOW HELP
Start
Development
Toolbox
myDataSourceView.cube [Design]
myDataSourceView.dsv [Design]
Cube Structure
Dimension Usage
Calculations
KPIs
Actions
Partitions
Aggregations
Perspectives
Translations
Browser
Language: Default
Edit as Text
Import...
myDataSourceView
Metadata
Measure Group:
<All>
myDataSourceView
Measures
g Values
g Values Count
KPIs
g Attacks Type
Attack
Attack Detail
attackHierarchy
Members
Attack
Attack Detail
g Flags
Flag
g Logged In
Logged In
g Protocols
Protocol Type
g Services
Service
KddcupDimensions
Count
Calculated Members
Dimension
Hierarchy
Operator
Filter Expression
Parame...
t dimension>
Attack
Attack Detail
g Values Count
dos back 2203
dos land
dos neptune
dos pod 264
dos smurf 280790
dos teardrop 979
normal normal 97278
probe ipsweep 1247
probe nmap 231
probe portsweep 1040
probe satan 1589
u2l ftp_write 8
u2l guess_passwd 53
u2l imap 12
u2l multihop 7
u2l phf 4
u2l spy 2
u2l warezclient 1020
[g Attacks Type].[attackHierarchy].[Attack Detail]
Level: Attack Detail
Solution Explorer
Search Solution Explorer (Ctrl+;)
Solution 'kddCup1999_Project' (1 project)
kddCup1999_Project
Data Sources
myDataSource.ds
Data Source Views
myDataSourceView.dsv
Cubes
myDataSourceView.cube
Dimensions
g Flags.dim
g Logged In.dim
g Protocols.dim
g Services.dim
KddcupDimensions.dim
g Attacks Type.dim
Mining Structures
Roles
Solution Explorer
Class View
Deployment Progress - kddCup1999_Project
Server : DISCO\SRV
Database : dataMining
Status:
No changes detected
Deployment Progress
Properties
Ready

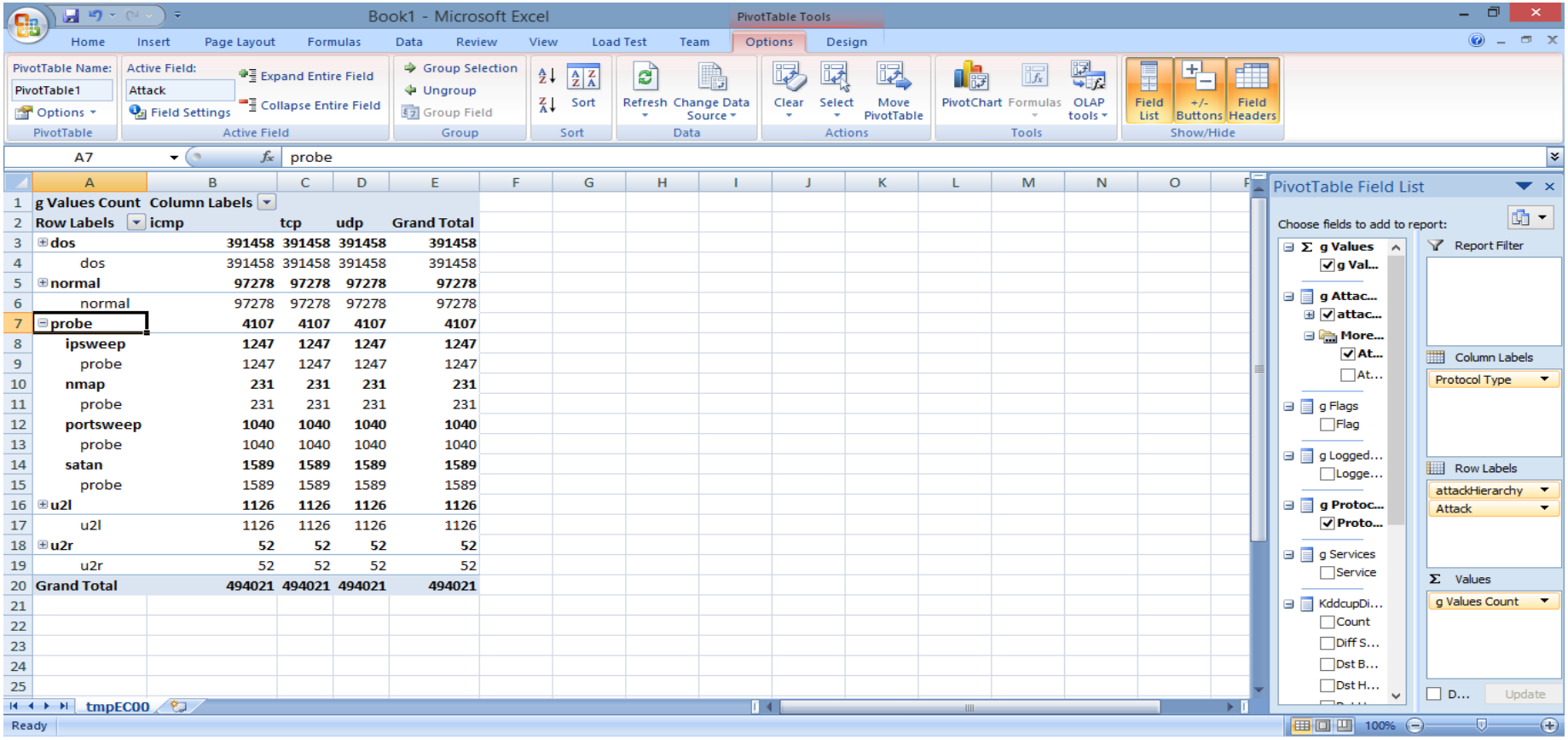
Εικόνα 60: Προβολή αριθμού επιθέσεων, ανά κατηγορία και μορφή επίθεσης

Η προβολή των στοιχείων της αποθήκης δεδομένων μπορεί να επεκταθεί με τη προσθήκη, οποιονδήποτε από τις διαστάσεις που ορίστηκαν. Έστω πως επιλέγεται η επιπλέον προσθήκη της διάστασης gProtocol, οπότε και θα εμφανιστούν οι εγγραφές συγχρόνως ομαδοποιημένες και ως προς το πρωτόκολλο που αντιστοιχεί η κάθε σύνδεση – εγγραφή του συνόλου δεδομένων:



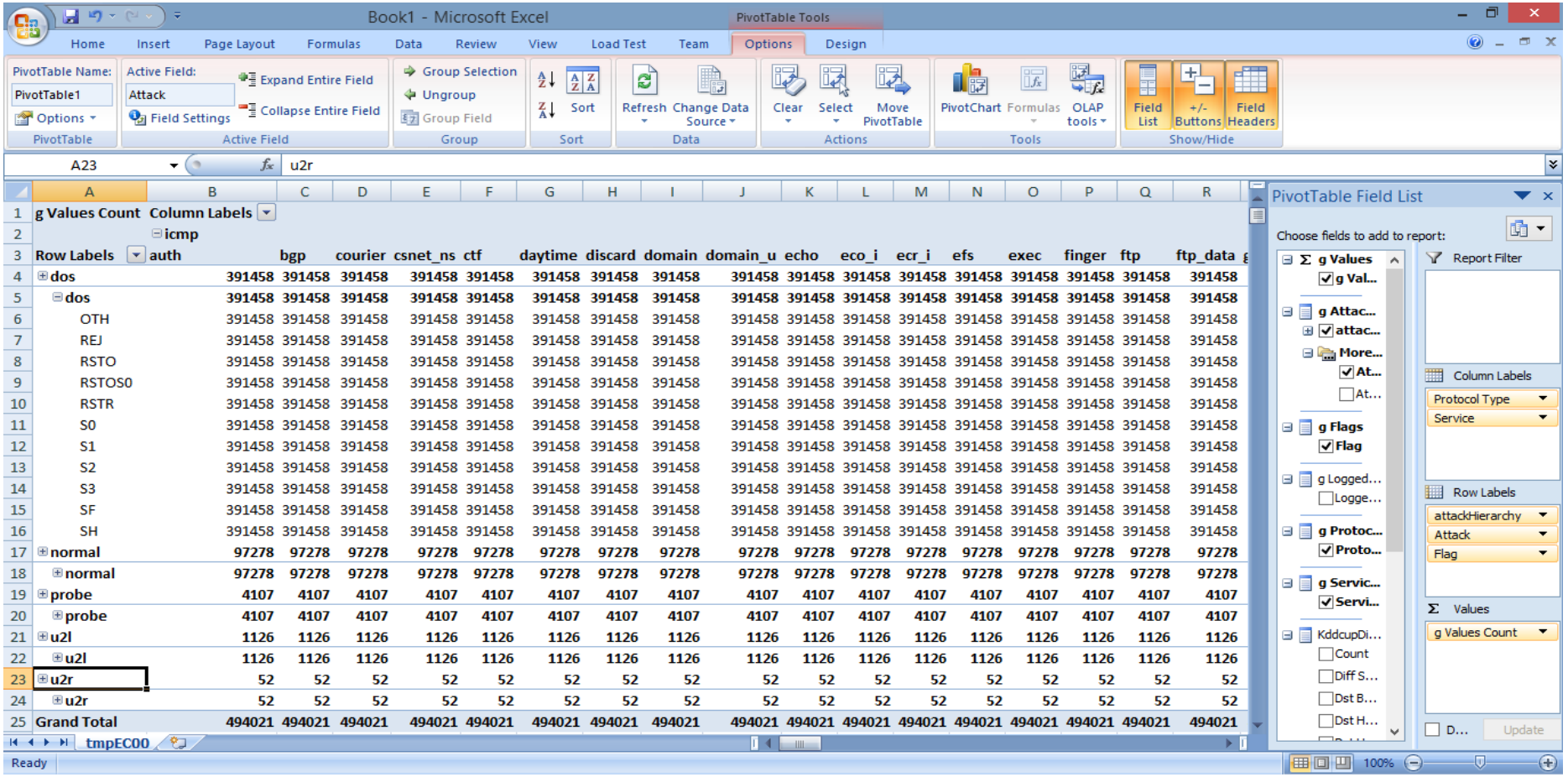
Εικόνα 61: Προσθήκη διάστασης στην αποθήκη δεδομένων

Αν επιθυμείται η προβολή της αποθήκης δεδομένων με τη παραδοσιακή μορφή (όπως στο Visual Studio 2008), θα πρέπει να πατηθεί το κουμπί *Analyze in Excel*, οπότε αρχικά μεταφέρονται τα δεδομένα στο Excel και στο πλαίσιο δεξιά δηλώνονται τα χαρακτηριστικά στις στήλες γραμμές του φύλλο εργασιών:



Εικόνα 62: Προβολή αποθήκης δεδομένων στο Excel

Γενικότερα μπορούμε να προσθέσουμε στη προβολή, όσες διαστάσεις από τις είκοσι διαθέσιμες επιθυμούμε, προκειμένου να διερευνήσουμε συσχετίσεις μεταξύ των χαρακτηριστικών:



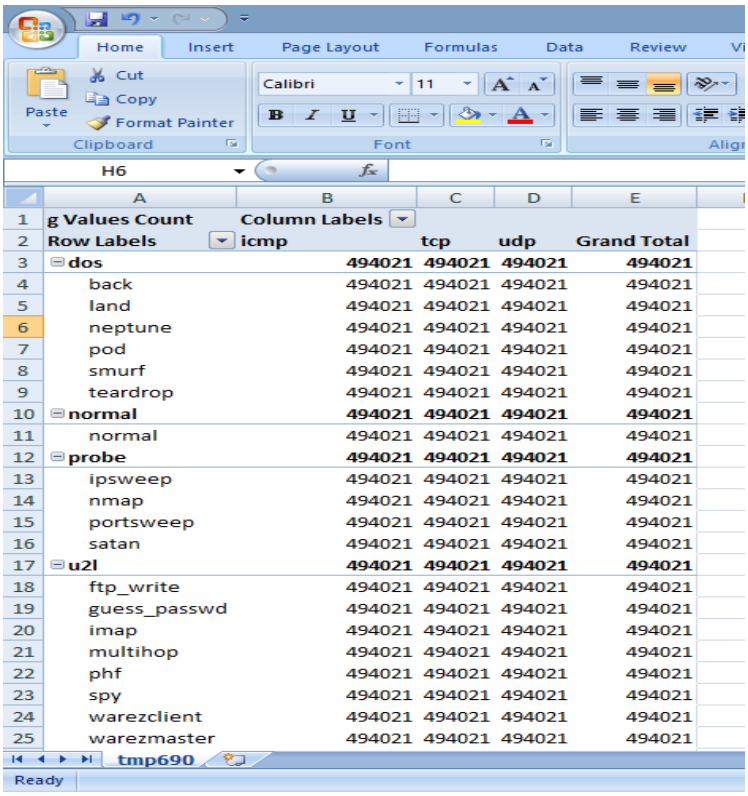
Εικόνα 63: Παράδειγμα εφαρμογής πολλαπλών διαστάσεων στην αποθήκη δεδομένων

Στο παραπάνω παράδειγμα στον οριζόντιο άξονα δηλώθηκα τα χαρακτηριστικά είδος επίθεσης (ιεραρχία) και τιμή πεδίου flag, ενώ στον κάθετο άξονα το είδος του πρωτοκόλλου που χρησιμοποιήθηκε και η υπηρεσία (service) χρήσης.

6.4 Αναλυτική επεξεργασία δεδομένων- OLAP

Οι αποθήκες δεδομένων, επιτρέπουν το χρήστη να κατανοήσει τα δεδομένα πριν την εφαρμογή μεθοδολογιών εξόρυξης δεδομένων. Οι βασικές πράξεις που εφαρμόζονται συνήθως σε μια αποθήκη δεδομένων είναι οι: *roll-up, drill-down, slice, dice, pivot*. Αν θεωρήσουμε πως οι διαστάσεις που θα χρησιμοποιήσουμε για τη προβολή της αποθήκης δεδομένων είναι οι μορφές επιθέσεων (ιεραρχία) ως γραμμές και το είδος πρωτοκόλλου ως στήλες, ακολουθεί παράδειγμα των πέντε (5) βασικών πράξεων σε αποθήκη δεδομένων (19), (22), (21):

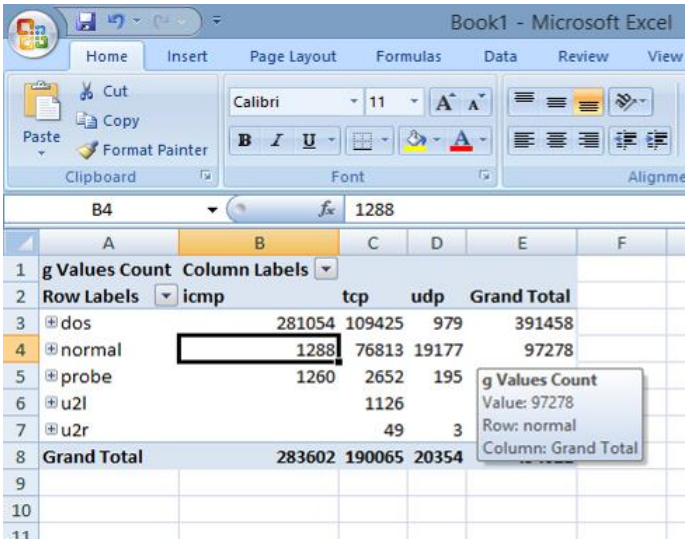
roll-up: αποτελεί μια πιο γενική προβολή των δεδομένων του κύβου και αντιστοιχεί συνήθως με τη μετάβαση ένα επίπεδο πάνω στην ιεραρχία διάστασης ενός κύβου. Για παράδειγμα στον κάτωθι κύβο (*Analyze in Excel*) προβάλλονται ο αριθμός επιθέσεων για κάθε δυνατή μορφή επίθεσης ανά πρωτόκολλο χρήσης:



	A	B	C	D	E
1	g Values Count	Column Labels			
2	Row Labels	icmp	tcp	udp	Grand Total
3	**dos**	**494021**	**494021**	**494021**	**494021**
4	back	494021	494021	494021	494021
5	land	494021	494021	494021	494021
6	neptune	494021	494021	494021	494021
7	pod	494021	494021	494021	494021
8	smurf	494021	494021	494021	494021
9	teardrop	494021	494021	494021	494021
10	**normal**	**494021**	**494021**	**494021**	**494021**
11	normal	494021	494021	494021	494021
12	**probe**	**494021**	**494021**	**494021**	**494021**
13	ipsweep	494021	494021	494021	494021
14	nmap	494021	494021	494021	494021
15	portsweep	494021	494021	494021	494021
16	satan	494021	494021	494021	494021
17	**u2l**	**494021**	**494021**	**494021**	**494021**
18	ftp_write	494021	494021	494021	494021
19	guess_passwd	494021	494021	494021	494021
20	imap	494021	494021	494021	494021
21	multihop	494021	494021	494021	494021
22	phf	494021	494021	494021	494021
23	spy	494021	494021	494021	494021
24	warezclient	494021	494021	494021	494021
25	warezmaster	494021	494021	494021	494021

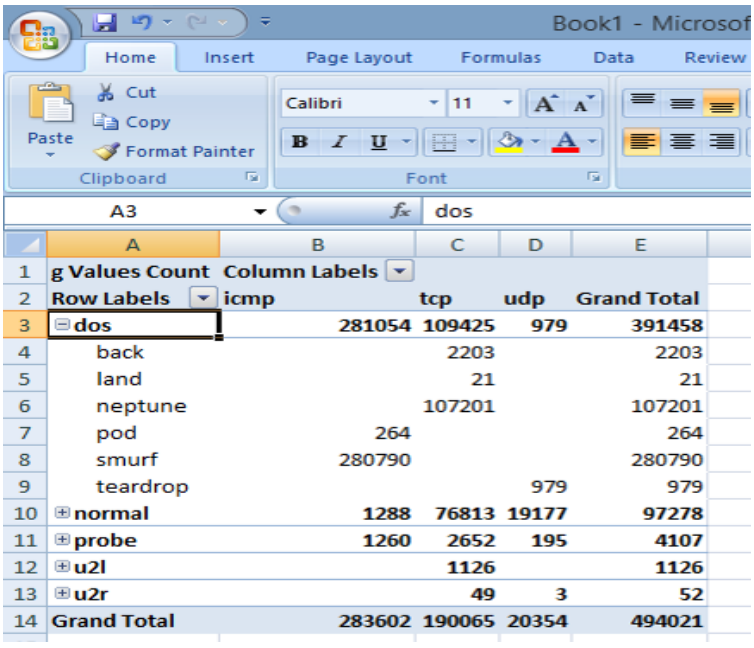
Εικόνα 64: Προβολή δεδομένων κύβου

Αν περιορίσουμε τη προβολή του αριθμού επιθέσεων, μόνο ανά κατηγορία επιθέσεων και όχι λεπτομερή περιγραφή ανά τίτλο επιθέσεων, εφαρμόζουμε με πράξη *roll-up*:



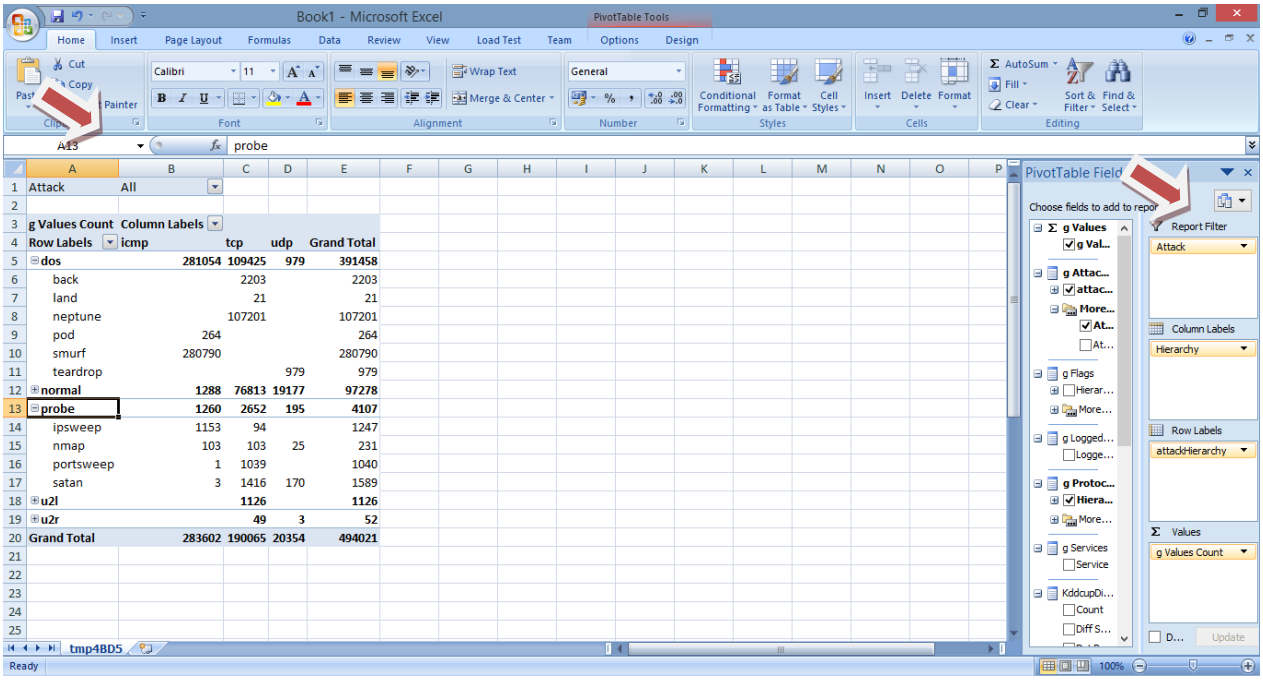
Εικόνα 65: Εφαρμογή πράξης roll-up

drill-down: αποτελεί την αντίθεση ακριβώς διαδικασία του roll-up. Μετάβαση ένα επίπεδο κάτω στην ιεραρχία διάστασης ενός κύβου. Αν στη παραπάνω εικόνα κάνουμε κλικ στον σταυρό δίπλα στη μορφή επιθέσεων *dos* και εμφανιστεί η συχνότητα εμφάνισης (ανά πρωτόκολλο χρήσης), τότε εκτελούμε μια πράξη *drill-down*:



Εικόνα 66: Εκτέλεση πράξης drill-down

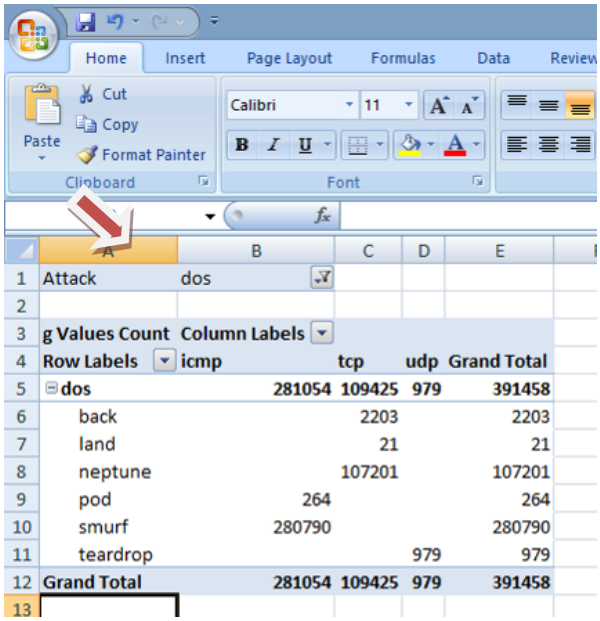
slice: αποτελεί τη πράξη εξέτασης ενός κύβου με την επιλογή μιας μόνο διάστασης. Αν στην αποθήκη δεδομένων που εμφανίζεται παρακάτω (*Analyze in Excel*), ορίσουμε ως φίλτρο την κατηγορία επίθεσης:



Attack	All				
g Values Count	Column Labels				
Row Labels	icmp	tcp	udp	Grand Total	
dos	281054	109425	979	391458	
back		2203		2203	
land		21		21	
neptune		107201		107201	
pod	264			264	
smurf	280790			280790	
teardrop			979	979	
normal	1288	76813	19177	97278	
probe	1260	2652	195	4107	
ipsweep	1153	94		1247	
nmap	103	103	25	231	
portsweep	1	1039		1040	
satan	3	1416	170	1589	
u2l		1126		1126	
u2r		49	3	52	
Grand Total	283602	190065	20354	494021	

Εικόνα 67: Ορισμό φίλτρων σε κύβο

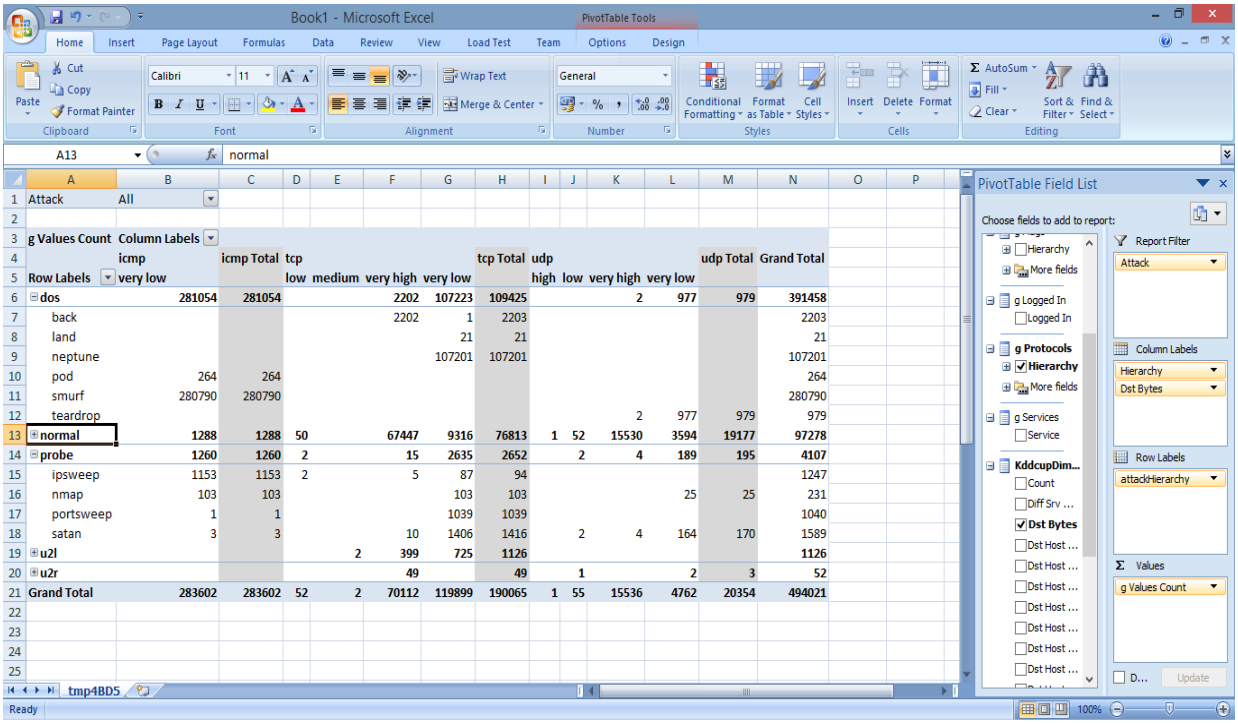
και επιλέξουμε την εφαρμογή φίλτρου στις κατηγορίες επίθεσης, αποκλειστικά ίση με κατηγορία επίθεσης *dos*, το αποτέλεσμα που λαμβάνεται αποτελεί εφαρμογή πράξης *slice*:



Attack	dos			
g Values Count	Column Labels			
Row Labels	icmp	tcp	udp	Grand Total
dos	281054	109425	979	391458
back		2203		2203
land		21		21
neptune		107201		107201
pod	264			264
smurf	280790			280790
teardrop			979	979
Grand Total	281054	109425	979	391458

Εικόνα 68: Εφαρμογή πράξης slice σε κύβο

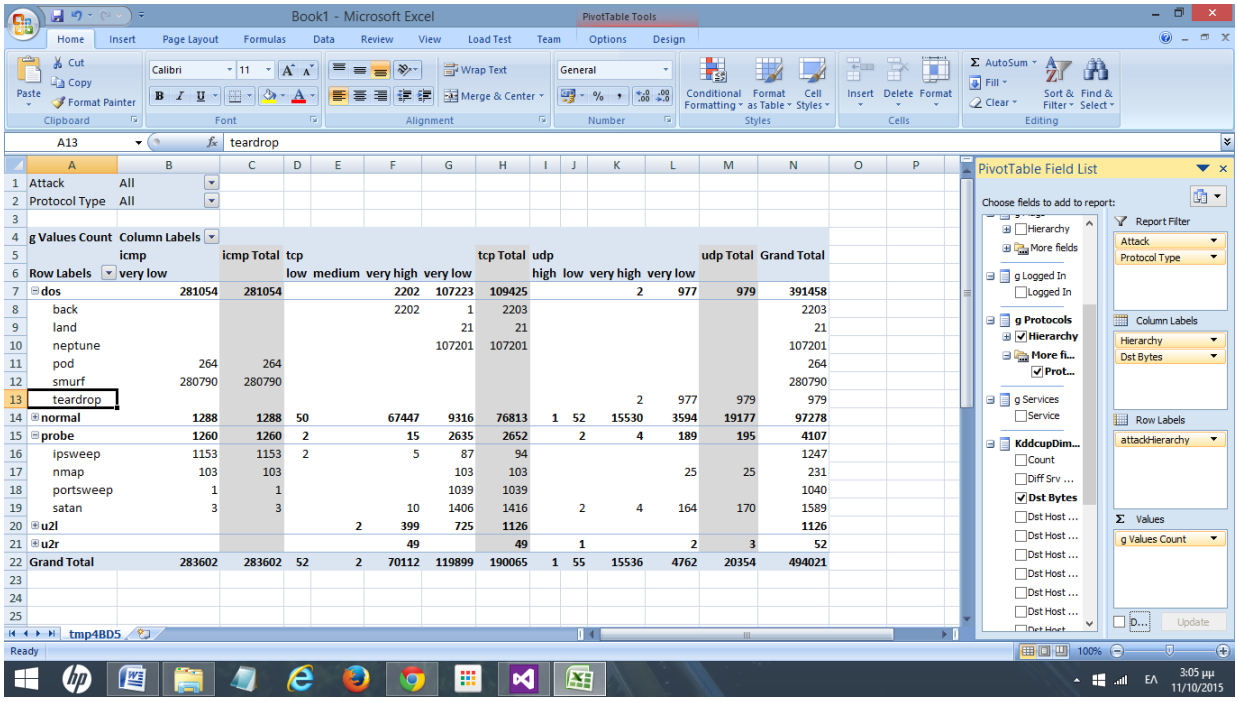
dice: αποτελεί τη πράξη εξέτασης ενός κύβου με την επιλογή δυο ή περισσότερων διαστάσεων. Αρχικά προσθέτουμε μια επιπλέον διάσταση στις στήλες του κύβου (το χαρακτηριστικό *dst_bytes* με πέντε πιθανές τιμές), οπότε ο αρχικός κύβος που προκύπτει είναι ο εξής:



Attack	All														
g Values Count	Column Labels														
	icmp		icmp Total	tcp				tcp Total	udp				udp Total	Grand Total	
Row Labels	very low			low	medium	very high	very low		high	low	very high	very low			
dos	281054		281054			2202	107223	109425			2	977	979	391458	
back						2202	1	2203						2203	
land							21	21						21	
neptune							107201	107201						107201	
pod	264		264											264	
smurf	280790		280790											280790	
teardrop												2	977	979	979
normal	1288		1288	50		67447	9316	76813	1	52	15530	3594	19177	97278	
probe	1260		1260	2		15	2635	2652			2	4	189	195	4107
ipsweep	1153		1153	2		5	87	94							1247
nmap	103		103				103	103					25	25	231
portsweep	1		1				1039	1039							1040
satan	3		3			10	1406	1416			2	4	164	170	1589
u2l					2	399	725	1126							1126
u2r						49		49			1		2	3	52
Grand Total	283602		283602	52	2	70112	119899	190065	1	55	15536	4762	20354	494021	

Εικόνα 69: Προσθήκη διάστασης στις στήλες του κύβου

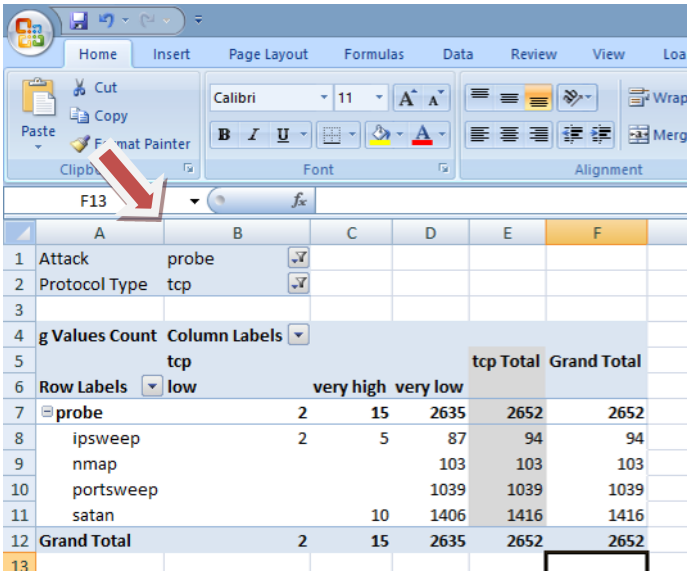
Στη συνέχεια ορίζουμε ένα ακόμα φίλτρο για τον κύβο (το χαρακτηριστικό είδος πρωτοκόλλου):



Attack	All													
Protocol Type	All													
g Values Count	Column Labels													
	icmp		icmp Total	tcp					tcp Total	udp				udp Total	Grand Total
Row Labels	very low			low	medium	very high	very low		high	low	very high	very low		
dos	281054		281054			2202	107223	109425			2	977	979	391458
back						2202	1	2203						2203
land							21	21						21
neptune							107201	107201						107201
pod	264		264											264
smurf	280790		280790											280790
teardrop											2	977	979	979
normal	1288		1288	50		67447	9316	76813	1	52	15530	3594	19177	97278
probe	1260		1260	2		15	2635	2652		2	4	189	195	4107
ipsweep	1153		1153	2		5	87	94						1247
nmap	103		103				103	103				25	25	231
portsweep	1		1				1039	1039						1040
satan	3		3			10	1406	1416		2	4	164	170	1589
u2l					2	399	725	1126						1126
u2r						49		49		1		2	3	52
Grand Total	283602		283602	52	2	70112	119899	190065	1	55	15536	4762	20354	494021

Εικόνα 70: Προσθήκη 2ου φίλτρου στον κύβο

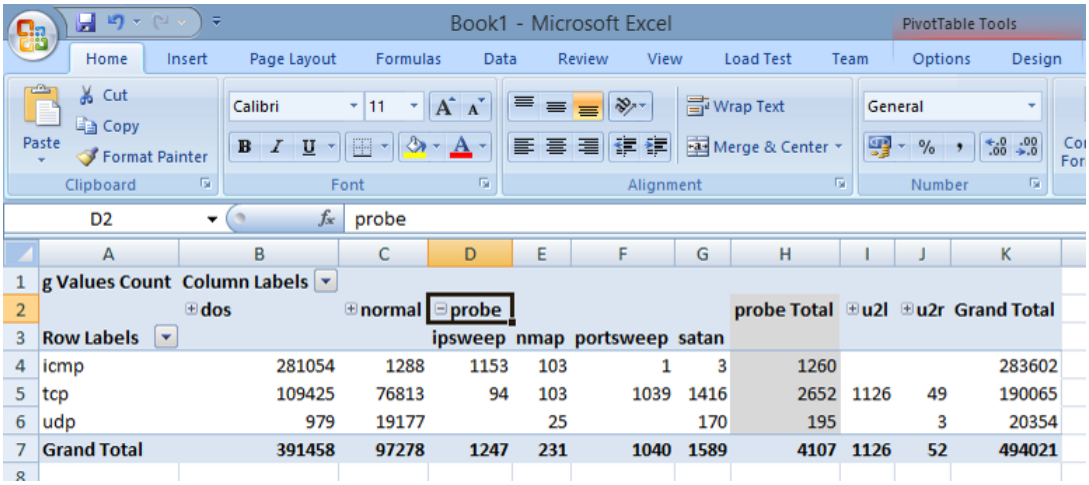
Στη συνέχεια αν εφαρμόσουμε φίλτρα για πρωτόκολλο ίσο με *tcp* και κατηγορία επίθεσης *probe*, έχουμε εφαρμόσει μια πράξη *dice*:



Attack	probe				
Protocol Type	tcp				
g Values Count	Column Labels				
	tcp			tcp Total	Grand Total
Row Labels	low	very high	very low		
probe	2	15	2635	2652	2652
ipsweep	2	5	87	94	94
nmap			103	103	103
portsweep			1039	1039	1039
satan		10	1406	1416	1416
Grand Total	2	15	2635	2652	2652

Εικόνα 71: Εκτέλεση πράξης dice

pivot: αποτελεί πρακτικά εναλλαγή μεταξύ των χαρακτηριστικών στήλης και χαρακτηριστικών γραμμών ενός κύβου, δηλ. αλλαγή διάταξης των διαστάσεων ενός κύβου. Στο παράδειγμά μας αν τοποθετούσαμε, όπως απεικονίζεται στη παρακάτω εικόνα, τα είδη πρωτοκόλλων στις γραμμές και το είδος επίθεσης στις στήλες, εκτελούμε μια πράξη *pivot*:



g Values Count	Column Labels									
	dos	normal	probe				probe Total	u2l	u2r	Grand Total
Row Labels			ipsweep	nmap	portsweep	satan				
icmp	281054	1288	1153	103	1	3	1260			283602
tcp	109425	76813	94	103	1039	1416	2652	1126	49	190065
udp	979	19177		25		170	195		3	20354
Grand Total	**391458**	**97278**	**1247**	**231**	**1040**	**1589**	**4107**	**1126**	**52**	**494021**

Εικόνα 72: Εκτέλεση πράξης pivot

6.5 Σημαντικά χαρακτηριστικά(feature selection)

Για τον υπολογισμό των σημαντικών χαρακτηριστικών του δείγματος θα χρησιμοποιηθεί η εφαρμογή WEKA. Αρχικά θα δημιουργηθεί το σχετικό αρχείο ARFF και θα αφαιρεθούν από το σύνολο δεδομένων τα εικοσιένα (21) χαρακτηριστικά που διαπιστώθηκε από τη παραπάνω ανάλυση πως παρουσιάζουν την ίδια τιμή σε ποσοστό τουλάχιστον 92% για το σύνολο των εγγραφών του συνόλου δεδομένων (σχεδόν 500,000 εγγραφές). Πιο συγκεκριμένα θα αφαιρεθούν τα χαρακτηριστικά με δείκτη: 1, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 27, 28, 31, 40, και 41.

Επίσης, η εξαρτημένη μεταβλητή είναι εκφρασμένη σε 23 διαφορετικές κατηγοριακές τιμές (22 μορφές επιθέσεων και η τιμή περί φυσιολογικής σύνδεσης -normal). Χρησιμοποιώντας τη κατηγοριοποίηση μορφών επιθέσεων σε τέσσερις (4) κατηγορίες σύμφωνα με τον Πίνακα 1 της παρούσας εργασία, η μεταβλητή κλάσης θα εκφραστεί με πέντε πιθανές τιμές (dos, probe, u2r, u2l, normal).

Η επικεφαλίδα του αρχείου ARFF που δημιουργήθηκε με την εκτέλεση της εντολής CSVLoader του WEKA, παρουσιάζεται παρακάτω:

```
@relation forWeka



@attribute protocol_type {tcp,udp,icmp}

@attribute service
```

```
{vmnet,smtp,ntp_u,shell,kshell,aol,imap4,urh_i,netbios_ssn,tftp_u,mtp,uucp,nnsp,echo,tim_i,s
sh,iso_tsap,time,netbios_ns,systat,hostnames,login,efs,supdup,http_8001,courier,ctf,finger,nn
tp,ftp_data,red_i,ldap,http,ftp,pm_dump,exec,klogin,auth,netbios_dgm,other,link,X11,discard,
private,remote_job,IRC,daytime,pop_3,pop_2,gopher,sunrpc,name,rje,domain,uucp_path,http
_2784,Z39_50,domain_u,csnet_ns,whois,eco_i,bgp,sql_net,printer,telnet,ecr_i,urp_i,netstat,ht
tp_443,harvest}

@attribute flag {RSTR,S3,SF,RSTO,SH,OTH,S2,RSTOS0,S1,S0,REJ}

@attribute src_bytes numeric

@attribute dst_bytes numeric

@attribute logged_in {1,0}

@attribute count numeric

@attribute srv_count numeric

@attribute serror_rate numeric

@attribute srv_serror_rate numeric

@attribute same_srv_rate numeric

@attribute diff_srv_rate numeric

@attribute dst_host_count numeric

@attribute dst_host_srv_count numeric

@attribute dst_host_same_srv_rate numeric

@attribute dst_host_diff_srv_rate numeric

@attribute dst_host_same_src_port_rate numeric

@attribute dst_host_srv_diff_host_rate numeric

@attribute dst_host_serror_rate numeric

@attribute dst_host_srv_serror_rate numeric

@attribute label {dos,probe,u2r,u2l,normal}
```

Και το τμήμα δεδομένων του αρχείου ARFF που δημιουργήθηκε παρουσιάζεται παρακάτω:

@data

tcp,http,SF,181,5450,1,8,8,0,0,1,0,9,9,1,0,0.11,0,0,0,normal

tcp,http,SF,239,486,1,8,8,0,0,1,0,19,19,1,0,0.05,0,0,0,normal

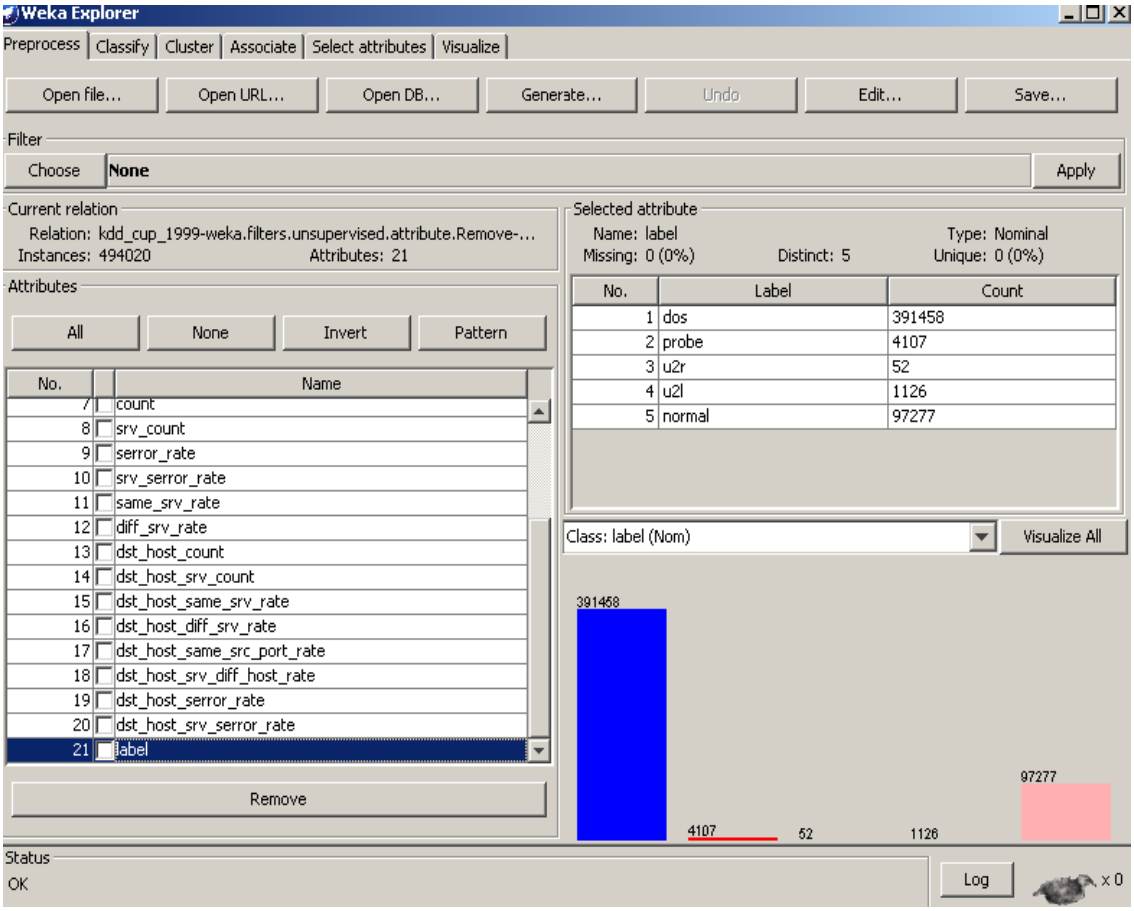
…..

…..

tcp,http,SF,291,1200,1,6,12,0,0,1,0,26,255,1,0,0.04,0.05,0.04,0.01,normal

tcp,http,SF,219,1234,1,6,35,0,0,1,0,6,255,1,0,0.17,0.05,0,0.01,normal

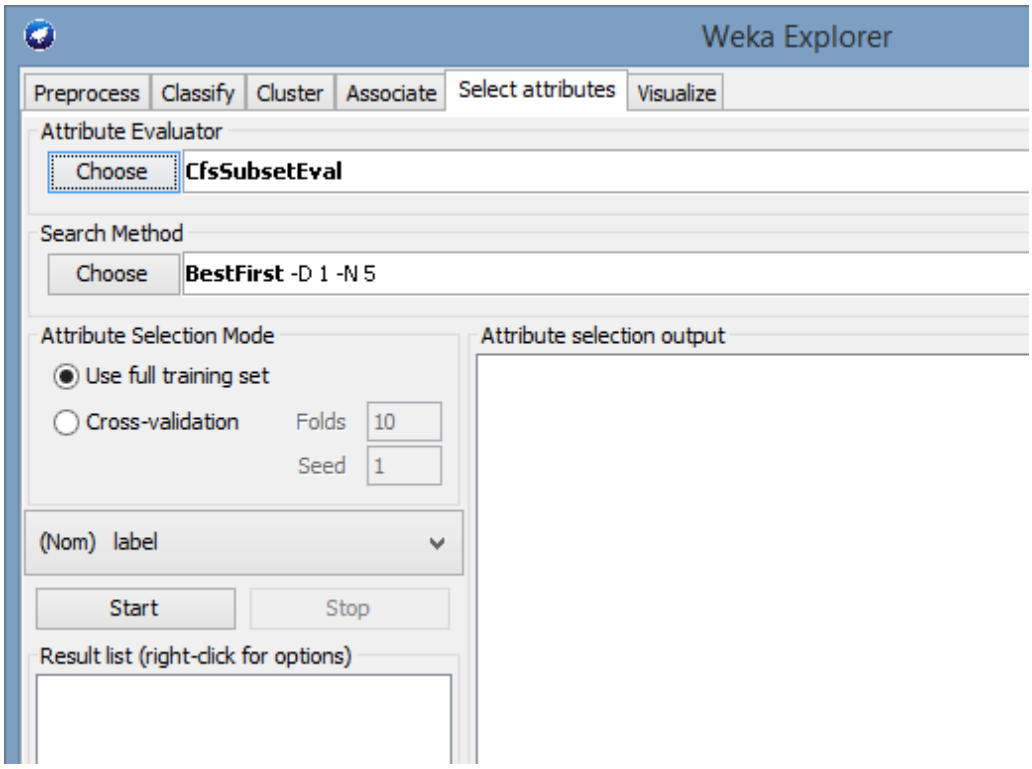
Στη συνέχεια φορτώνοντας το αρχείο ARFF στην εφαρμογή WEKA, εμφανίζονται πληροφορίες για τα χαρακτηριστικά του συνόλου δεδομένων:



Εικόνα 73: Φόρτωση αρχείου ARFF στην εφαρμογή WEKA

Από όπου διαπιστώνεται πως η πλειοψηφία των εγγραφών αναφέρονται σε επιθέσεις της κατηγορίας DOS (391.458), ακολουθούμενες από τις μη κακόβουλες συνδέσεις (97.277), ενώ οι άλλες τρεις κατηγορίες επιθέσεων (probe, u2r u2l) αντιπροσωπεύουν χαμηλό ποσοστό των εγγραφών του συνόλου δεδομένων.

Για τον προσδιορισμό των σημαντικών χαρακτηριστικών του συνόλου δεδομένων, αρχικά επιλέγουμε τη καρτέλα Select attribute, την μέθοδο αξιολόγησης χαρακτηριστικών *CfsSubsetEval*, ως μέθοδο αναζήτησης τη BestFirst, δηλώνοντας το χαρακτηριστικό *label* ως μεταβλητή κλάσης και χρησιμοποιώντας στην αναζήτηση των καλύτερων χαρακτηριστικών το σύνολο όλων των εγγραφών, πατάμε το κουμπί *Start* για τον προσδιορισμό των σχετικών χαρακτηριστικών:



Εικόνα 74: Καθορισμό παραμέτρων προσδιορισμού σημαντικών χαρακτηριστικών

Μετά την αξιολόγηση των χαρακτηριστικών από τη εφαρμογή, με βάση τους δηλωμένους (προεπιλεγμένες τιμές) παραμέτρους, προσδιορίζονται τα εξής τέσσερα (4) σημαντικά χαρακτηριστικά για το σύνολο δεδομένων KDD CUP 1999:

1. service
2. dst_bytes
3. logged_in

4. dst_host_count

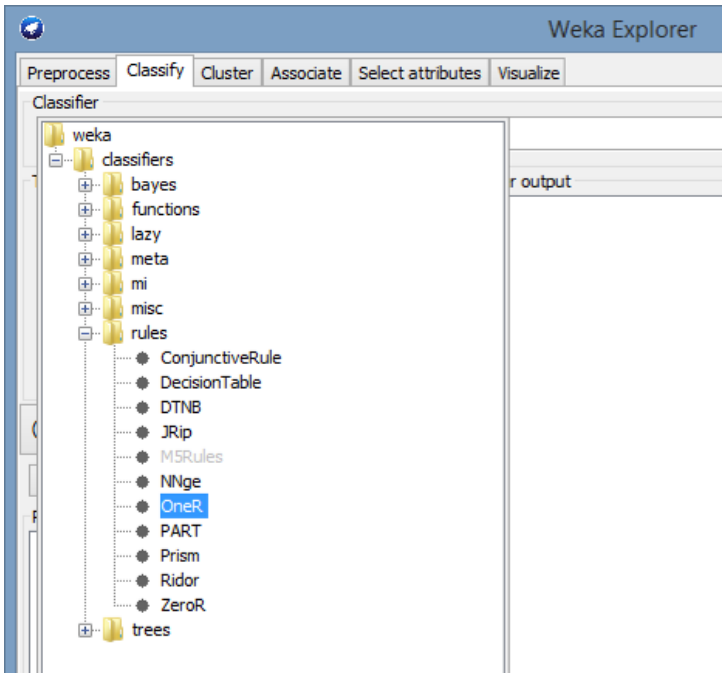
Αυτά τα τέσσερα χαρακτηριστικά κατά κύριο λόγο καθορίζουν αν μια σύνδεση, αντιστοιχεί σε κακόβουλη επίθεση ή σε φυσιολογική σύνδεση χρήστη.

6.6 Κατηγοριοποίηση – Πρόβλεψη (classification - prediction)

Συνολικά θα εφαρμόζονται στο σύνολο δεδομένων με τα 21 χαρακτηριστικά (συμπεριλαμβανομένης της μεταβλητής), δέκα (10) δημοφιλής αλγόριθμοι κατηγοριοποίησης. Πιο συγκεκριμένα θα εφαρμοστούν αλγόριθμοι των εξής κατηγοριών αλγορίθμων κατηγοριοποίησης (25), (26):

- Αλγόριθμοι κατηγοριοποίησης Bayesian (2)
- Αλγόριθμοι κατηγοριοποίησης META (2)
- Αλγόριθμοι παραγωγής κανόνων ταξινόμησης (3)
- Αλγόριθμοι με χρήση δέντρων αποφάσεων (3)

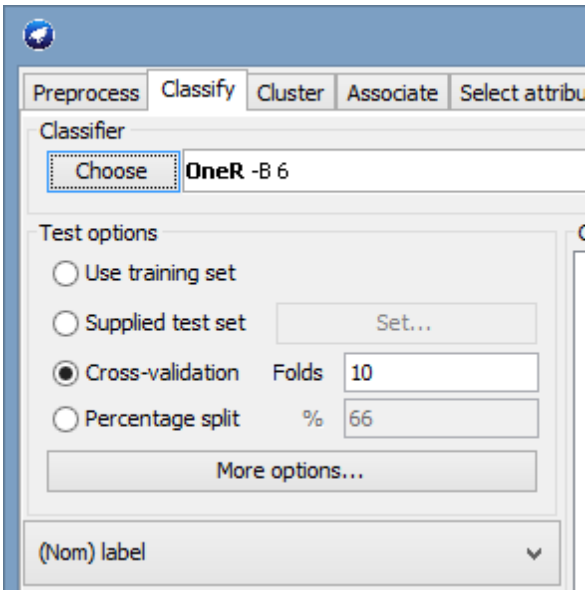
Για την επιλογή του αλγόριθμο που θα εκτελεστεί κάθε φορά, θα πρέπει αρχικά να επιλεγεί η καρτέλα Classify και στη συνέχεια επιλέγεται ο αλγόριθμος έχοντας πατήσει το κουμπί Choose και δηλώσει τον επιθυμητό αλγόριθμοι. Στο αντίστοιχο πλαίσιο διαλόγου οι αλγόριθμοι κατηγοριοποίησης εμφανίζονται ταξινομημένη ανά κατηγορία αλγορίθμων ταξινόμησης:



Εικόνα 75: Επιλογή αλγορίθμου κατηγοριποίησης

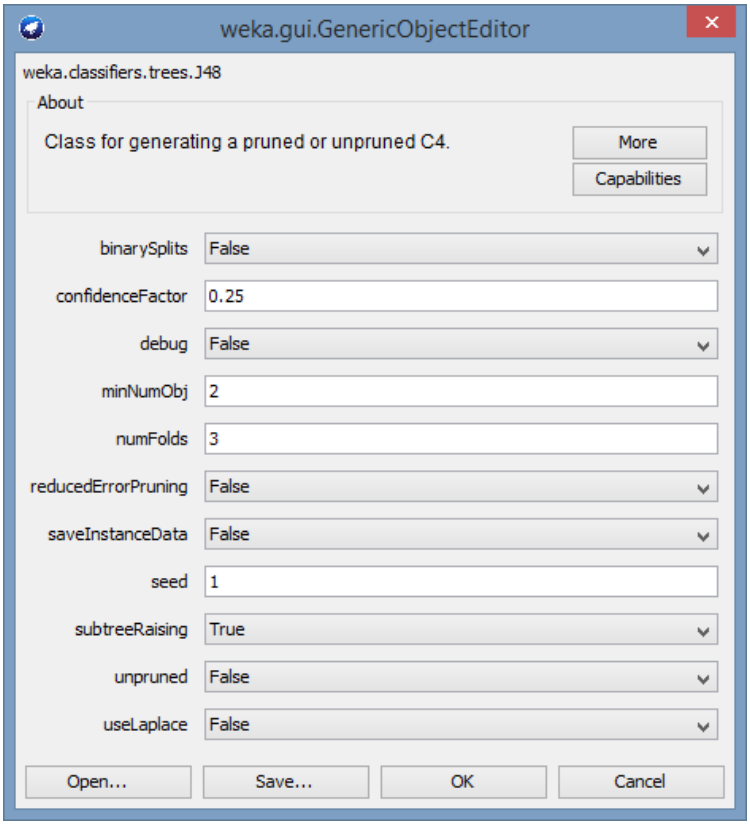
Εξαιτίας του μεγάλου αριθμού εγγραφών (σχεδόν 500.000 εγγραφές), παρά τον περιορισμό των χαρακτηριστικών του συνόλου δεδομένων που χρησιμοποιούνται ο χρόνος εκτέλεσης των αλγορίθμων είναι ιδιαίτερα υψηλός, ενδεικτικά για την εκτέλεση του αλγορίθμου Decision Table από την εφαρμογή WEKA σε υπολογιστή με 4GB RAM και 3.1 GHz, χρειάστηκαν πάνω από 30 λεπτά για την ολοκλήρωση εκτέλεσης του αλγορίθμου. Επίσης ενδεικτικά, η εκτέλεση του αλγορίθμου Bagging, χρειάστηκε πάνω από 60 λεπτά στην εφαρμογή WEKA. Για τον παραπάνω λόγο, αλλά και λόγω των ιδιαίτερα υψηλών ποσοστών σωστά ταξινομημένων εγγραφών κατά την εφαρμογή των αλγορίθμων, δεν έγιναν επιπλέον πειραματισμοί στην εκτέλεση των αλγορίθμων κατηγοριοποίησης με διαφοροποίησης στις τιμές των παραμέτρων τους.

Σε κάθε εκτέλεση αλγορίθμου θα χρησιμοποιηθούν οι προεπιλεγμένες τιμές παραμέτρων των αλγορίθμων και μέθοδο αναζήτησης 10 folds:



Εικόνα 76: Ρυθμίσεις συνόλου εκπαίδευσης και ελέγχου

Για πιθανή τροποποίηση των παραμέτρων ενός αλγορίθμου, θα πρέπει να γίνει κλικ στον τίτλο του αλγορίθμου για την εμφάνιση και τη τροποποίηση των σχετικών παραμέτρων:



Εικόνα 77: Παράμετροι αλγορίθμου κατηγοριοποίησης J48

Τα αποτελέσματα που λαμβάνονται για κάθε αλγόριθμο παρουσιάζονται στον παρακάτω πίνακα:

Πίνακας 5: Ποσοστό σωστά ταξινομημένων εγγραφών και F-measure ανά αλγόριθμο κατηγοριοποίησης που εφαρμόστηκε

Μεταβλητή κλάσης	BayesNet	NaiveBayes	Bagging	Filtered Classifier	Decision Table	PART	OneR	J48	Random Forest	Random Tree
Label	99.968 % F: 1	97.3556 % F: 0.978	99.951 % F: 1	99.9344 % F: 0.999	99.7644% F: 0.998	99.968 % F: 1	98.1667 % F: 0.98	99.9686 % F:	99.9737 % F: 1	99.9549 % F: 1

Όλοι οι αλγόριθμοι κατηγοριοποίησης που εξετάστηκαν, παρουσιάζουν ιδιαίτερα υψηλά ποσοστά ορθής ταξινόμησης εγγραφών και μπορούν να χρησιμοποιηθούν για την αξιολόγηση νέων συνδέσεων. Το μεγαλύτερο ποσοστό ορθής ταξινόμησης εγγραφών παρουσιάζει ο αλγόριθμος Random Forest, με τιμή F-measure ίση με την μονάδα (1) και ποσοστό σωστά ταξινομημένων εγγραφών 99.9737 %. Τα δέντρα αποφάσεων που προκύπτουν από τους παραπάνω αλγόριθμους είναι ιδιαίτερα περίπλοκα με αριθμό κόμβων που ξεπερνά τους 400 κόμβους, καθιστώντας δύσκολη την αναπαράστασή τους σε χαρτί. Ωστόσο, αν μια νέα σύνδεση ταξινομηθεί από τα αντίστοιχα δέντρα με βάση την τιμή στα χαρακτηριστικά της, με μεγάλο ποσοστό επιτυχίας θα αξιολογηθεί σωστά ως φυσιολογική σύνδεση ή ως κακόβουλη επίθεση. Το ίδιο ισχύει και για τις άλλες κατηγορίες αλγορίθμων ταξινόμησης, που παράγουν πολύπλοκα μοντέλα πρόβλεψης για την ορθή ταξινόμηση των εγγραφών. Ενδεικτικά ο απλούστερος μεταξύ των παραπάνω αλγορίθμων, ο ταξινομητής OneR, που παράγει μονάχα έναν κανόνα για τον χαρακτηρισμό των εγγραφών, ταξινομεί τις εγγραφές σύμφωνα με τον αριθμό bytes από την πηγή στον προορισμό, με βάση τον παρακάτω κανόνα ταξινόμησης:

```
src_bytes:
 < 0.5    -> dos
 < 2.5    -> probe
 < 7.5    -> normal
 < 8.5    -> probe
```

```
< 17.5   -> normal

< 18.5   -> probe

< 27.5   -> normal

< 28.5   -> dos

< 122.5 -> normal

< 126.5 -> u2l

< 333.5 -> normal

< 334.5 -> u2l

< 519.5 -> normal

< 520.5 -> dos

< 831.5 -> normal

< 832.5 -> u2l

< 1031.5          -> normal

< 1032.5          -> dos
```

```
< 1221.5        -> normal

< 1222.5        -> u2l

< 1223.5        -> normal

< 1224.5        -> u2l

< 1230.5        -> normal

< 1231.5        -> u2l

< 1232.5        -> normal

< 1233.5        -> u2l

< 1234.5        -> normal

< 1235.5        -> u2l

< 1236.5        -> normal

< 1237.5        -> u2l

< 1241.5        -> normal

< 1242.5        -> u2l
```

```
< 1479.5          -> normal
< 1480.5          -> dos
< 7033.5          -> normal
< 7061.0          -> u2l
< 35725.5         -> normal
< 39133.0         -> dos
< 40565.0         -> normal
< 54614.5         -> dos
< 2347338.5       -> normal
< 3.49255659E8            -> u2l
>= 3.49255659E8           -> probe

(485018/494020 instances correct)
```

Ακολουθούν οι πίνακες συσχέτισης για κάθε αλγόριθμο που εφαρμόστηκε στον πίνακα δεδομένων, από όπου διαφαίνεται (με εξαίρεση τον αλγόριθμο NaiveBayes) πως οι λανθασμένα ταξινομημένες εγγραφές αφορούν κυρίως τον λανθασμένο χαρακτηρισμό μιας σύνδεση ως φυσιολογική, ενώ στην πραγματικότητα πρόκειται για κακόβουλη σύνδεση:

Πίνακας 6: Πίνακας συσχέτισης BayesNet

```
=== Confusion Matrix ===

     a      b      c      d      e   <-- classified as
391444      5      0      1      8 |      a = dos
     9   4081      2      2     13 |      b = probe
     0      1     34      4     13 |      c = u2r
     1      2      8   1087     28 |      d = u21
    11     18     12     20  97216 |      e = normal
```

Πίνακας 7: Πίνακας συσχέτισης NaiveBayes

```
=== Confusion Matrix ===

     a      b      c      d      e   <-- classified as
385617   2750     98     12   2981 |      a = dos
    21   3591    276      4    215 |      b = probe
     1      0     40      2      9 |      c = u2r
     5     17    611    172    321 |      d = u21
   877   2343   2195    326  91536 |      e = normal
```

Πίνακας 8: Πίνακας συσχέτισης Bagging

```
=== Confusion Matrix ===

     a      b      c      d      e   <-- classified as
391423     25      0      3      7 |      a = dos
    57   4010      0      4     36 |      b = probe
     2      3     27      5     15 |      c = u2r
     1      2      6   1092     25 |      d = u21
    15      8      4     24  97226 |      e = normal
```

Πίνακας 9: Πίνακας συσχέτισης Filtered Classifier

```
=== Confusion Matrix ===

     a      b      c      d      e   <-- classified as
391417      6      0      0     35 |      a = dos
    79   3995      1      0     32 |      b = probe
     0      1     26      0     25 |      c = u2r
     2      0      0   1061     63 |      d = u21
    39     11      6     24  97197 |      e = normal
```

Πίνακας 10: Πίνακας συσχέτισης Decision Table

```
=== Confusion Matrix ===

      a      b      c      d      e   <-- classified as
 391438      8      0      0     12 |      a = dos
     58   4023      0      0     26 |      b = probe
      4      0      0      0     48 |      c = u2r
     35      0      0    879    212 |      d = u21
    428     26      0    307  96516 |      e = normal
```

Πίνακας 11: Πίνακας συσχέτισης PART

```
=== Confusion Matrix ===

      a      b      c      d      e   <-- classified as
 391444      5      0      1      8 |      a = dos
      9   4081      2      2     13 |      b = probe
      0      1     34      4     13 |      c = u2r
      1      2      8   1087     28 |      d = u21
     11     18     12     20  97216 |      e = normal
```

Πίνακας 12: Πίνακας συσχέτισης OneR

```
=== Confusion Matrix ===

      a      b      c      d      e   <-- classified as
 391431      0      0      0     27 |      a = dos
   2551   1512      0      1     43 |      b = probe
     12      0      0      0     40 |      c = u2r
     31      0      0    743    352 |      d = u21
   5570    127      0    303  91277 |      e = normal
```

Πίνακας 13: Πίνακας συσχέτισης J48

```
=== Confusion Matrix ===

      a      b      c      d      e   <-- classified as
 391444      1      0      2     11 |      a = dos
      2   4083      2      2     18 |      b = probe
      0      0     36      5     11 |      c = u2r
      6      0      3   1079     38 |      d = u21
     12      7     12     23  97223 |      e = normal
```

Πίνακας 14: Πίνακας συσχέτισης Random Forest

```
=== Confusion Matrix ===

     a      b      c      d      e   <-- classified as
391456      0      0      0      2 |      a = dos
    29   4055      0      0     23 |      b = probe
     0      0     40      2     10 |      c = u2r
     0      0      8   1105     13 |      d = u2l
     8      7     11     17  97234 |      e = normal
```

Πίνακας 15: Πίνακας συσχέτισης Random Tree

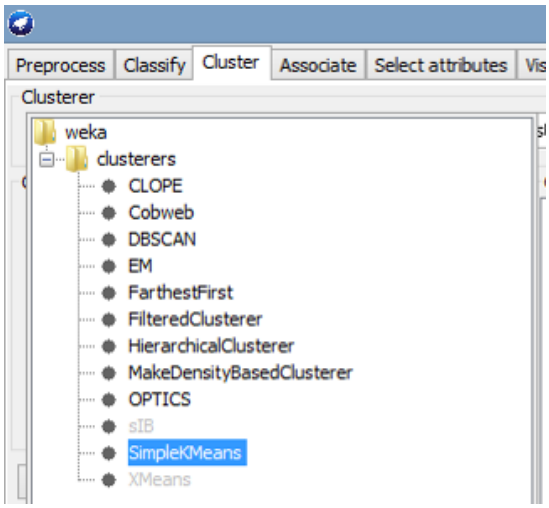
```
=== Confusion Matrix ===

     a      b      c      d      e   <-- classified as
391427     14      0      0     17 |      a = dos
    35   4047      0      1     24 |      b = probe
     0      1     33      5     13 |      c = u2r
     1      2      7   1088     28 |      d = u2l
    23     15     11     26  97202 |      e = normal
```

Από τις παραπάνω εφαρμογές αλγορίθμων, διαπιστώνεται πως αν περιοριστεί η κατηγοριοποίηση στα χαρακτηριστικά που προέκυψαν κατά την αρχική προπαρασκευή των δεδομένων (αφαίρεση χαρακτηριστικών με πολύ μεγάλο ποσοστό κοινών τιμών στις περισσότερες εγγραφές του συνόλου δεδομένων), προκύπτουν ιδιαίτερα υψηλά ποσοστά σωστά ταξινομημένων εγγραφών, κατά την εφαρμογή των διαφόρων αλγορίθμων κατηγοριοποίησης. Αυτό σημαίνει πως εφαρμόζοντας τις τιμές μια καινούργιας σύνδεσης στο μοντέλο που προκύπτει από κάθε αλγόριθμο κατηγοριοποίησης, μπορεί να προβλεφθεί με επιτυχία (με χαμηλό ποσοστό σφάλματος) αν η σύνδεση αποτελεί προσπάθεια επίθεση ή όχι.

6.7 Συσταδοποίηση(clustering)

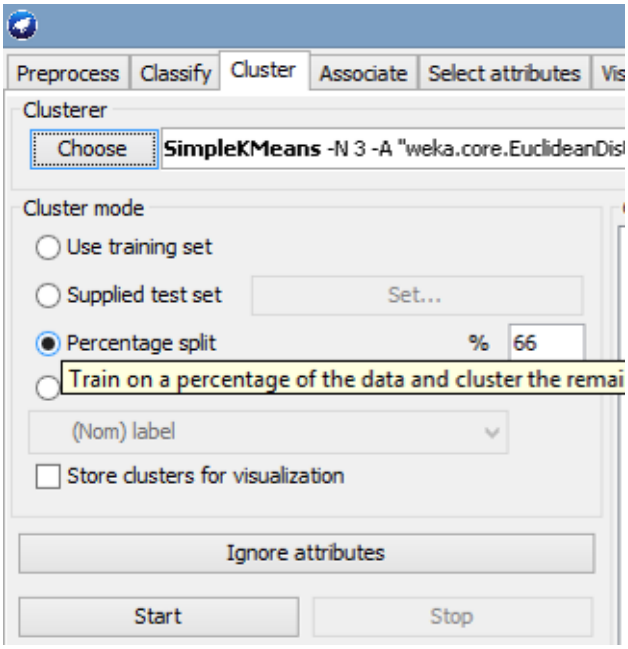
Το σύνολο δεδομένων που θα χρησιμοποιηθεί για τη συσταδοποίηση είναι το σύνολο δεδομένων με όλα τα χαρακτηριστικά εκφρασμένα σε κατηγοριακή μορφή που προέκυψε κατά τη σχεδίαση αποθήκης δεδομένων. Προτιμάται το συγκεκριμένο σύνολο δεδομένων σε σχέση με το σύνολο δεδομένων που χρησιμοποιήθηκε κατά την εφαρμογή αλγορίθμων κατηγοριοποίησης (αποτελούνται από τα ίδια χαρακτηριστικά), για να είναι εκφρασμένες οι τιμές χαρακτηριστικών ανά κατηγορία επίθεσης ή με τις ακραίες τιμές ενός χαρακτηριστικού σε κατηγοριακή μορφή (εύρος τιμών) και όχι αριθμητικός μέσος.

Για την υλοποίηση συσταδοποίησης, θα εφαρμοστεί στα δεδομένα ο αλγόριθμος SimpleKMeans:



Εικόνα 78: Επιλογή αλγόριθμου συσταδοποίησης

Με επιλογή χρήσης ως δείγμα εκπαίδευσης το 66% των εγγραφών του συνόλου δεδομένων:



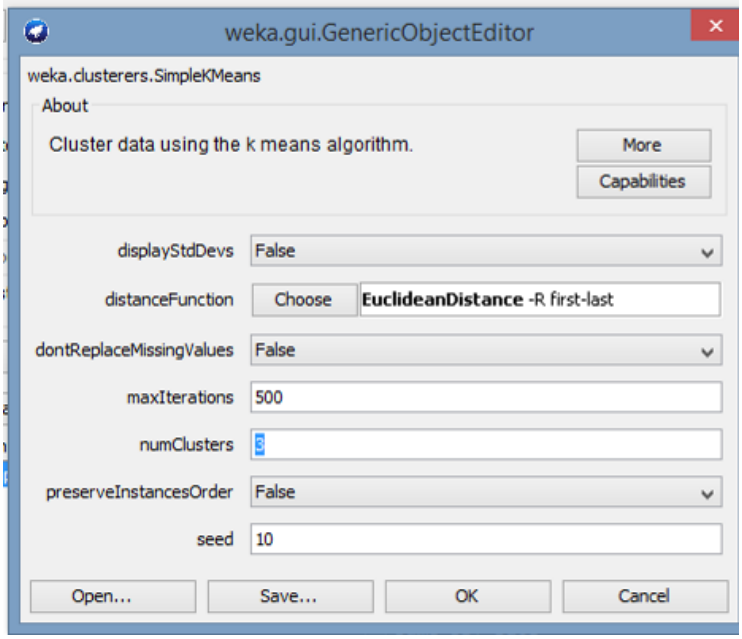
Εικόνα 79: Καθορισμός δείγματος εκπαίδευσης συσταδοποίησης

Ο προσδιορισμός του κατάλληλου αριθμού συστάδων που θα πρέπει να δημιουργηθούν, θα προσδιοριστεί με την βοήθεια της τιμής του Αθροίσματος Τετραγωνικού Σφάλματος (SSE - Sum of Squared Error) για κάθε δοκιμαστικό αριθμό συστάδων. Η τιμή του SSE για κάθε δοκιμαστικό αριθμό συστάδων παρουσιάζεται στον παρακάτω πίνακα:

Πίνακας 16: SSE ανά αριθμό συστάδων

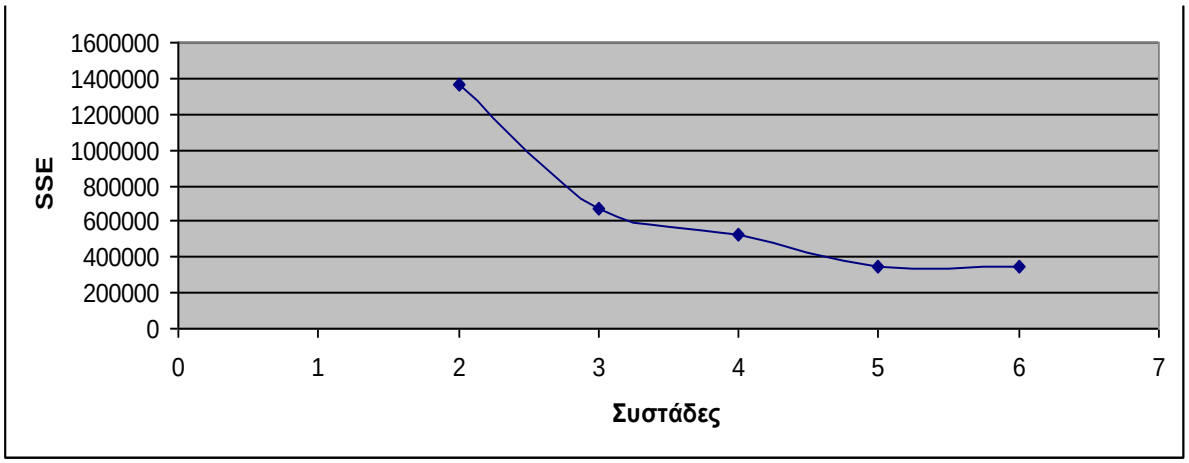
Αριθμός συστάδων	SSE
2	1367946
3	668454
4	520427
5	347892
6	344954
7	295018

Ο αριθμός των συστάδων που καλείται να δημιουργήσει σε κάθε εκτέλεση ο αλγόριθμος Apriori, ρυθμίζεται από τις παραμέτρους του αλγόριθμου και πιο συγκεκριμένα από τη παράμετρο numClusters:



Εικόνα 80: Καθορισμός αριθμού συστάδων αλγόριθμου Apriori

Ο ιδανικός αριθμός συστάδων για ένα σύνολο δεδομένων, αντιστοιχεί στο σημείο που η καμπύλη απεικόνισης του SSE σε σχέση με τον σχετικό αριθμό συστάδων, σχηματίζει κούρμπα και μειώνεται ο ρυθμός μείωσης του SSE. Η σχετική καμπύλη απεικόνισης της μεταβολής του SSE, καθώς αυξάνεται ο αριθμός συστάδων παρουσιάζεται στο παρακάτω Εικόνα :



Εικόνα 1: Ιστόγραμμα SSE ανά αριθμό συστάδων

Από το παραπάνω Εικόνα , διαπιστώνεται πως η καμπύλη εμφανίζει καμπή στη μετακίνηση του SSE από αριθμό συστάδων 3 σε αριθμό συστάδων 4. Επίσης, θα μπορούσε να θεωρηθεί πως και στο σημείο που αντιστοιχεί σε αριθμό συστάδων 5 η καμπύλη παρουσιάζει εκ νέου καμπή, συνεπώς εφαρμόζεται συσταδοποίηση και για τους δυο πιθανούς αριθμούς συστάδων (3, 5).

Στον παρακάτω πίνακα, παρουσιάζονται οι τιμές χαρακτηριστικών για κάθε μια από τις τρεις συστάδες που προέκυψαν από την εφαρμογή του αλγορίθμου Apriori:

Πίνακας 17: Τιμές χαρακτηριστικών συστάδων (3 συστάδες)

		Cluster#		
Attribute	Full Data	0	1	2
	(494021)	(112187)	(99325)	(282509)
protocol_type	icmp	tcp	tcp	icmp
service	ecr_i	private	http	ecr_i
flag	SF	S0	SF	SF
src_bytes	medium	very low	low	medium
dst_bytes	very low	very low	very high	very low
logged_in	0	0	1	0
count	other	other	other	511.0
srv_count	other	other	other	511.0
serror_rate	0.0	100.0	0.0	0.0
srv_serror_rate	0.0	100.0	0.0	0.0
same_srv_rate	100.0	other	100.0	100.0
diff_srv_rate	0.0	other	0.0	0.0
dst_host_count	255.0	255.0	other	255.0
dst_host_srv_count	255.0	other	255.0	255.0
dst_host_same_srv_rate	100.0	other	100.0	100.0
dst_host_diff_srv_rate	0.0	other	0.0	0.0
dst_host_same_src_port_rate	100.0	0.0	other	100.0
dst_host_srv_diff_host_rate	0.0	0.0	other	0.0
dst_host_serror_rate	0.0	100.0	0.0	0.0
dst_host_srv_serror_rate	0.0	100.0	0.0	0.0
label	dos	dos	normal	dos

Στον παρακάτω πίνακα, παρουσιάζονται οι τιμές χαρακτηριστικών για κάθε μια από τις πέντε συστάδες που προέκυψαν από την εφαρμογή του αλγορίθμου Apriori:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(326053)	(186498)	(3096)	(24823)	(57535)	(54101)
protocol_type	icmp	icmp	tcp	tcp	tcp	tcp
service	ecr_i	ecr_i	private	private	http	private
flag	SF	SF	S0	REJ	SF	S0
src_bytes	medium	medium	very low	very low	low	very low
dst_bytes	very low	very low	very low	very low	very high	very low
logged_in	0	0	0	0	1	0
count	other	511.0	other	other	other	other
srv_count	other	511.0	other	other	other	other
serror_rate	0.0	0.0	100.0	0.0	0.0	100.0
srv_serror_rate	0.0	0.0	100.0	0.0	0.0	100.0
same_srv_rate	100.0	100.0	other	other	100.0	other
diff_srv_rate	0.0	0.0	other	other	0.0	other
dst_host_count	255.0	255.0	255.0	255.0	other	255.0
dst_host_srv_count	255.0	255.0	other	other	255.0	other
dst_host_same_srv_rate	100.0	100.0	0.0	other	100.0	other
dst_host_diff_srv_rate	0.0	0.0	other	other	0.0	other
dst_host_same_src_port_rate	100.0	100.0	0.0	0.0	other	0.0
dst_host_srv_diff_host_rate	0.0	0.0	0.0	0.0	other	0.0
dst_host_serror_rate	0.0	0.0	100.0	0.0	0.0	100.0
dst_host_srv_serror_rate	0.0	0.0	100.0	0.0	0.0	100.0
label	dos	dos	dos	dos	normal	dos

Πίνακας 18: Τιμές χαρακτηριστικών συστάδων (5 συστάδες)

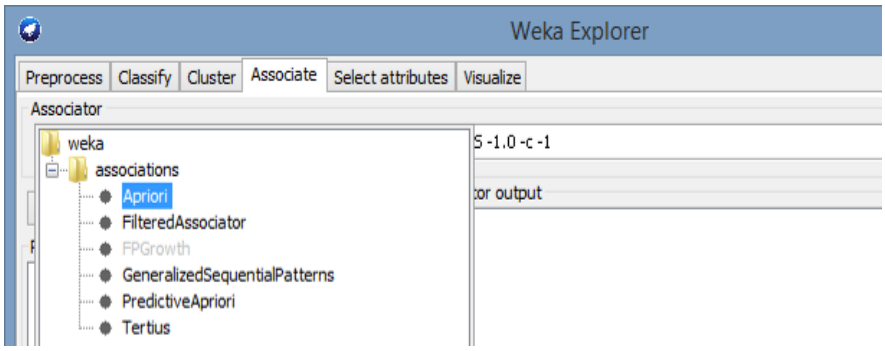
Παρατηρώντας τις τιμές των χαρακτηριστικών και για τις δυο περιπτώσεις συσταδοποίησης, διαπιστώνουμε πως είναι σε απόλυτη συμφωνία μεταξύ τους:

- στις συνδέσεις που θεωρούνται ασφαλείς, ο χρήστης έχει πρότερα συνδεθεί με επιτυχία (πεδίο logged_in)
- όταν το πεδίο dst_host_same_src_port_rate (% συνδέσεων με ίδιο source port) κυμαίνεται σε ενδιάμεσε τιμές και όχι σε ακραίες τιμές (min, max), πρόκειται για φυσιολογική σύνδεση
- όταν το πεδίο dst_host_srv_diff_host_rate (% συνδέσεων σε άλλο host) κυμαίνεται σε ενδιάμεσε τιμές και όχι σε ακραίες τιμές (min, max), πρόκειται για φυσιολογική σύνδεση

Για τα υπόλοιπα χαρακτηριστικά σύμφωνα με τα αποτελέσματα συσταδοποίησης δεν παρατηρείται ιδιαίτερη διαφοροποίηση μεταξύ κακόβουλων συνδέσεων και φυσιολογικών συνδέσεων.

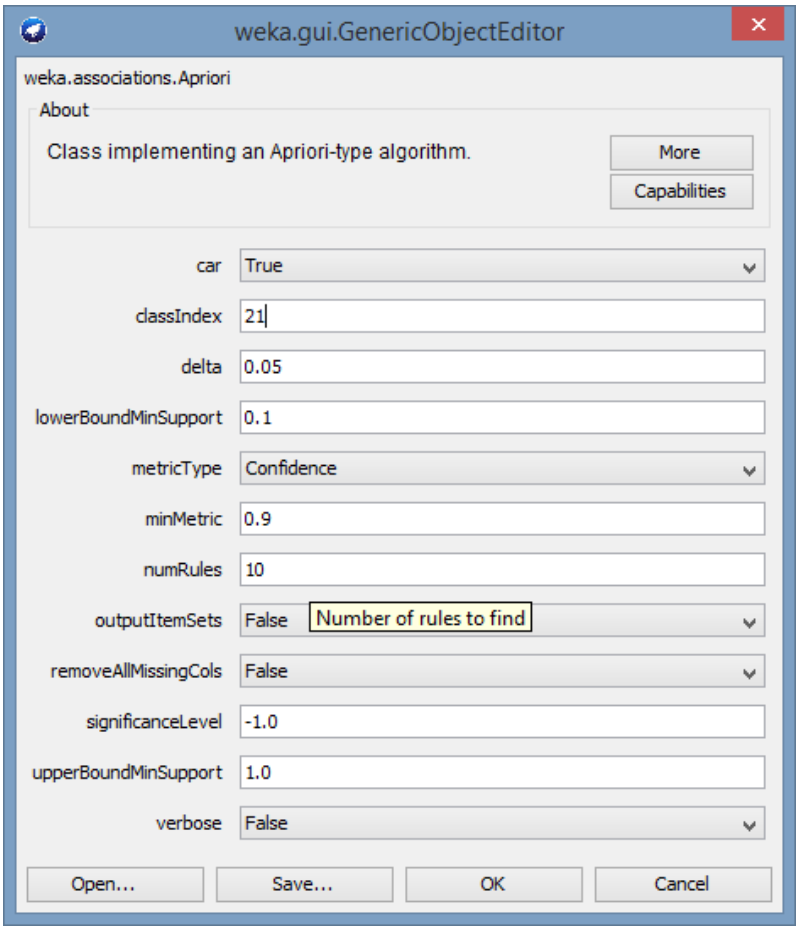
6.8 Κανόνες συσχέτισης (association mining)

Το σύνολο δεδομένων που θα χρησιμοποιηθεί, είναι το ίδιο με εκείνο που χρησιμοποιήθηκε κατά την εφαρμογή συσταδοποίησης στα δεδομένα, καθώς ο αλγόριθμος που θα χρησιμοποιηθεί (Apriori), απαιτεί τις τιμές των χαρακτηριστικών να είναι εκφρασμένες σε κατηγοριακή μορφή (11) (16):



Εικόνα 81: Επιλογή αλγορίθμου Apriori

Για την εξαγωγή κανόνων συσχέτισης θα εφαρμοστεί ο αλγόριθμος Apriori, με την επιλογή, τη δημιουργία κανόνων συσχέτισης για την εξαρτημένη μεταβλητή (label) η οποία έχει δείκτη 21 στο σύνολο δεδομένων arff. Για να δημιουργηθούν κανόνες συσχέτισης για την εξαρτημένη μεταβλητή label θα πρέπει να θέσουμε τη τιμή της μεταβλητής car ως αληθές (true):



Εικόνα 82: Ρύθμιση παραμέτρων εκτέλεσης αλγορίθμου Apriori

Οι δέκα κανόνες συσχέτισης (με τον υψηλότερο δείκτη εμπιστοσύνης) για την πρόβλεψη του είδους επίθεσης που αντιστοιχεί σε μια σύνδεση είναι οι εξής:

1. dst_bytes=very low logged_in=0 serror_rate=0.0 dst_host_count=255.0 dst_host_serror_rate=0.0 dst_host_srv_serror_rate=0.0 306123 ==> label=dos 301623 conf:(0.99)

2. dst_bytes=very low logged_in=0 serror_rate=0.0 srv_serror_rate=0.0 dst_host_count=255.0 dst_host_serror_rate=0.0 dst_host_srv_serror_rate=0.0 306123 ==> label=dos 301623 conf:(0.99)

3. dst_bytes=very low logged_in=0 serror_rate=0.0 dst_host_count=255.0 dst_host_srv_diff_host_rate=0.0 dst_host_serror_rate=0.0 dst_host_srv_serror_rate=0.0 306123 ==> label=dos 301623 conf:(0.99)

4. dst_bytes=very low logged_in=0 serror_rate=0.0 srv_serror_rate=0.0 dst_host_count=255.0 dst_host_srv_diff_host_rate=0.0 dst_host_serror_rate=0.0 dst_host_srv_serror_rate=0.0 306123 ==> label=dos 301623 conf:(0.99)

5. dst_bytes=very low logged_in=0 serror_rate=0.0 dst_host_count=255.0 dst_host_serror_rate=0.0 306127 ==> label=dos 301623 conf:(0.99)

6. dst_bytes=very low logged_in=0 serror_rate=0.0 srv_serror_rate=0.0 dst_host_count=255.0 dst_host_serror_rate=0.0 306127 ==> label=dos 301623 conf:(0.99)

7. dst_bytes=very low logged_in=0 serror_rate=0.0 dst_host_count=255.0 dst_host_srv_diff_host_rate=0.0 dst_host_serror_rate=0.0 306127 ==> label=dos 301623 conf:(0.99)

8. dst_bytes=very low logged_in=0 serror_rate=0.0 srv_serror_rate=0.0 dst_host_count=255.0 dst_host_srv_diff_host_rate=0.0 dst_host_serror_rate=0.0 306127 ==> label=dos 301623 conf:(0.99)

9. dst_bytes=very low logged_in=0 serror_rate=0.0 dst_host_count=255.0 dst_host_srv_serror_rate=0.0 306437 ==> label=dos 301847 conf:(0.99)

10. dst_bytes=very low logged_in=0 serror_rate=0.0 srv_serror_rate=0.0 dst_host_count=255.0 dst_host_srv_serror_rate=0.0 306437 ==> label=dos 301847 conf:(0.99)

Παρατηρώντας τους κανόνες συσχέτισης, διαπιστώνεται πως όλοι τους αναφέρονται σε συνθήκες που μια σύνδεση αποτελεί επίθεση της κατηγορίας επιθέσεων dos. Ο λόγος που συμβαίνει αυτό, είναι πως οι αντίστοιχοι κανόνες έχουν μεγαλύτερο βαθμός κάλυψης, καθώς οι εγγραφές που αντιστοιχούν σε επίθεση της κατηγορίας dos, αποτελούν την συντριπτική πλειοψηφία του συνόλου δεδομένων.

Οι κανόνες συσχέτισης για κακόβουλες επιθέσεις (κατηγορίας dos), παρουσιάζουν κοινά χαρακτηριστικά με τα αποτελέσματα συσταδοποίησης. Παρατηρούμε πως βασικό κριτήριο για να αντιστοιχία μια σύνδεση σε κακόβουλη επίθεση αποτελεί να μην έχει συνδεθεί με επιτυχία στο σύστημα (is_logged=0) και οι τιμές των πεδίων dst_host_same_src_port_rate και dst_host_srv_diff_host_rate να αντιστοιχούν σε ακραίες τιμές (min, max). Επίσης, σημάδι πως μια σύνδεση είναι πιθανότατα κακόβουλη, αποτελεί ο αριθμός των bytes δεδομένων από τον προορισμό στην πηγή να είναι μηδενικός ή ιδιαίτερα χαμηλός (dst_bytes). Επίσης, οι κακόβουλες επιθέσεις δεν τείνουν να προκαλούν σφάλματα σύνδεσης (SYN error, REJ error) κατά την προσπάθεια σύνδεσής τους με τον προορισμό (πεδία: serror_rate, srv_serror_rate, dst_host_serror_rate, dst_host_srv_serror_rate). Τέλος, ο αριθμός των συνδέσεων στον ίδιο host με την τρέχουσα σύνδεση κατά τα τελευταία δύο δευτερόλεπτα είναι ιδιαίτερα αυξημένος (dst_host_count) και ίσος με τη μέγιστη τιμή αριθμού συνδέσεων.

7 Συμπεράσματα

Από την εφαρμογή αλγορίθμων κατηγοριοποίησης, προκύπτουν μοντέλα πρόβλεψης (πολύπλοκα και με πολυάριθμους κανόνες), τα οποία ωστόσο παρουσιάζουν ιδιαίτερα υψηλή αποτελεσματικότητα πρόβλεψης κακόβουλων συνδέσεων. Η εφαρμογή σε εκείνα, των στοιχείων – χαρακτηριστικών μια νέας σύνδεσης, μπορεί με ιδιαίτερα υψηλό ποσοστό αξιοπιστίας (μεγαλύτερο του 99.5%) να αναγνωρίσει μια κακόβουλη επίθεση.

Επίσης, η συσταδοποίησης των εγγραφών του συνόλου δεδομένων, εντόπισε χαρακτηριστικά (is_logged, dst_host_same_src_port_rate και dst_host_srv_diff_host_rate), των οποίων οι τιμές συνηγορούν αξιόπιστα υπέρ ή κατά του χαρακτηρισμού μιας σύνδεσης ως κακόβουλης.

Επιπροσθέτως, οι κανόνες συσχέτισης που προέκυψαν, συμφωνούν ως προς την βαρύτητα των πεδίων που αναφέρθηκα παραπάνω σχετικά με την αξιολόγηση μιας καινούργιας σύνδεσης ως κακόβουλης ή όχι, καθώς επίσης και πρόσθεσαν χαρακτηριστικά, των οποίων συγκεκριμένες τιμές συνηγορούν υπέρ του χαρακτηρισμού μιας σύνδεσης ως κακόβουλη επίθεση.

Τέλος, η υλοποίηση αποθήκης δεδομένων για τα χαρακτηριστικά του δείγματος που δεν παρουσιάζουν μεγάλη ομοιότητα στις τιμές τους για το σύνολο των εγγραφών (μεγαλύτερο του 90%) του δείγματος, επιτρέπουν μια επιχείρηση να κατανοήσει καλύτερα τα χαρακτηριστικά των συνδέσεων και πιθανές συσχετίσεις μεταξύ των χαρακτηριστικών που συνηγορούν υπέρ του χαρακτηρισμού μιας σύνδεσης ως επίθεση.

8 Βιβλιογραφικές Πηγές

1. **UCI ML Repository.** ***KDD Cup 1999 Data Set.*** **[Ηλεκτρονικό] 1999. https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/kddcup99.html.**
2. William, Stalling. ***Βασικές Αρχές Ασφάλειας Δικτύων.*** **3η αμερικανική έκδοση. s.l. : Κλειδαριθμος, 2011.**
3. Γκρίτζαλη, Κάτσικας Σωκρ. - Δ. Γκρίτζαλης - Στεφ. ***Ασφάλεια Πληροφοριακών Συστημάτων.***
4. Σουρής Ανδρ., Πατσός Δ., Γρηγοριάδης Ν. ***Ασφάλεια της Πληροφορίας.*** **s.l. : Εκδόσεις Νέων Τεχνολογιών, 2004.**
5. Πολέμη Ν., Κοτζανικολάου Π.. ***Τεχνολογίες και Πολιτικές Ασφάλειας.*** **Πειραιάς : s.n., 2007.**
6. Πολέμη Νινέτα, Αλέξανδρος Καλιοντζόγλου. ***Πρακτικά Θέματα Ασφάλειας Πληροφοριακών Συστημάτων & Εφαρμογών.*** **1η Έκδοση. Αθήνα : Εκδόσεις Νέων Τεχνολογιών, 2008. σ. 228.**
7. Λάζος, Γρ. ***Πληροφορική και Έγκλημα.*** **s.l. : Νομική Βιβλιοθήκη, 2001.**
8. Μάγκος, Εμμ. ***Ασφάλεια Υπολογιστών και Προστασία Δεδομένων.*** **Πληροφορική , Ιόνιο Πανεπιστήμιο . Κέρκυρα : s.n., 2011.**
9. Νάκος, Γρ. Ηλεκτρονική Βιβλιοθήκη. ***Κεφάλαιο 3. "Ασφάλεια στα Δίκτυα Υπολογιστών".*** **[Ηλεκτρονικό] http://greg61.gr/blog/%CE%B7-%CE%B3%CF%89%CE%BD%CE%B9%CE%AC-%CF%84%CE%BF%CF%85-%CF%85%CF%80%CE%BF%CE%BB%CE%BF%CE%B3%CE%B9%CF%83%CF%84%CE%AE-%CE%BA%CE%B1%CE%B9-%CF%84%CE%BF%CF%85-%CE%B4%CE%B9%CE%B1%CE%B4%CE%B9%CE%BA%CF%84/%CE%B1%CF%83%CF%86%CE%AC%CE%BB%CE%.**
10. Μανωλόπουλος Ι., Νανόπουλος Αλ. ***"Εισαγωγή στην Εξόρυξη Δεδομένων και τις Αποθήκες Δεδομένων".*** **s.l. : Εκδόσεις Νέων Τεχνολογιών, 2009.**
11. Roiger R.G., Geatz M.W. ***"Εξόρυξη Πληροφορίας – Ένας Εισαγωγικός Οδηγός με Παραδείγματα". Εκδόσεις Κλειδάριθμος.*** **s.l. : Εκδόσεις Κλειδάριθμος, 2008.**
12. Dunham, M. H. ***"Data Mining – Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα".*** **s.l. : Εκδόσεις Νέων Τεχνολογιών, 2004.**
13. Frawley W., Piatetsky-Shapiro G. , Matheus C. ***"Knowledge Discovery in Databases - An Overview".*** **1991. σσ. 1-30.**
14. Αθ., Μπαμπαλιάρης. ***Εξόρυξη Γνώσης από Βάσεις Δεδομένων .*** **Καβάλα : s.n., 2011.**
15. 15. Brachman, R. J., Khabaza, T., Kloesgen, W., Pieatetsky-Shapiro, G., and Simoudis, E. ***Mining Business Databases. Communications of the ACM, 39, 11,42-8.*** **1996.**
16. Agrawal, R., Imielinski, T., and Swami A. ***Mining Association Rules Between Sets of Items in Large Databases. In Buneman P. and Jajordia S. , eds., Proceedings of the ACM Sigmoid International Conference on Man-agement of Data.*** **New York : ACM, 1993.**
17. Cox, E. Free-Form Text Data Mining Integrating Fuzzy Systems, Self-Organizing Neural Nets and Rule-Based Knowledge Bases. ***PC AI.*** **2000, September-October, σσ. 22-26.**
18. Chester, M. ***Neural Networks—A Tutorial.*** **Upper Saddle River, NJ : PTR Prentice Hall, 1993.**

19. Chaudhuri S., Dayal U. ***'Data warehousing and OLAP for Decision Support'.***
20. Red Brick Systems Inc. ***Red Brick Warehouse 5.0.*** **[Ηλεκτρονικό] 1997. http://www.redbrick.com/rbsg/html/whouse50.html .**
21. Inmon, W. H. ***Building the Data Warehouse.*** **second edition. s.l. : John Wiley & Sons, 1996.**
22. Chaudhuri S., Dayal. U. ***An Overview of Data Warehousing and OLAP.***
23. Smyth, Stephen D. Bay and Dennis F. Kibler and Michael J. Pazzani and Padhraic. The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation. SIGKDD Explorations, 2. 2000.
24. Stolfo, J., Fan, W., Lee, W., Prodromidis, A., & Chan, P. K. . Cost-based modeling and evaluation for data mining with application to fraud and intrusion detection. Results from the JAM Project by Salvatore. ***UCI Machine Learning Repository: KDD Cup 1999 Data Data Set.*** **[Ηλεκτρονικό] 2000. http://rexa.info/paper/fa81d2810f9e8900d92eac1c4292291840b2f7a7.**
25. Breiman, L., Friedman, J., Olshen, R., and Stone C. ***Classification and Regression Trees.*** **Monterey, CA : Wadsworth International Group, 1984.**
26. Breiman, L. ***Bagging Predictors. Machine Learning.*** **1996. σσ. 123-140.**
27. Γκρίτζαλης Στέφανος, Λαμπρινουδάκης Κωνσταντίνος, Κάτσικας Σωκράτης, Μήτρου Λ. ***Προστασία της Ιδιωτικότητας & Τεχνολογίες Πληροφορικής και Επικοινωνιών.***
28. Lee W., & Stolfo, S. J. ***A framework for constructing features and models for intrusion detection systems. ACM transactions on Information and system security (TiSSEC), 3(4).*** **2000. σσ. 227-261.**
29. Gill, H. S., and Rao, P. C. ***The Official Guide to Data Warehousing.*** **Indianapolis, IN : Que Publishin, 1996.**
30. Kimball, R., Reeves, L., Ross, M., and Thomthwaite, W. ***18. Kimball, R., Reeves, L., Ross, M., and Thomthwaite, W. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses. New York: John Wiley & Sons.*** **New York : John Wiley & Sons, 1998.**
31. D., Byard J. and Schneider. ***The Ins & Outs (and everything in between) of Data Warehousing. Tutorials of ACM SIGMOD International Conference on Management of Data.*** **Montreal : s.n., 1996.**
32. An Introduction to Multidimensional Database Technology. Kenan Technologies. [Ηλεκτρονικό] 1995. http://www.kenan.com/content/compinfo/whitepapers/white.htm .
33. Case, S., Azarmi, N., Thint, M., and Ohtani, T. Enhancing E-Communities with Agent-Based Systems. ***Computer.*** **2001, July, σσ. 64-69.**
34. Chester, M. ***Neural Networks—A Tutorial.*** **Upper Saddle River, NJ : PTR Prentice Hall, 1993.**
35. Dasgupta, A., and Raftery, A. E. Detecting Features in Spatial Point Processes with Clutter via Model-based Clustering. ***Journal of the American Statistical Association.*** **1998, 93, σσ. 294-302.**
36. Pang - Tan Ning, Steinbach Michael, Kumar Vipin. ***Introduction to Data Mining.*** **s.l. : Pearson International Edition, 2006.**
37. Σ., Λιγουδιστιάνος. Αποθήκες Δεδομένων. 8, σσ. 195-210.
38. Βαζιργιάννης, Μιχάλης , Χαλκίδη, Μαρία. ***Εξόρυξη γνώσης από βάσεις δεδομένων.***