Εννοιολογικό Μοντέλο για Διεργασίες Εξαγωγής, Μετασχηματισμού και Φόρτωσης Δεδομένων

Άλκης Σιμιτσής
Εθνικό Μετσόβιο Πολυτεχνείο,
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Τομέας Πληροφορικής,
Ηρώων Πολυτεχνείου 9,
15772,Αθήνα,Ελλάδα
asimi@dblab.ntua.gr

Πάνος Βασιλειάδης
Πανεπιστήμιο Ιωαννίνων,
Τμήμα Πληροφορικής,
45110,Ιωάννινα,Ελλάδα
pvassil@cs.uoi.gr

Τίμος Σελλής
Εθνικό Μετσόβιο Πολυτεχνείο,
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Τομέας Πληροφορικής,
Ηρώων Πολυτεχνείου 9,
15772,Αθήνα,Ελλάδα
timos@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ

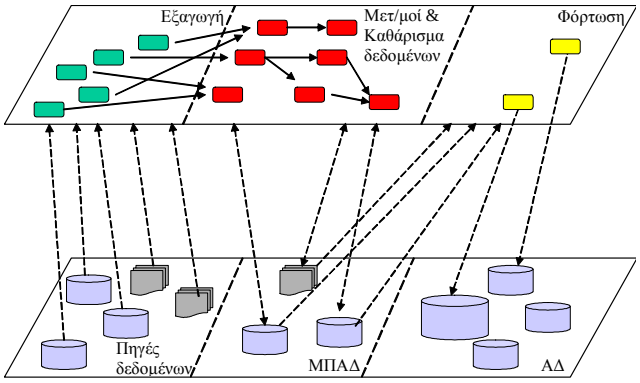
Το άρθρο αυτό αφορά στο εννοιολογικό σχεδιασμό σεναρίων Εξαγωγής-Μετασχηματισμού-Φόρτωσης (ΕΜΦ) δεδομένων για αποθήκες δεδομένων. Προτείνεται ένα εννοιολογικό μοντέλο χρήσιμο στην αρχική φάση ενός έργου σχεδιασμού και ανάπτυξης μίας αποθήκης δεδομένων. Οι βασικές απαιτήσεις σε αυτό το στάδιο του έργου είναι η ταυτοποίηση των αλληλοσυσχετίσεων των γνωρισμάτων και η καταγραφή των απαραίτητων διεργασιών ΕΜΦ. Για την ικανοποίηση των απαιτήσεων αυτής της φάσης, το μοντέλο παρουσιάζει τρία βασικά χαρακτηριστικά: απλότητα, γενικότητα και επεκτασιμότητα. Παράλληλα, παρουσιάζεται μία μεθοδολογία χρήσης του, ενώ περιγράφεται ένα γενικότερο πλαίσιο μοντελοποίησης για να δειχθεί η αποτελεσματικότητα και η λειτουργικότητά του σε πραγματικές εφαρμογές.

1. ΕΙΣΑΓΩΓΗ

Για ένα μεγάλο διάστημα στο παρελθόν, η έρευνα αντιμετώπιζε τις αποθήκες δεδομένων ως συλλογές υλοποιημένων όψεων. Αν και αυτή η θεώρηση είναι κομψή και πιθανώς επαρκής για τον σκοπό της εξέτασης εναλλακτικών στρατηγικών για την διατήρηση όψεων, κρίνεται ανεπαρκής για την περιγραφή της δομής και των περιεχομένων μίας αποθήκης δεδομένων. Στην αγορά κυκλοφορούν εξειδικευμένα εργαλεία, υπό τον γενικό χαρακτηρισμό εργαλεία *ΕΜΦ, Εξαγωγής-Μετασχηματισμού-Φόρτωσης (ETL, Extraction-Transformation-Loading)*[1], με σκοπό να διευκολύνουν και να χειριστούν τις λειτουργικές διαδικασίες των αποθήκες δεδομένων. Τα ΕΜΦ εργαλεία αναλαμβάνουν το έργο της διατήρησης της ομοιογένειας, του καθαρισμού και της φόρτωσης μίας αποθήκης δεδομένων. Το κόστος αυτών των εργασιών υπολογίζεται ότι ανέρχεται στο ένα τρίτο της συνολικής εργασίας και των εξόδων στον προϋπολογισμό μίας αποθήκης δεδομένων, ενώ ο χρόνος ανάπτυξής τους μπορεί να φτάσει ως και το 80% του συνολικού χρόνου ανάπτυξης μίας αποθήκης δεδομένων. Παρόλα αυτά, κυρίως λόγω της πολυπλοκότητας, του κόστους και της δύσκολης προσαρμογής των έτοιμων πακέτων στις ανάγκες κάθε εταιρίας, πολλοί οργανισμοί αναπτύσσουν μόνοι τους εργαλεία για την εκτέλεση ΕΜΦ εργασιών προσαρμοσμένα στις ανάγκες των εκάστοτε περιστάσεων.

Για να δώσουμε μία γενική ιδέα της λειτουργικότητας αυτών των εργαλείων αναφέρουμε τις πιο χαρακτηριστικές λειτουργίες τους που περιλαμβάνουν: (α) τον εντοπισμό της σχετικής πληροφορίας στην πλευρά της πηγής, (β) την εξαγωγή της πληροφορίας αυτής, (γ) την προσαρμογή και την ενοποίηση πληροφορίας που προέρχεται από πολλαπλές πηγές σε μία κοινή μορφή, (δ) τον καθαρισμό του παραγόμενου συνόλου δεδομένων, με βάση τους κανόνες της βάσης δεδομένων, αλλά και άλλους λογικούς κανόνες, και τέλος, (ε) την προώθηση των δεδομένων στην αποθήκη δεδομένων και/ή στις αποθηκευμένες όψεις.

Στη συνέχεια, θα υιοθετήσουμε το όνομα ΕΜΦ τόσο για τις διεργασίες ΕΜΦ, όσο και για αυτές του καθαρισμού των δεδομένων.



Σχήμα 1. Το περιβάλλον των διεργασιών ΕΜΦ

Στο Σχήμα 1 περιγράφεται αφηρημένα το γενικό πλαίσιο των διεργασιών ΕΜΦ. Στο κατώτερο επίπεδο, απεικονίζονται τα σημεία αποθήκευσης δεδομένων που εμπλέκονται στην όλη διαδικασία. Στην αριστερή πλευρά, παρατηρούμε τους αρχικούς προμηθευτές – πηγές δεδομένων. Συνήθως, ως πηγές δεδομένων θεωρούμε σχεσιακές βάσεις δεδομένων και αρχεία. Τα δεδομένα από τις πηγές αυτές εξάγονται από ρουτίνες *εξαγωγής* (όπως φαίνεται στο πάνω αριστερό μέρος του Σχήματος 1), οι οποίες παρέχουν είτε ολόκληρα στιγμιότυπα των πηγών είτε τις εκάστοτε διαφορές αυτών. Στη συνέχεια, τα εξαγόμενα δεδομένα

[1] Έγινε μια προσπάθεια απόδοσης των όρων στην ελληνική γλώσσα. Για όσες απορίες προκύψουν στον αναγνώστη, τον παραπέμπουμε στο γλωσσάριο στο τέλος του άρθρου.

μεταφέρονται στη *Μεταβατική Περιοχή Αποθήκευσης Δεδομένων*, *ΜΠΑΔ*, όπου μετασχηματίζονται και καθαρίζονται πριν φορτωθούν στην αποθήκη δεδομένων. Η αποθήκη δεδομένων, *ΑΔ*, απεικονίζεται στο δεξί μέρος του κατώτερου επιπέδου και αποτελεί τον τελικό προορισμό αποθήκευσης των δεδομένων, αφενός σε πίνακες πληροφοριών για την αποθήκευση της πληροφορίας και αφετέρου σε πίνακες διαστάσεων με την περιγραφή και τις πολυδιάστατες ιεραρχίες των αποθηκευμένων αυτών δεδομένων. Η *φόρτωση* της κεντρικής αποθήκης δεδομένων γίνεται από αντίστοιχες ρουτίνες και εργαλεία που απεικονίζονται στο πάνω δεξί μέρος του σχήματος.

Στο [14], κάποιος μπορεί να βρει μία λεπτομερή περιγραφή των διαφορετικών πτυχών των διεργασιών ΕΜΦ: τον καθορισμό της εργασίας της διεργασίας, τον προγραμματισμό με βάση το χρόνο ή κάποιο γεγονός, την παρακολούθηση (όπως, συνεχή παρακολούθηση της προόδου/κατάστασης της διεργασίας), τον χειρισμό λαθών (για εισερχόμενα δεδομένα που παραβιάζουν τους περιορισμούς ακεραιότητας ή τους επιχειρησιακούς κανόνες), την επαναφορά μετά από κατάρρευση και ικανότητες έναρξης/τερματισμού (όπως, για παράδειγμα, καταχώρηση μίας δοσοληψίας κάθε μερικές εγγραφές) καθώς και διάφορες άλλες δυνατότητες.

Στην εργασία αυτή, εστιάζουμε την προσοχή μας στο εννοιολογικό τμήμα του ορισμού των διεργασιών ΕΜΦ. Συγκεκριμένα, ασχολούμαστε με τα αρχικά στάδια του σχεδιασμού της αποθήκης δεδομένων. Κατά τη διάρκεια αυτής της φάσης, ο σχεδιαστής της αποθήκης δεδομένων ενδιαφέρεται για δύο διαδικασίες, οι οποίες εκτελούνται παράλληλα. Η πρώτη διαδικασία περιλαμβάνει τη *συλλογή των απαιτήσεων του χρήστη*. Η δεύτερη διαδικασία, η οποία είναι ίσης σπουδαιότητας για την επιτυχία της ανάπτυξης της αποθήκης δεδομένων, περιλαμβάνει την *ανάλυση της δομής των πηγών δεδομένων που υπάρχουν ήδη και την τελική αντιστοίχησή τους στο μοντέλο της αποθήκης δεδομένων*. Η σχετική βιβλιογραφία [14, 27] αναφέρει ότι ο σχεδιασμός διεργασιών ΕΜΦ στοχεύει την παραγωγή ενός κρίσιμου παραδοτέου: την απεικόνιση των γνωρισμάτων των πηγών δεδομένων στα γνωρίσματα των πινάκων της αποθήκης δεδομένων. Η παραγωγή αυτού του παραδοτέου περιλαμβάνει πολλές συνομιλίες οι οποίες καταλήγουν συχνά σε αναθεώρηση και επαναδιατύπωση των αρχικών υποθέσεων και απεικονίσεων. Κατά συνέπεια, είναι εύλογο ότι πρέπει να υιοθετηθεί ένα απλό εννοιολογικό μοντέλο, το οποίο (α) διευκολύνει τις προσπάθειες επαναπροσδιορισμού και αναθεώρησης και (β) χρησιμεύει ως τρόπος επικοινωνίας μεταξύ των συμβαλλόμενων μερών.

Πιστεύουμε ότι η αυστηρή μοντελοποίηση των αρχικών εννοιών του σχεδιασμού μίας αποθήκης δεδομένων δεν έχει εξεταστεί επαρκώς από την ερευνητική κοινότητα. Για το σκοπό αυτό, προτείνουμε ένα εννοιολογικό μοντέλο με στόχο:

- την εύρεση των αλληλεξαρτήσεων μεταξύ γνωρισμάτων και εννοιών, και
- τον προσδιορισμό των απαιτουμένων μετασχηματισμών που χρειάζεται να πραγματοποιηθούν κατά τη διάρκεια της φόρτωσης της αποθήκης δεδομένων.

Υιοθετούμε τον όρο *μετασχηματισμός* ως ένα γενικό όρο για την αναδόμηση του σχήματος και των τιμών ή για την επιλογή και τον μετασχηματισμό των δεδομένων. Η αλληλεξάρτηση των γνωρισμάτων εκφράζεται μέσω *σχέσεων παροχής*, οι οποίες αντιστοιχίζουν γνωρίσματα των πηγών δεδομένων με τα αντίστοιχα γνωρίσματα στην αποθήκη δεδομένων. Εκτός από αυτό το θεμελιώδες είδος σχέσης, το προτεινόμενο μοντέλο είναι ικανό να συλλάβει πάσης φύσης περιορισμούς αλλά και σύνθεση μετασχηματισμών. Λόγω της φύσης της διαδικασίας σχεδιασμού, παρουσιάζονται τα χαρακτηριστικά γνωρίσματα του εννοιολογικού μοντέλου ως ένα σύνολο βημάτων σχεδίασης, τα οποία οδηγούν στο βασικό στόχο: τις αλληλεξαρτήσεις των γνωρισμάτων. Τα βήματα αυτά συνιστούν τη μεθοδολογία για το σχεδιασμό του εννοιολογικού μοντέλου για διεργασίες ΕΜΦ.

Το γενικό μετα-μοντέλο περιλαμβάνει ένα μικρό σύνολο *γενικών κατασκευαστών*, που είναι ικανοί για την αντιμετώπιση όλων των περιπτώσεων. Στην αρχιτεκτονική μας, αυτές οι οντότητες συνιστούν το *Επίπεδο Μετα-Μοντέλου*. Επιπλέον, εισάγουμε ένα μηχανισμό που επιτρέπει τη δημιουργία μίας «παλέτας» *πρότυπων συχνά χρησιμοποιούμενων μετασχηματισμών* (όπως ανάθεση υποκατάστατου κλειδιού, παραβιάσεις πρωτεύοντος κλειδιού, κ.λπ.). Αυτό το σύνολο των πρότυπων μετασχηματισμών, προσανατολισμένων σε διεργασίες ΕΜΦ, αποτελεί υποσύνολο του Επιπέδου Μετα-Μοντέλου και καλείται *Επίπεδο Προτύπων*. Αν ο σχεδιαστής χρειαστεί επιπλέον μετασχηματισμούς από αυτούς που βρίσκονται ήδη στην «παλέτα», μέσω του μηχανισμού που αναφέραμε μπορεί να δημιουργήσει καινούριους πρότυπους μετασχηματισμούς.

Η έρευνά μας επικεντρώνεται γύρω από το ερευνητικό έργο Αρκτος ΙΙ, το οποίο έχει ως κύριο στόχο τη μοντελοποίηση και τη βελτιστοποίηση σεναρίων ΕΜΦ για αποθήκες δεδομένων. Το συγκεκριμένο άρθρο βασίζεται κυρίως στα [29] και [26] και επικεντρώνεται στον εννοιολογικό σχεδιασμό σεναρίων ΕΜΦ για αποθήκες δεδομένων.

Η συνεισφορά του άρθρου μπορεί να καταγραφεί ως εξής:

- Πρόταση ενός καινούριου εννοιολογικού μοντέλου που είναι προσαρμοσμένο στην ταυτοποίηση των αλληλοσυσχετίσεων των γνωρισμάτων και στην αναγνώριση των απαραίτητων διεργασιών ΕΜΦ, που απαιτούνται στα αρχικά στάδια του έργου σχεδίασης μίας αποθήκης δεδομένων.
- Περιγραφή μίας μεθοδολογίας για το σχεδιασμό του εννοιολογικού τμήματος διεργασιών ΕΜΦ.
- Κατασκευή του προτεινόμενου μοντέλου με κύρια χαρακτηριστικά τη γενικότητα και την επεκτασιμότητα, ώστε ο σχεδιαστής να μπορεί να το εμπλουτίσει με καινούριες πρότυπες διεργασίες ΕΜΦ.
- Εισαγωγή μίας «παλέτας», ενός συνόλου συχνά χρησιμοποιουμένων διεργασιών ΕΜΦ, όπως η ανάθεση υποκατάστατων κλειδιών, ο έλεγχος περιορισμών αναφορικής ακεραιότητας, κ.λπ..

Το άρθρο αυτό οργανώνεται ως εξής. Στην ενότητα 2 παρουσιάζεται η σχετική με θέματα ΕΜΦ εργασία. Στην ενότητα 3 αναπτύσσεται το εννοιολογικό μοντέλο για διεργασίες ΕΜΦ και στην ενότητα 4 παρουσιάζεται μία μεθοδολογία για τη χρήση του εννοιολογικού μοντέλου. Στην ενότητα 5 παρουσιάζονται δύο βασικά χαρακτηριστικά του μοντέλου: η γενίκευση και η επεκτασιμότητα. Τέλος, στην ενότητα 6 καταγράφονται τα συμπεράσματα της συγκεκριμένης μελέτης, καθώς και κάποια ερευνητικά θέματα που προκύπτουν και επιζητούν περαιτέρω μελέτη στο μέλλον.

2. ΣΧΕΤΙΚΗ ΕΡΓΑΣΙΑ

Σ' αυτήν την ενότητα παρουσιάζονται ερευνητικές προσπάθειες στο πεδίο του εννοιολογικού σχεδιασμού για αποθήκες δεδομένων και διεργασίες ΕΜΦ γενικότερα. Για περισσότερες πληροφορίες, παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στο [32] για μία εκτεταμένη ανάλυση όσων παρουσιάζονται σε αυτήν την ενότητα.

Εννοιολογικά Μοντέλα για Αποθήκες Δεδομένων. Το front end της αποθήκης δεδομένων, έχει μονοπωλήσει την έρευνα στο εννοιολογικό κομμάτι για τη μοντελοποίηση των αποθηκών δεδομένων. Στην πραγματικότητα, το μεγαλύτερο μέρος της εργασίας για την μοντελοποίηση του εννοιολογικού μέρους στο πεδίο των αποθηκών δεδομένων έχει αφιερωθεί στη σύλληψη των εννοιολογικών χαρακτηριστικών του σχήματος αστέρα της αποθήκης, των επιμέρους συλλογών δεδομένων και των συναθροίσεων (για μία πιο αναλυτική συζήτηση, δες [24]). Οι ερευνητικές προσπάθειες μπορούν να ομαδοποιηθούν σε τέσσερις κατηγορίες: (α) *μοντελοποίηση των διαστάσεων* [13, 14], (β) *επεκτάσεις του τυπικού μοντέλου Ο/Σ* [21, 4, 5, 9, 16, 22], (γ) *μοντελοποίηση UML* [18, 23] (δ) *ιδιαίτερα μοντέλα* [7, 8, 24] χωρίς, όμως, ξεκάθαρο νικητή. Οι υποστηρικτές της μοντελοποίησης των διαστάσεων υποστηρίζουν ότι το μοντέλο είναι ελαχιστοποιημένο και κατανοητό (ειδικά από τους τελικούς χρήστες) καθώς και ότι αντιστοιχίζεται άμεσα με λογικές δομές. Οι υποστηρικτές του τυπικού μοντέλου Ο/Σ καθώς και του μοντέλου UML, βασίζουν τις απόψεις τους στη δημοτικότητα των μοντέλων τους, καθώς και στα αυστηρά σημασιολογικά θεμέλια και στην καλή διαμόρφωση των εννοιολογικών σχημάτων των αποθηκών δεδομένων. Τα μοντέλα της τελευταίας κατηγορίας, που δεν εμπίπτουν σε μία από τις προηγούμενες γενικές κατηγορίες, στηρίζονται κυρίως στην καινοτομία και στην ικανότητα να προσαρμόζονται στις ιδιαιτερότητες του περιβάλλοντος ΣΑΕΔ.

Εννοιολογικά μοντέλα για διεργασίες ΕΜΦ. Υπάρχουν μερικές προσπάθειες σχετικά με το συγκεκριμένο πρόβλημα, αν και δε γνωρίζουμε κάποια άλλη προσπάθεια που να εξετάζει διεξοδικά τις λεπτομέρειες των διεργασιών ΕΜΦ σε εννοιολογικό επίπεδο. Μπορούμε να αναφερθούμε στο [1] ως μία πρώτη προσπάθεια ξεκάθαρου διαχωρισμού της διαδικασίας ανανέωσης της αποθήκης δεδομένων από την παραδοσιακή επεξεργασία ως συντήρηση όψεων ή ως διαδικασία μαζικής φόρτωσης. Όμως, το προτεινόμενο μοντέλο είναι ανεπίσημο και εστιάζεται περισσότερο στην παρουσίαση αποδείξεων για την πολυπλοκότητα της προσπάθειας αυτής παρά στην τυπική μοντελοποίηση των διεργασιών αυτών καθ' αυτών. Στα [4, 5] εισάγεται η έννοια των *ισχυρισμών* μέσα στο μοντέλο, προκειμένου να συλλάβουν τις αντιστοιχίσεις μεταξύ των πηγών και της αποθήκης δεδομένων. Εντούτοις, οποιοσδήποτε μετασχηματισμός πρέπει να οριστεί εκ νέου για τη μετάβαση στο λογικό μοντέλο. Εκτός αυτών των ερευνητικών προσεγγίσεων, από το χώρο της αγοράς, το μοντέλο που περιγράφεται στο [14] μπορεί να θεωρηθεί ως μία άτυπη τεκμηρίωση της γενικής διαδικασίας ΕΜΦ.

Σχετικές εργασίες για ΕΜΦ σε λογικό και φυσικό επίπεδο. Τέλος, εκτός από την πληθώρα εμπορικών εργαλείων ΕΜΦ [25] υπάρχουν και κάποιες ερευνητικές προσπάθειες, όπως [2, 6, 15, 17, 19, 20, 32, 30, 28]. Η διαχείριση της ποιότητας στις αποθήκες δεδομένων, συζητείται εκτενώς στα [10, 11, 12].

Τονίζουμε ότι αυτό το άρθρο δεν περιγράφει ακόμη ένα μοντέλο διαδικασιών ή ροής έργου. Κατά συνέπεια, δεν καλύπτει τη ροή έργου των διεργασιών ΕΜΦ για τη φόρτωση δεδομένων στην αποθήκη. Υπάρχουν δύο βασικοί λόγοι για την προσέγγιση αυτή. Κατ' αρχάς, το εννοιολογικό μοντέλο για τις διεργασίες ΕΜΦ εστιάζεται στην τεκμηρίωση και τυποποίηση των ιδιαιτεροτήτων των πηγών των δεδομένων σε σχέση με την αποθήκη δεδομένων και όχι στην παροχή μίας τεχνικής λύσης για την εφαρμογή της διαδικασίας. Αφετέρου, το εννοιολογικό μοντέλο ΕΜΦ κατασκευάζεται στα αρχικά στάδια ενός έργου αποθήκης δεδομένων, κατά τη διάρκεια των οποίων οι χρονικοί περιορισμοί του έργου απαιτούν περισσότερο μία γρήγορη τεκμηρίωση των εμπλεκομένων πηγών δεδομένων και των αλληλοσυσχετίσεών τους, παρά μία σε βάθος περιγραφή της σύνθετης ροής έργου. Από αυτή την άποψη, η προσέγγιση μας είναι συμπληρωματική ως προς τα προαναφερθέντα λογικά πρότυπα, δεδομένου ότι αφορά σε ένα αρχικό στάδιο της διαδικασίας σχεδιασμού. Παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στα [28, 31] για ένα τυπικό μοντέλο που περιγράφει τη ροή έργου σε διεργασίες ΕΜΦ.

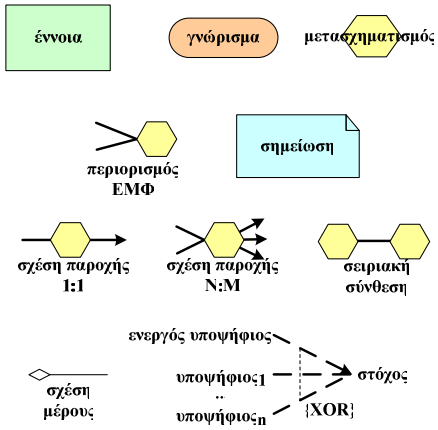
3. ΕΝΝΟΙΟΛΟΓΙΚΟ ΜΟΝΤΕΛΟ

Στην ενότητα αυτή παρουσιάζεται ένα εννοιολογικό μοντέλο για διεργασίες ΕΜΦ. Ο σκοπός είναι να εντοπιστούν σε υψηλό επίπεδο οι οντότητες που χρησιμοποιούνται για τη σύλληψη της σημασιολογίας των διεργασιών ΕΜΦ. Αρχικά, παρουσιάζονται οι γραφικοί συμβολισμοί και το μετα-μοντέλο. Έπειτα, διευκρινίζονται τα δομικά συστατικά του μοντέλου μέσω ενός χαρακτηριστικού παραδείγματος.

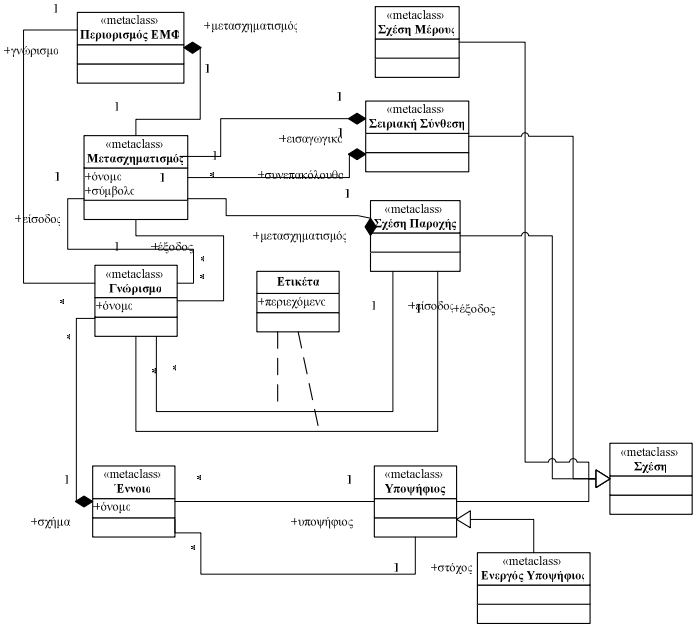
Στο Σχήμα 2 απεικονίζονται γραφικά οι διάφορες οντότητες για το προτεινόμενο μοντέλο. Δεν υιοθετούμε συμβολισμούς UML για τις έννοιες και τα γνωρίσματα, καθώς θεωρούμε ότι τα γνωρίσματα στις διεργασίες ΕΜΦ θα πρέπει να αντιμετωπίζονται ως «πολίτες πρώτης τάξης». Κατά συνέπεια, τα γνωρίσματα δεν περιλαμβάνονται στον ορισμό της οντότητας όπου ανήκουν, όπως για παράδειγμα μία κλάση UML ή ένας σχεσιακός πίνακας. Το προτεινόμενο μοντέλο είναι ορθόγωνο με τα εννοιολογικά μοντέλα που έχουν προταθεί για αποθήκες δεδομένων σχήματος αστέρα. Στην πραγματικότητα, η σχετική εργασία που έχει γίνει για το front end των αποθηκών δεδομένων, μπορεί να συνδυαστεί με το προτεινόμενο μοντέλο, που είναι σαφώς προσανατολισμένο στο back end της αποθήκης δεδομένων.

Στο Σχήμα 3 απεικονίζονται οι βασικές οντότητες του προτεινόμενου μετα-μοντέλου ως διάγραμμα UML. Όλα τα συστατικά του εννοιολογικού μοντέλου που εισάγονται στη συνέχεια, αναφέρονται στις οντότητες του Σχήματος 3.

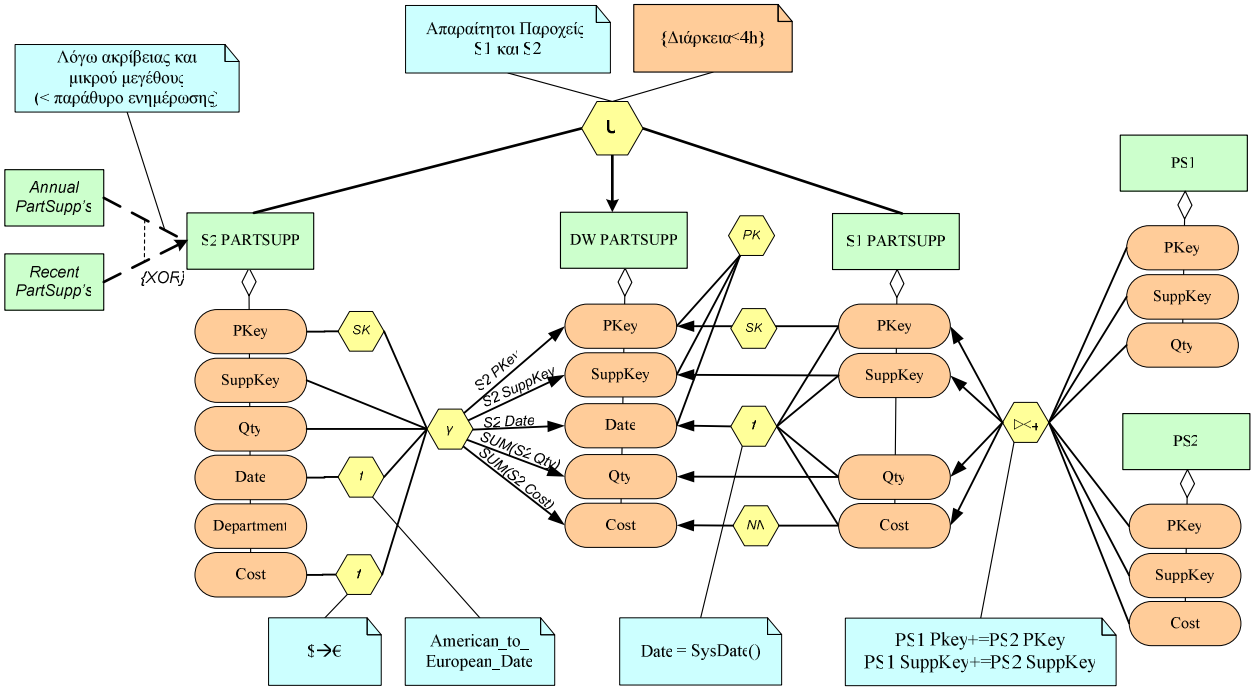
Παράδειγμα. Για τη διευκόλυνση της παρουσίασης του μοντέλου, παρουσιάζουμε το ακόλουθο παράδειγμα. Θεωρούμε δύο βάσεις δεδομένων, `S1` και `S2`, που στο εξής θα αναφέρονται ως πηγές `S1` και `S2`, και μία αποθήκη δεδομένων, `DW`, που στο εξής θα αναφέρεται ως έννοια-στόχος `DW`. Το συγκεκριμένο παράδειγμα περιλαμβάνει τη διάδοση δεδομένων από την έννοια `PARTSUPP(PKey,SuppKey,Qty,Cost)` της πηγής S_1, αλλά και από την έννοια `PARTSUPP(PKey,Department, SuppKey,Date,Qty,Cost)` της πηγής S_2 στο στόχο `DW`. Στην αποθήκη δεδομένων, η έννοια `DW.PARTSUPP(PKey, SuppKey,Date,Qty,Cost)` αποθηκεύει καθημερινά (`Date`)



Σχήμα 2. Γραφικοί συμβολισμοί για το εννοιολογικό μοντέλο των διεργασιών ΕΜΦ



Σχήμα 3. Το προτεινόμενο μετα-μοντέλο ως διάγραμμα UML



Σχήμα 4. Διάγραμμα του εννοιολογικού μοντέλου για το παράδειγμα

πληροφορία για τη διαθέσιμη ποσότητα (Qty) και κόστος (Cost) των εξαρτημάτων (PKey) ανά προμηθευτή (SuppKey). Υποθέτουμε ότι ο πρώτος προμηθευτής είναι Ευρωπαίος, ενώ ο δεύτερος Αμερικανός, κατά συνέπεια, τα δεδομένα που προέρχονται από τη δεύτερη πηγή θα πρέπει να προσαρμοστούν στις ευρωπαϊκές μονάδες και τυποποιήσεις. Επίσης, για τον πρώτο προμηθευτή θεωρούμε ότι απαιτείται ο συνδυασμός πληροφορίας από δύο διαφορετικές έννοιες στην αρχική βάση δεδομένων, στοιχείο που επιτυγχάνεται με μία εξωτερική συνένωση των εννοιών PS_1 και PS_2. Το Σχήμα 4 απεικονίζει το πλήρως ανεπτυγμένο διάγραμμα του εννοιολογικού μοντέλου για αυτό το παράδειγμα. Στο υπόλοιπο αυτής της ενότητας θα εξηγήσουμε κάθε τμήμα του λεπτομερώς.

3.1 Συστατικά του μοντέλου

Σε αυτή την ενότητα παρουσιάζονται τα επιμέρους δομικά συστατικά του μοντέλου. Περιγράφονται τα γνωρίσματα, οι έννοιες, οι μετασχηματισμοί, οι περιορισμοί ΕΜΦ και τα σχόλια. Επίσης, περιγράφονται τα διαφορετικά είδη σχέσεων που προβλέπει το προτεινόμενο εννοιολογικό μοντέλο: σχέση μέρους, σχέση υποψηφίου και σχέση παροχής. Τέλος, παρουσιάζεται η σειριακή σύνθεση μετασχηματισμών με την οποία επεκτείνεται τη δυνατότητα του μοντέλου για την κάλυψη περισσότερο σύνθετων περιπτώσεων διεργασιών ΕΜΦ.

Γνωρίσματα. Το *γνώρισμα* αποτελεί τη στοιχειώδη μονάδα πληροφορίας στο μοντέλο. Ο ρόλος των γνωρισμάτων είναι ο ίδιος όπως στα τυποποιημένα διαγράμματα Ο/Σ. Υιοθετούμε τη γραφική αναπαράσταση των γνωρισμάτων του τυποποιημένου πρότυπου Ο/Σ και απεικονίζουμε τα γνωρίσματα με οβάλ σχήμα.

Έννοιες. Η *έννοια* αντιπροσωπεύει μία οντότητα στις πηγές ή στην αποθήκη δεδομένων. Παραδείγματα εννοιών είναι τα αρχεία δεδομένων στις πηγές, οι πίνακες γεγονότων και διαστάσεων στην αποθήκης δεδομένων κ.λπ.. Η έννοια ορίζεται τυπικά από (α) ένα όνομα και (β) ένα σύνολο γνωρισμάτων, ενώ γραφικά απεικονίζεται με ένα ορθογώνιο σχήμα. Με βάση το μοντέλο Ο/Σ, η έννοια αποτελεί μία γενίκευση των οντοτήτων και των συσχετίσεων. Ανεξάρτητα όμως από το χρησιμοποιούμενο μοντέλο, είτε πρόκειται για επέκταση του μοντέλου Ο/Σ είτε για το μοντέλο διαστάσεων, όλες οι οντότητες που συντίθενται από ένα σύνολο γνωρισμάτων, είναι γενικά στιγμιότυπα της κλάσης «έννοια».

Όπως αναφέρεται στο [30], μπορούμε να χρησιμοποιήσουμε διάφορες *φυσικές* δομές αποθήκευσης ως πεπερασμένες λίστες γνωρισμάτων, συμπεριλαμβανομένων των σχεσιακών βάσεων δεδομένων, αρχείων COBOL ή ASCII, πολυδιάστατων κύβων, καθώς και διαστάσεων. Οι έννοιες είναι απόλυτα ικανές να μοντελοποιήσουν τέτοιου είδους δομές, πιθανότατα μέσω ενός γενικευμένου (IsA) μηχανισμού. Ας θεωρήσουμε για παράδειγμα τις δομές ΣΑΕΔ. Οι συσχετίσεις των επιπέδων και των τιμών, οι οποίες αποτελούν τον πυρήνα όλων των προσεγγίσεων που αναφέρονται στην Ενότητα 2, δε σχετίζονται με την περίπτωση των διεργασιών ΕΜΦ. Η υιοθέτηση των εννοιών είναι αρκετή για το πρόβλημα μοντελοποίησης διεργασιών ΕΜΦ. Εξάλλου, μπορούμε να διασπάσουμε τη γενική δομή Έννοια σε υποκλάσεις που θα φέρουν τα χαρακτηριστικά οποιασδήποτε από τις προαναφερθείσες προσεγγίσεις (π.χ. υποκλάσεις Πίνακας Γεγονότων και Πίνακας Διάσταση), επιτυγχάνοντας έτσι ομοιογένεια στη μεταχείριση μοντέλων ΣΑΕΔ και ΕΜΦ.

Επιστρέφοντας στο παράδειγμα του Σχήματος 4 παρατηρούμε τις έννοιες PS1, PS2, S1.PARTSUPP, S2.PARTSUPP και DW.PARTSUPP μαζί με τα γνωρίσματά τους.

Μετασχηματισμοί. Ο *μετασχηματισμός* αντιπροσωπεύει μέρος ή πλήρη ενότητα κώδικα, εκτελώντας μία συγκεκριμένη διαδικασία. Διαχωρίζουμε δύο γενικές κατηγορίες μετασχηματισμών: (α) μετασχηματισμούς που συντηρούν το σχήμα των εισερχομένων δεδομένων, όπως το φιλτράρισμα ή ο καθαρισμός δεδομένων (π.χ. έλεγχος για παραβιάσεις πρωτεύοντος ή δευτερεύοντος κλειδιού), και (β) μετασχηματισμούς που μεταβάλλουν το σχήμα των εισερχόμενων δεδομένων (π.χ. συνάθροιση). Τυπικά, ένας μετασχηματισμός ορίζεται από (α) ένα πεπερασμένο σύνολο γνωρισμάτων εισόδου, (β) ένα πεπερασμένο σύνολο γνωρισμάτων εξόδου, και (γ) ένα γραφικό σύμβολο που χαρακτηρίζει τη φύση του μετασχηματισμού. Ο μετασχηματισμός απεικονίζεται γραφικά με ένα εξάγωνο σχήμα.

Στο παράδειγμα του Σχήματος 4, μπορούμε να δούμε αρκετούς μετασχηματισμούς. Παρατηρήστε, λόγου χάρη, αυτούς που σχετίζονται με τη διάδοση δεδομένων από την έννοια S1.PARTSUPP στην DW.PARTSUPP. Υπάρχει ένας μετασχηματισμός ανάθεσης υποκατάστατου κλειδιού (SK), μία συνάρτηση υπολογισμού της ημερομηνίας συστήματος (f) και ένας έλεγχος για κενές τιμές (NN) στο γνώρισμα Cost.

Οι μετασχηματισμοί δε χρησιμοποιούνται ως αυτόνομες οντότητες στο μοντέλο, αντίθετα αποτελούν μέρος άλλων δομικών συστατικών αυτού. Κατά συνέπεια, περισσότερα για τη λειτουργία τους θα αναφερθούν κατά την περιγραφή των συστατικών που τους χρησιμοποιούν.

Περιορισμοί ΕΜΦ. Σε αρκετές περιπτώσεις, ο σχεδιαστής θέλει να τονίσει το γεγονός ότι τα δεδομένα μίας συγκεκριμένης έννοιας οφείλουν να πληρούν κάποιες απαιτήσεις. Για παράδειγμα, ο σχεδιαστής μπορεί να θέλει να επιβάλει ένα περιορισμό πρωτεύοντος κλειδιού ή ένα περιορισμό μη-μηδενικής τιμής σε ένα σύνολο γνωρισμάτων. Για να καλυφθούν τέτοιες ανάγκες εισάγουμε τους *περιορισμούς ΕΜΦ*, που τυπικά ορίζονται ως: (α) ένα πεπερασμένο σύνολο γνωρισμάτων στα οποία εφαρμόζεται ο περιορισμός, και (β) ένας μετασχηματισμός ο οποίος επιβάλλει τον περιορισμό. Σ' αυτό το σημείο πρέπει να τονιστεί ότι παρ' όλη την ομοιότητα στο όνομα, οι περιορισμοί ΕΜΦ είναι διαφορετικά στοιχεία μοντελοποίησης από τους γνωστούς περιορισμούς UML. Ένας περιορισμός ΕΜΦ απεικονίζεται γραφικά ως ένα σύνολο από συνεχείς γραμμές οι οποίες ξεκινούν από τα εμπλεκόμενα γνωρίσματα και καταλήγουν στο μετασχηματισμό υλοποίησης του περιορισμού. Στο παράδειγμα του Σχήματος 4, εφαρμόζεται ένας περιορισμός ΕΜΦ πρωτεύοντος κλειδιού (PK) στην έννοια DW.PARTSUPP που αφορά στα γνωρίσματα PKey, SuppKey, Date.

Σχόλια. Το *σχόλιο*, ακριβώς όπως και στη μοντελοποίηση UML, αποτελεί άτυπη ετικέτα για να αποτυπωθούν πρόσθετα σχόλια που επιθυμεί να κάνει ο σχεδιαστής στη φάση σχεδιασμού, ή για να αποτυπώσει περιορισμούς σε κάποια στοιχεία ή σύνολα στοιχείων [3]. Όπως και στη UML, τα σχόλια απεικονίζονται ως ορθογώνια με την πάνω δεξιά γωνία διπλωμένη. Στο προτεινόμενο μοντέλο, τα σχόλια χρησιμοποιούνται για:

- Σύντομα σχόλια που εξηγούν τις σχεδιαστικές αποφάσεις.
- Επεξηγήσεις για τη σημασιολογία των εφαρμοσμένων μετασχηματισμών. Για παράδειγμα, στην περίπτωση

σχεσιακών επιλογών ή συνενώσεων, με σχόλιο αποδίδονται οι συνθήκες της αντίστοιχης επιλογής ή συνένωσης, ενώ στην περίπτωση των συναρτήσεων, με σχόλιο αποδίδεται ο καθορισμός του είδους της συνάρτησης που εφαρμόζεται.

- Επισήμανση των περιορισμών χρόνου εκτέλεσης που εμπλέκονται σε διάφορες πτυχές των διεργασιών ΕΜΦ, όπως ο προγραμματισμός με βάση το χρόνο ή κάποιο γεγονός, η επίβλεψη, η καταγραφή συμβάντων, η αντιμετώπιση λαθών, η ανάνηψη μετά από κατάρρευση, κ.λπ..

Επιστρέφοντας στο παράδειγμα του Σχήματος 4, παρατηρούμε στο πάνω τμήμα έναν περιορισμό χρόνου εκτέλεσης, σύμφωνα με τον οποίο ο συνολικός χρόνος εκτέλεσης για τη φόρτωση του `DW.PARTSUPP` (που περιλαμβάνει τη φόρτωση των `S1.PARTSUPP` και `S2.PARTSUPP`) δεν μπορεί να υπερβεί τις 4 ώρες (ή αλλιώς το χρονικό παράθυρο φόρτωσης της αποθήκης δεδομένων είναι διάρκειας 4 ωρών).

Σχέσεις Μέρους. Η *σχέση μέρους* υπογραμμίζει το γεγονός ότι μία έννοια αποτελείται από ένα σύνολο γνωρισμάτων. Γενικά, το μοντέλο Ο/Σ δε μεταχειρίζεται αυτό το είδος σχέσης ως «πολίτη πρώτης τάξης», ενώ η μοντελοποίηση με UML, κρύβει τα γνωρίσματα μέσα στις κλάσεις και χρησιμοποιεί τις σχέσεις μέρους με μία πολύ ευρύτερη έννοια. Δεν επιθυμούμε να ορίσουμε ξανά τις σχέσεις μέρους της UML, αλλά να τονίσουμε τη σχέση μίας έννοιας με τα γνωρίσματά της, με δεδομένο ότι χρειαζόμαστε τα γνωρίσματα ως «πολίτες πρώτης τάξης» για να απεικονίσουμε τις αλληλοσυσχετίσεις μεταξύ τους. Η σχέση μέρους απεικονίζεται ως μία συνεχής γραμμή μεταξύ μίας έννοιας και ενός γνωρίσματος που φέρει ένα ρόμβο στην πλευρά της έννοιας που περιέχει τα γνωρίσματα. Για λόγους σχεδιαστικής απλότητας, συχνά εμφανίζεται μία σχέση μέρους μεταξύ μίας έννοιας και όλων των γνωρισμάτων που ανήκουν σε αυτή.

Σχέσεις Υποψηφίου. Στη σχεδίαση των αποθηκών δεδομένων, είναι πολύ συνηθισμένο, ειδικά στα πρώτα στάδια σχεδιασμού του έργου, να υπάρχουν περισσότερες από μία πιθανές έννοιες (πηγές) για τη φόρτωση μίας έννοιας (στόχου) στην αποθήκη δεδομένων. Κάθε μία από αυτές τις πιθανές έννοιες ονομάζεται *υποψήφια*. Επομένως, οι *σχέσεις υποψηφίου* χρησιμοποιούνται για να υποδηλώσουν το γεγονός ότι ενδέχεται να υπάρχει ένα σύνολο πηγών ικανών να διαδώσουν δεδομένα σε μία συγκεκριμένη έννοια. Τυπικά, μία σχέση υποψηφίου περιλαμβάνει (α) ακριβώς μία υποψήφια έννοια και (β) ακριβώς μία έννοια στόχο. Οι σχέσεις υποψηφίου απεικονίζονται με έντονες διακεκομμένες γραμμές μεταξύ των υποψηφίων και της έννοιας-στόχου. Όταν ακριβώς μία από αυτές πρέπει να επιλεγεί, σημειώνουμε το σύνολο των σχέσεων για τη συγκεκριμένη έννοια με έναν περιορισμό UML `{XOR}`.

Σχέσεις Ενεργού Υποψηφίου. Μία *σχέση ενεργού υποψηφίου* εκφράζει το γεγονός ότι, από ένα σύνολο υποψηφίων πηγών για μία έννοια, μόνο ένας υποψήφιος επιλέγεται για να τροφοδοτήσει τη συγκεκριμένη έννοια. Ο υποψήφιος που τελικά επιλέγεται ονομάζεται *ενεργός υποψήφιος*. Κατά συνέπεια, μία σχέση ενεργού υποψηφίου είναι μία εξειδίκευση της σχέσης υποψηφίου με την ίδια δομή αλλά πιο αυστηρή σημασιολογία. Η σχέση ενεργού υποψηφίου συμβολίζεται γραφικά με ένα έντονο διακεκομμένο βέλος από την έννοια-πηγή προς την έννοια-στόχο.

Στο παράδειγμα του Σχήματος 4, υποθέτουμε ότι η πηγή `S2` διαθέτει περισσότερα από ένα συστήματα παραγωγής (π.χ. αρχεία COBOL), τα οποία χαρακτηρίζονται ως υποψήφια για το `S2.PARTSUPP`. Έτσι οι διαθέσιμοι υποψήφιοι (που απεικονίζονται στο πάνω αριστερά τμήμα του Σχήματος 4) είναι:

- Μία έννοια `AnnualPartSupp's`, η οποία περιέχει το πλήρες ετήσιο ιστορικό των προμηθευτών εξαρτημάτων (στην πράξη αντιπροσωπεύει ένα αρχείο `F1`). Η κύρια χρήση της αφορά σε αναφορές και περιέχει ένα υπερσύνολο των γνωρισμάτων που χρειάζονται για τις ανάγκες της αποθήκης δεδομένων.
- Μία έννοια `RecentPartSupp's`, η οποία περιέχει μόνο τα στοιχεία του τελευταίου μήνα (στην πράξη αντιπροσωπεύει ένα αρχείο `F2`). Χρησιμοποιείται συνεχώς από τους τελικούς χρήστες για την εισαγωγή ή την ανανέωση των δεδομένων, όπως επίσης και από μερικά προγράμματα αναφορών.

Στο Σχήμα 4 παρατηρούμε, επίσης, ότι η έννοια `RecentPartSupp's` επιλέχθηκε ως ενεργή υποψήφια. Παράλληλα, ένα σχόλιο περιγράφει τις λεπτομέρειες αυτής της σχεδιαστικής επιλογής.

Σχέσεις Παροχής. Μία *σχέση παροχής* απεικονίζει ένα σύνολο γνωρισμάτων εισόδου σε ένα σύνολο γνωρισμάτων εξόδου μέσω ενός μετασχηματισμού. Στην απλή `1:1` περίπτωση, οι σχέσεις παροχής εκφράζουν το γεγονός ότι ένα γνώρισμα εισόδου των πηγών δεδομένων μεταφέρεται σε ένα γνώρισμα εξόδου στην αποθήκη δεδομένων. Αν τα γνωρίσματα είναι φυσικά και σημασιολογικά ισοδύναμα, τότε δεν απαιτείται μετασχηματισμός. Σε αντίθετη περίπτωση, η αντιστοίχηση μεταξύ των γνωρισμάτων γίνεται μέσω κατάλληλου μετασχηματισμού (π.χ. μετατροπή ημερομηνίας από ευρωπαϊκή σε αμερικάνικη μορφή, έλεγχος για κενές τιμές, κ.λπ.).

Γενικά, υπάρχει η περίπτωση αλλαγής του σχήματος των δεδομένων εισόδου. Κατά συνέπεια, οι σχέσεις παροχής τυπικά ορίζονται από (α) ένα πεπερασμένο σύνολο γνωρισμάτων εισόδου, (β) ένα πεπερασμένο σύνολο γνωρισμάτων εξόδου, και (γ) ένα κατάλληλο μετασχηματισμό (τέτοιο ώστε τα γνωρίσματα εισόδου και εξόδου του να απεικονίζονται ένα προς ένα με τα αντίστοιχα γνωρίσματα της σχέσης).

Στην περίπτωση `1:1`, μία σχέση παροχής απεικονίζεται με ένα έντονο βέλος από το γνώρισμα εισόδου στο γνώρισμα εξόδου, ενώ πάνω στο βέλος σημειώνεται και ο μετασχηματισμός που χρησιμοποιείται.

Στη γενική περίπτωση `N:M`, μία σχέση προμηθευτή απεικονίζεται ως ένα σύνολο από έντονα βέλη με αρχή τα γνωρίσματα εισόδου, και τέλος τα γνωρίσματα εξόδου, μέσω του κατάλληλου μετασχηματισμού. Η γραφική αναπαράσταση της σχέσης παροχής `N:M` αποκρύπτει την απεικόνιση μεταξύ των γνωρισμάτων εισόδου και εξόδου. Για να αντισταθμιστεί αυτό το μειονέκτημα, η σύνδεση της σχέσης παροχής με κάθε ένα από τα εμπλεκόμενα γνωρίσματα εξόδου συνοδεύεται από μία ετικέτα, ώστε να μην υπάρχει αμφιβολία για τον προμηθευτή ενός γνωρίσματος εξόδου. Για παράδειγμα, στο Σχήμα 4 η σχέση παροχής που αφορά στο μετασχηματισμό γ είναι τύπου `N:M`. Για να μην υπάρχει αμφιβολία για την ένα προς ένα απεικόνιση των γνωρισμάτων εξόδου σε αυτά της εισόδου, η σύνδεση της σχέσης με τα γνωρίσματα εξόδου συνοδεύεται από μία ετικέτα για κάθε γνώρισμα εξόδου που φέρει το όνομα του αντίστοιχου γνωρίσματος εισόδου.

Τέλος, πρέπει να σημειώσουμε μία συντακτική προσθήκη στο μοντέλο μας. Μερικές φορές τυχαίνει μία συγκεκριμένη σχέση παροχής να περιλαμβάνει όλα τα γνωρίσματα ενός συνόλου από έννοιες. Για παράδειγμα, στην περίπτωση της ένωσης, όλες τα γνωρίσματα των εννοιών εισόδου και εξόδου συμμετέχουν στο μετασχηματισμό. Για να αποφύγουμε την υπερφόρτωση του σχήματος με υπερβολικά πολλές σχέσεις, εισάγουμε ένα συντακτικό συμβολισμό που απεικονίζει τις έννοιες εισόδου στις έννοιες εξόδου (αντί της απεικόνισης των γνωρισμάτων εισόδου στα γνωρίσματα εξόδου). Αυτό μπορεί να χρησιμοποιηθεί σαν μία λειτουργία σμίκρυνσης – μεγέθυνσης στο διάγραμμα. Στο πρώτο επίπεδο απεικονίζονται μόνο οι έννοιες και δίνεται μία γενική επισκόπηση του σεναρίου. Στο δεύτερο και πιο λεπτομερές επίπεδο, οι σχέσεις μεταξύ των εννοιών επεκτείνονται στις σχέσεις μεταξύ των εμπλεκόμενων γνωρισμάτων, οπότε και παρέχεται το σενάριο ΕΜΦ σε όλη του τη λεπτομέρεια.

Επιστρέφοντας και πάλι στο Σχήμα 4, εξετάζουμε τις σχέσεις μεταξύ των γνωρισμάτων των εννοιών `S1.PARTSUPP` και `S2.PARTSUPP`. Αρχικά, αγνοούμε τη συνάθροιση γ που πραγματοποιείται στα δεδομένα της πηγής `S2` και εξετάζουμε τους υπόλοιπους μετασχηματισμούς.

- Το γνώρισμα `PKey` δέχεται δεδομένα απ' ευθείας από το ομώνυμο γνώρισμα των `S1` και `S2`, μέσω ενός μετασχηματισμού ανάθεσης υποκατάστατου κλειδιού (`SK`). Η ανάθεση ενός υποκατάστατου κλειδιού αποτελεί συνηθισμένη τακτική στις αποθήκες δεδομένων και χρησιμοποιείται για την αντικατάσταση των κλειδιών των συστημάτων παραγωγής με ένα ομοιόμορφο κλειδί. Γενικά, οι βασικοί λόγοι για τη χρήση ενός μετασχηματισμού ανάθεσης υποκατάστατου κλειδιού είναι τόσο η επίδοση, όσο και η ομοιογένεια της σήμανσης. Για παράδειγμα, γνωρίσματα τύπου συμβολοακολουθίας γενικά δεν κρίνονται ως καλές περιπτώσεις κλειδιού για χρήση σε ευρετήριο και συνήθως απαιτείται η αντικατάσταση τους με κλειδιά αριθμητικού τύπου. Παράλληλα, διαφορετικά συστήματα παραγωγής ενδέχεται να χρησιμοποιούν διαφορετικά κλειδιά για τα ίδια αντικείμενα ή το ίδιο κλειδί για διαφορετικά αντικείμενα, με αποτέλεσμα την ανάγκη για συνολική αλλαγή αυτών των τιμών στην αποθήκη δεδομένων. Ας θεωρήσουμε την περίπτωση όπου το εξάρτημα `Τιμόνι` έχει `Pkey = 30` στην πηγή `S1` και `Pkey = 40` στην πηγή `S2`, ενώ στην πηγή `S2` το `Pkey = 30` να αντιστοιχεί στο εξάρτημα `Πόρτα`. Τέτοιες συγκρούσεις είναι εύκολο να λυθούν με ένα συνολικό μηχανισμό αντικατάστασης, μέσω της αντιστοίχησης ενός ομοιόμορφου υποκατάστατου κλειδιού.
- Το γνώρισμα `SuppKey` συνδέεται με τα ομώνυμα γνωρίσματα στις πηγές χωρίς να απαιτείται η χρήση μετασχηματισμού.
- Το γνώρισμα `Date` δέχεται δεδομένα από το ομώνυμο γνώρισμα της `S2`, μέσω ενός μετασχηματισμού `American_to_European_Date`. Παράλληλα, η ημερομηνία των εγγραφών που προέρχονται από την `S1` καθορίζεται από την εφαρμογή της συνάρτησης `SysDate()` (διότι η έννοια `S1.PARTSUPP` δεν περιέχει το γνώρισμα αυτό). Παρατηρήστε τη λειτουργία που εφαρμόζεται για τις εγγραφές που προέρχονται από την πηγή `S1`: δέχονται ως είσοδο όλα τα γνωρίσματα της `S1.PARTSUPP` (για να διαπιστωθεί ότι η παραγόμενη τιμή είναι ένα νέο γνώρισμα), περνούν από ένα μετασχηματισμό τύπου συνάρτησης που υπολογίζει την ημερομηνία συστήματος, και από εκεί καταλήγουν στο γνώρισμα `DW.PARTSUPP.Date`.
- Το γνώρισμα `Qty` παίρνει τις τιμές του απ' ευθείας από τα ομώνυμα γνωρίσματα των δύο πηγών χωρίς την ανάγκη κάποιου μετασχηματισμού.
- Το γνώρισμα `Cost` δέχεται δεδομένα από τα ομώνυμα γνωρίσματα των δύο πηγών. Όσον αφορά στην πηγή `S2`, εφαρμόζεται ένας μετασχηματισμός `$2€` για τη μετατροπή του κόστους των εξαρτημάτων σε ευρωπαϊκές τιμές. Όσον αφορά στην πηγή `S1`, εφαρμόζεται ένας μετασχηματισμός μη-κενής τιμής (`NN`), για να αποφευχθεί η φόρτωση στην αποθήκη δεδομένων εγγραφών που δεν περιλαμβάνουν κόστος εξαρτημάτων.

Σημειώστε ότι ενδέχεται να υπάρχουν γνωρίσματα εισόδου τα οποία αγνοούνται κατά τη διάρκεια διεργασιών ΕΜΦ, όπως για παράδειγμα το `S2.PARTSUPP.Department`.

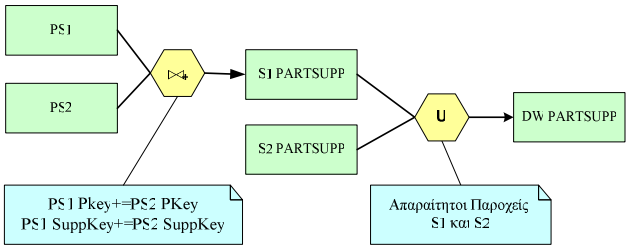
Σειριακή Σύνθεση Μετασχηματισμών. Συνήθως, οι διεργασίες ΕΜΦ περιλαμβάνουν περισσότερους από έναν μετασχηματισμούς για κάθε γνώρισμα. Επομένως, σε μία σχέση παροχής υπάρχει η ανάγκη για περισσότερους από έναν μετασχηματισμούς. Για παράδειγμα, θα μπορούσαμε να ομαδοποιήσουμε τα δεδομένα εισόδου ως προς κάποιο σύνολο γνωρισμάτων, έχοντας εξασφαλίσει, ταυτόχρονα, ότι δεν εμπλέκονται κενές τιμές στη διαδικασία αυτή. Σε μία τέτοια περίπτωση χρειάζεται να εφαρμόσουμε ένα μετασχηματισμό μη-κενής τιμής σε κάθε ένα από τα γνωρίσματα και στη συνέχεια να μεταφέρουμε μόνο τις σωστές σχέσεις στην ομαδοποίηση. Για την επίτευξη του σκοπού αυτού χρειάζεται η σειριακή εφαρμογή των προαναφερόμενων μετασχηματισμών. Ένα πρόβλημα που διαφαίνεται αφορά στην απαίτηση, σύμφωνα με τον ορισμό, ένας μετασχηματισμός να έχει ένα σύνολο γνωρισμάτων ως είσοδο και ένα σύνολο γνωρισμάτων ως έξοδο. Έτσι, η απλή σύνδεση δύο μετασχηματισμών φαίνεται ασυνεπής. Για να ξεπεραστεί αυτό, εισάγεται η έννοια της *σειριακής σύνθεσης των μετασχηματισμών*. Τυπικά, η σειριακή σύνθεση μετασχηματιστών περιλαμβάνει (α) ένα μοναδικό μετασχηματισμό έναρξης και (β) ένα μοναδικό μετασχηματισμό συνέχειας. Η σειριακή σύνθεση γραφικά απεικονίζεται με έντονες συνεχείς γραμμές που ενώνουν τους εμπλεκόμενους μετασχηματισμούς.

Ένα πιο σύνθετο μέρος του παραδείγματος του Σχήματος 4, είναι η συνάθροιση που εφαρμόζεται στα δεδομένα από την πηγή `S2`. Στο παράδειγμα αυτό, η πηγή `S2` κρατά πληροφορία για τους προμηθευτές εξαρτημάτων ανάλογα με το τμήμα στο οποίο ανήκουν. Η φόρτωση δεδομένων στην αποθήκη δεδομένων, η οποία αγνοεί αυτή τη λεπτομέρεια, απαιτεί την ομαδοποίηση των δεδομένων ανά `PKey`, `SuppKey` και `Date`, και τη συνάθροιση των `Cost` και `Qty`. Η συγκεκριμένη λειτουργία επιτυγχάνεται από το μετασχηματισμό συνάθροισης γ. Παράλληλα, οι προαναφερθέντες μετασχηματισμοί δεν αγνοούνται, αλλά ο καθένας μαζί με το μετασχηματισμό συνάθροισης γ αποτελούν σειριακή σύνθεση μετασχηματισμών. Σημειώνουμε, επίσης, τις ετικέτες στην έξοδο του μετασχηματισμού συνάθροισης που επισημαίνουν τους προμηθευτές δεδομένων στα αντίστοιχα γνωρίσματα εξόδου (π.χ. `S2.PARTSUPP.PKey` για το `DW.PARTSUPP.PKey` και `SUM(S2.PARTSUPP.Qty)` για το `DW.PARTSUPP.Qty`).

4. ΜΕΘΟΔΟΛΟΓΙΑ ΧΡΗΣΗΣ ΤΟΥ ΕΝΝΟΙΟΛΟΓΙΚΟΥ ΜΟΝΤΕΛΟΥ

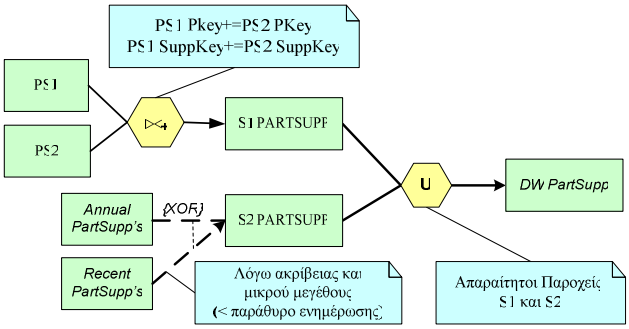
Στην ενότητα αυτή, παραθέτουμε τη σειρά βημάτων που ένας σχεδιαστής ακολουθεί κατά τη διάρκεια κατασκευής της αποθήκης δεδομένων. Κάθε βήμα αυτής της μεθοδολογίας θα παρουσιαστεί με αναφορά στο παράδειγμα του Σχήματος 4. Όπως έχει ήδη διευκρινιστεί, ο τελικός στόχος της διαδικασίας αυτής είναι η δημιουργία των απεικονίσεων μεταξύ των γνωρισμάτων, ο προσδιορισμός των απαραίτητων μετασχηματισμών, καθώς και η συλλογή οποιασδήποτε σχετικής βοηθητικής πληροφορίας.

Βήμα 1ο: Αναγνώριση των κατάλληλων πηγών δεδομένων. Το πρώτο πράγμα που αντιμετωπίζει ένας σχεδιαστής κατά την περίοδο ανάλυσης και συλλογής των απαιτήσεων της αποθήκης δεδομένων, είναι η αναγνώριση των σχετικών πηγών δεδομένων. Υποθέστε ότι για ένα υποσύνολο της αποθήκης δεδομένων, έχουμε προσδιορίσει την έννοια `DW.PARTSUPP` ως τον πίνακα συμβάντων για το πώς διανέμονται τα εξαρτήματα ανάλογα με τον προμηθευτή τους. Υποθέστε επίσης, ότι έχουμε αποφασίσει για λόγους πληρότητας, πως πρέπει να γεμίζουμε τον πίνακα του αποτελέσματος με δεδομένα και από τις δύο πηγές `S1` και `S2`, δηλαδή χρειαζόμαστε την ένωση των δεδομένων από τις δύο αυτές πηγές. Επιπλέον, για να φορτώσουμε δεδομένα στην έννοια `S1.PARTSUPP` απαιτείται μία εξωτερική συνένωση με τις έννοιες `PS1` και `PS2`. Με βάση τις προηγούμενες παρατηρήσεις, το διάγραμμα των πηγών που επιλέξαμε και των σχέσεων μεταξύ αυτών, φαίνεται στο Σχήμα 5.



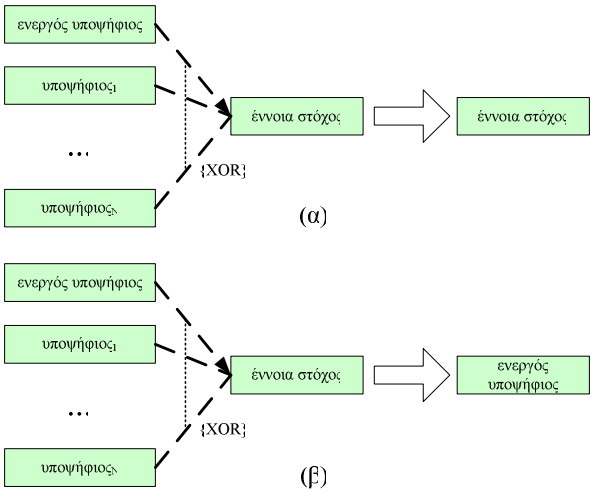
Σχήμα 5. Αναγνώριση των κατάλληλων πηγών δεδομένων

Βήμα 2ο: Υποψήφιοι και ενεργοί υποψήφιοι των εμπλεκομένων πηγών δεδομένων. Όπως ήδη έχουμε αναφέρει, κατά τη διάρκεια της ανάλυσης απαιτήσεων, ο σχεδιαστής ενδέχεται να ανακαλύψει ότι περισσότερες από μία πηγές δεδομένων μπορούν να είναι υποψήφιες για να συμπληρώσουν κάποια συγκεκριμένη έννοια. Ο τρόπος με τον οποίο λαμβάνεται μία τέτοια απόφαση είναι πέραν του σκοπού αυτής της εργασίας. Επιγραμματικά, όμως, αναφέρουμε ότι σημαντικό ρόλο σε τέτοιες αποφάσεις παίζουν κατά κύριο λόγο το πλήθος, η ποιότητα των δεδομένων, αλλά και η διαθεσιμότητα των πηγών. Στο παράδειγμα του Σχήματος 4 υποθέσαμε ότι η πηγή `S2` έχει περισσότερα από ένα συστήματα παραγωγής (όπως, για παράδειγμα, αρχεία COBOL), τα οποία είναι υποψήφια για την έννοια `S2.PARTSUPP`. Όπως είδαμε οι διαθέσιμες πηγές είναι: η έννοια `AnnualPartSupp's` (που περιέχει το πλήρες, ετήσιο, ιστορικό των προμηθευτών εξαρτημάτων) και η έννοια `RecentPartSupp's` (που περιέχει μόνο τα στοιχεία του τελευταίου μήνα). Οι υποψήφιες έννοιες για την έννοια `S2.PARTSUPP` και η (με βάση το σενάριο του παραδείγματος) ενεργή υποψήφια `RecentPartSupp's`, καθώς και τα αίτια, υπό τη μορφή σχολίου, για την επιλογή αυτή, απεικονίζονται στο Σχήμα 6.



Σχήμα 6. Υποψήφιοι και ενεργοί υποψήφιοι για τις εμπλεκόμενες πηγές δεδομένων

Προφανώς, όταν ληφθεί η απόφαση για τον ενεργό υποψήφιο, μπορεί να δημιουργηθεί ένα απλοποιημένο «λειτουργικό αντίγραφο» του σεναρίου στο οποίο αποβάλλονται όλοι οι άλλοι υποψήφιοι. Όπως φαίνεται στο Σχήμα 7, υπάρχουν δύο τρόποι να επιτευχθεί αυτό (α) αγνοώντας όλες τις πληροφορίες που αφορούν τους υποψηφίους για την έννοια – στόχο, και (β) αντικαθιστώντας την έννοια – στόχο με τον ενεργό υποψήφιο. Με άλλα λόγια, αντί ο σχεδιαστής να φορτώνει το διάγραμμα με πληροφορία όχι άμεσα χρήσιμη, μπορεί να χρησιμοποιήσει μία απλουστευμένη μορφή του. Οι κρυμμένοι υποψήφιοι δε σβήνονται, αντιθέτως παραμένουν μέρη του εννοιολογικού μοντέλου της εξεταζόμενης αποθήκης δεδομένων, ώστε να είναι δυνατό να χρησιμοποιηθούν σε μετέπειτα στάδια της λειτουργίας της αποθήκης δεδομένων. Ως παράδειγμα, αναφέρουμε την περίπτωση όπου πιθανές αλλαγές στον ενεργό υποψήφιο ενδέχεται να οδηγήσουν σε αναθεώρηση της επιλογής του ως ενεργού υποψηφίου. Το μόνο που αλλάζει είναι ότι απλά οι κρυμμένοι υποψήφιοι δε φαίνονται στο σχεδιαστή.



Σχήμα 7. Απλοποίηση διαγράμματος (α) αγνοώντας υποψήφιους, (β) θεωρώντας τον ενεργό υποψήφιο

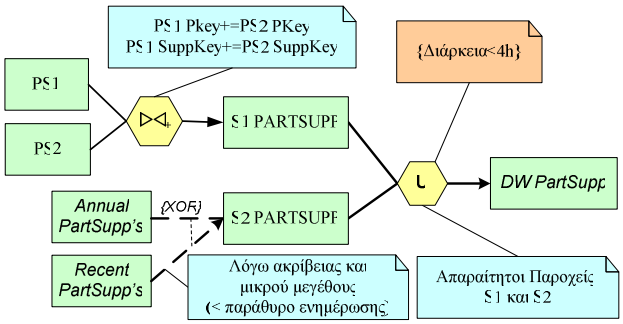
Βήμα 3ο: Αντιστοίχηση γνωρισμάτων μεταξύ πηγών και στόχων. Η πιο δύσκολη διαδικασία για τον σχεδιαστή της

αποθήκης δεδομένων είναι να καθορίσει την αντιστοίχηση των γνωρισμάτων των πηγών με αυτών της αποθήκης δεδομένων. Η διαδικασία αυτή περιλαμβάνει αρκετές συζητήσεις με τους διαχειριστές των πηγών δεδομένων ώστε να αποσαφηνιστούν σημεία όπως ο κώδικας, οι κανόνες ή οι τιμές, στοιχεία δηλαδή που συχνά είναι κρυμμένα στα δεδομένα ή στα προγράμματα των πηγών. Ακόμη, η φάση αυτή περιλαμβάνει αρκετές προσπάθειες «προεπισκόπησης δεδομένων» (με τη μορφή δειγματοληψίας ή με απλές ερωτήσεις καταμέτρησης) για να διαπιστωθούν πιθανά προβλήματα των δεδομένων που παρέχονται. Όπως είναι φυσικό, αυτή η διαδικασία είναι αλληλεπιδραστική και επιρρεπής σε λάθη. Η υποστήριξη που μπορεί να δοθεί στο χρήστη από ένα εργαλείο, εναπόκειται κυρίως στην αντιστοίχηση μεταξύ των εμπλεκομένων γνωρισμάτων, με το ενδιαφέρον να επικεντρώνεται στην επισήμανση των μετασχηματισμών και των εργασιών καθαρισμού που ενδέχεται να περικλείει αυτή η αντιστοίχηση.

Για κάθε γνώρισμα – στόχο πρέπει να οριστεί ένα σύνολο από σχέσεις παροχής. Σε απλές περιπτώσεις, η σχέση παροχής ορίζεται κατευθείαν μεταξύ αρχικού και τελικού γνωρίσματος. Στις περιπτώσεις όπου απαιτείται η χρήση μετασχηματισμού, ο μετασχηματισμός αυτός παρεμβάλλεται μεταξύ του αρχικού και του τελικού γνωρίσματος. Στις περιπτώσεις που απαιτούνται περισσότεροι του ενός μετασχηματισμοί, μεταξύ των εμπλεκόμενων γνωρισμάτων παρεμβάλλεται μία ακολουθία μετασχηματισμών, με τη μορφή σειριακής σύνθεσης,. Επίσης, σε αυτό το βήμα σχεδιασμού καθορίζονται και οι περιορισμοί ΕΜΦ. Ένα παράδειγμα των σχέσεων μεταξύ των γνωρισμάτων των εννοιών `PS1`, `PS2`, `S1.PARTSUPP`, `S2.PARTSUPP` και `DW.PARTSUPP` απεικονίζεται στο Σχήμα 8.

Βήμα 4ο: Εμπλουτισμός του διαγράμματος με περιορισμούς εκτέλεσης των διεργασιών. Εκτός από τον καθορισμό των εργασιών για ένα σενάριο ΕΜΦ που καθορίζει την απεικόνιση των πηγών στην αποθήκη δεδομένων μαζί με τους κατάλληλους μετασχηματισμούς, υπάρχουν και άλλες παράμετροι που πιθανόν να χρειάζεται να διευκρινιστούν για το περιβάλλον εκτέλεσης. Αυτού του είδους οι *περιορισμοί εκτέλεσης* περιλαμβάνουν:

- *Προγραμματισμός με βάση το χρόνο ή κάποιο γεγονός*. Ο σχεδιαστής χρειάζεται να αποφασίσει τη συχνότητα με την οποία θα εκτελείται η διεργασία ΕΜΦ, έτσι ώστε τα δεδομένα να είναι πάντα ανανεωμένα και η όλη διεργασία να εκτελείται στο διαθέσιμο χρόνο ανανέωσης.
- *Επίβλεψη*. Είναι απαραίτητη η συνεχής ροή πληροφοριών για την πρόοδο και την κατάσταση της διεργασίας, έτσι ώστε ο διαχειριστής να είναι ενήμερος για το στάδιο που βρίσκεται το φόρτωμα, το χρόνο έναρξής του, τη διάρκεια, κ.λπ.. Για το σκοπό αυτό, χρησιμοποιούνται μεταξύ άλλων: εγγραφές σε αρχεία, μηνύματα ενημέρωσης στην οθόνη ή με ηλεκτρονικό ταχυδρομείο, εκτυπώσεις ή γραφικές απεικονίσεις κ.α..
- *Καταγραφή Ιστορικού*. Καταγραφή πληροφοριών, οι οποίες παρουσιάζονται στο τέλος της εκτέλεσης του σεναρίου. Όλες οι σχετικές πληροφορίες (όπως τα προηγούμενα στοιχεία ανάλυσης) παρουσιάζονται χρησιμοποιώντας τις τεχνικές που προαναφέραμε.
- *Χειρισμός Εξαιρέσεων*. Η αντιμετώπιση της παραβίασης των κανόνων της βάσης δεδομένων ή των επιχειρησιακών κανόνων κατά εγγραφή είναι αναγκαίες για τη σωστή εκτέλεση των διεργασιών ΕΜΦ. Χρήσιμες για το σκοπό αυτό θεωρούνται πληροφορίες, όπως το ποιες εγγραφές είναι προβληματικές ή το πόσες εγγραφές είναι αποδεκτές (ποσοστό επιτυχίας).
- *Χειρισμός Σφαλμάτων*. Ανάνηψη από κατάρρευση και ικανότητες έναρξης και τερματισμού (αποθήκευση των συναλλαγών κάθε μερικές εγγραφές) είναι απολύτως απαραίτητες για τη στιβαρότητα και την αποδοτικότητα της διεργασίας.



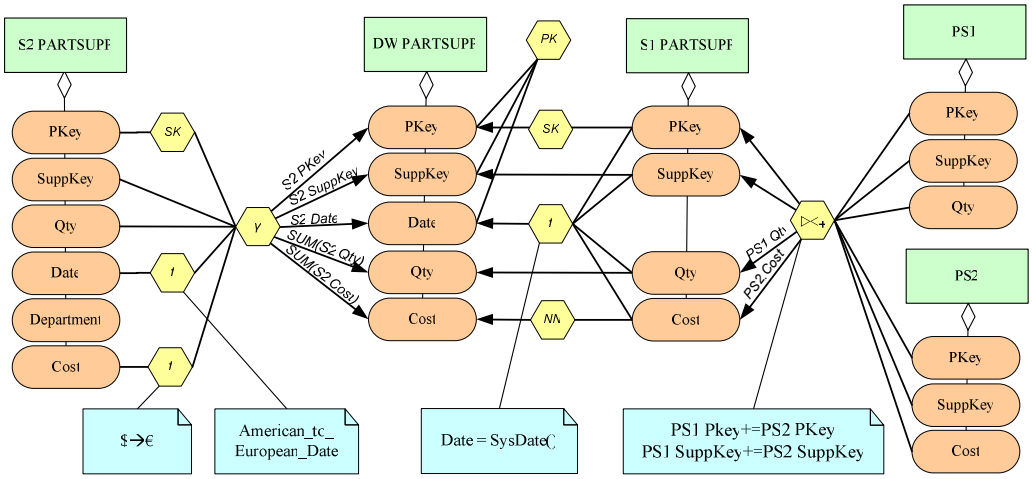
Σχήμα 9. Εμπλουτισμός του Σχήματος 6 με περιορισμό εκτέλεσης

Για την καταγραφή όλων των πληροφοριών, υιοθετούμε σημειώσεις τοποθετημένες στις κατάλληλες έννοιες, μετασχηματισμούς ή σχέσεις. Στο Σχήμα 9 απεικονίζουμε το διάγραμμα του Σχήματος 6, σημειώνοντας ένα περιορισμό χρόνου εκτέλεσης, δηλώνοντας ότι ο χρόνος εκτέλεσης για το φόρτωμα της `DW.PARTSUPP` (που περιλαμβάνει το φόρτωμα των `S1.PARTSUPP` και `S2.PARTSUPP`) δεν μπορεί να υπερβεί τις 4 ώρες.

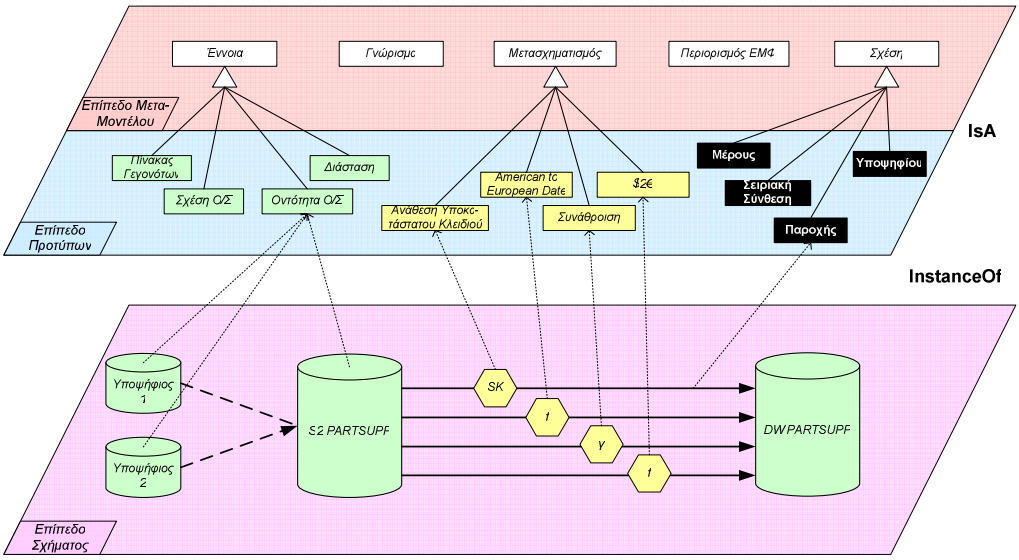
5. ΜΕΤΑ-ΜΟΝΤΕΛΟ

Πιστεύουμε ότι το σημείο κλειδί στην εννοιολογική απεικόνιση των διεργασιών ΕΜΦ στηρίζεται (α) στην αναγνώριση ενός μικρού συνόλου από *γενικούς κατασκευαστές*, ικανών να καλύψουν όλες τις περιπτώσεις (γενικότητα), και (β) στον προσδιορισμό ενός μηχανισμού που να επιτρέπει τη δημιουργία μίας «παλέτας» από *συχνά χρησιμοποιούμενους τύπους*, όπως μετασχηματισμούς, χώρους αποθήκευσης δεδομένων, κ.λπ. (επεκτασιμότητα).

Το Σχήμα 10 απεικονίζει το πλαίσιο μετα-μοντελοποίησης που εισάγουμε. Το κατώτατο επίπεδο του Σχήματος 10, που ονομάζεται *Επίπεδο Σχήματος*, περιλαμβάνει ένα σενάριο ΕΜΦ. Όλες οι οντότητες που υπάρχουν στο Επίπεδο Σχήματος αποτελούν στιγμιότυπα των κλάσεων *Έννοια*, *Γνώρισμα*, *Μετασχηματισμός*, *Περιορισμός ΕΜΦ* και *Σχέση*. Κατά συνέπεια, όπως φαίνεται στο πάνω μέρος του Σχήματος 10, εισάγουμε ένα επίπεδο μετα-κλάσης, που ονομάζουμε *Επίπεδο Μετα-Μοντέλου*, το οποίο περιλαμβάνει τις προαναφερθείσες κλάσεις. Η σύνδεση μεταξύ των δύο επιπέδων επιτυγχάνεται μέσω *Σχέσεων Στιγμιότυπων* (`InstanceOf`). Με την εισαγωγή του Επιπέδου Μετα-Μοντέλου ικανοποιούμε τη συνθήκη γενικότητας: οι πέντε κλάσεις που περιλαμβάνονται στο Επίπεδο Μετα-Μοντέλου είναι αρκετά γενικές για να μοντελοποιήσουν οποιοδήποτε σενάριο ΕΜΦ, μέσω της δημιουργίας κατάλληλων στιγμιότυπων.



Σχήμα 8. Αντιστοιχήσεις γνωρισμάτων για τη φόρτωση της έννοιας DW.PARTSUPP



Σχήμα 10. Το πλαίσιο μοντελοποίησης διεργασιών ΕΜΦ

Φίλτρα	**Μοναδιαίοι μετασχηματισμοί**	**Δυαδικοί μετασχηματισμοί**
Selection (σ)	Push	Union (U)
Not null (NN)	Aggregation (γ)	Join (⋈)
Primary key violation (PK)	Projection (π)	Diff (Δ)
Foreign key violation (FK)	Function application (f)	Update Detection (Δ_{UPD})
Unique value (UN)	Surrogate key assignment (SK)	
Domain mismatch (DM)	Tuple normalization (N)	**Σύνθετοι μετασχηματισμοί**
	Tuple denormalization (DN)	Slowly changing dimension (Type 1,2,3)(SDC-1/2/3)
Λειτουργίες Μεταφοράς		Format mismatch (FM)
Ftp (FTP)	**Λειτουργίες Επεξεργασίας Αρχείων**	Data type conversion (DTC)
Compress/Decompress (Z/dZ)	EBCDIC to ASCII conversion (EB2AS)	Switch (σ*)
Encrypt/Decrypt (Cr/dCr)	Sort file (Sort)	Extended union (U)

Σχήμα 11. Πρότυπα μετασχηματισμών και οι συμβολισμοί τους, ανά κατηγορίες

Ωστόσο, μπορούμε να βελτιώσουμε το μετα-μοντέλο για να καλύψουμε και τη συνθήκη επεκτασιμότητας. Προκειμένου να καταστήσουμε το μοντέλο μας πραγματικά χρήσιμο για πρακτικές περιπτώσεις διεργασιών ΕΜΦ, το εμπλουτίζουμε με ένα σύνολο εξειδικευμένων κατασκευαστών ΕΜΦ και εισάγουμε ένα τρίτο επίπεδο: το *Επίπεδο Προτύπων*. Οι κατασκευαστές στο Επίπεδο Προτύπων είναι επίσης μετα-κλάσεις, αρκετά παραμετροποιήσιμες και αποτελούν υποσύνολο των μετα-κλάσεων του Επιπέδου Μετα-Μοντέλου. Κατά συνέπεια, οι κλάσεις του Επιπέδου Προτύπων, αποτελούν εξειδικεύσεις (υποκλάσεις) των γενικών κλάσεων του επιπέδου Μετα-Μοντέλου, κάτι που φαίνεται ως σχέση IsA στο Σχήμα 10. Μέσω αυτού του μηχανισμού προσαρμογής, ο σχεδιαστής επιλέγει τα στιγμιότυπα του Επιπέδου Σχήματος από μία πλουσιότερη «παλέτα» κατασκευαστών. Με αυτή τη ρύθμιση, οι οντότητες στο Επίπεδο Σχήματος είναι στιγμιότυπα, όχι μόνο των αντίστοιχων κλάσεων του επιπέδου Μετα-Μοντέλου, αλλά και των υποκλάσεων του Επιπέδου Προτύπων.

Στο παράδειγμα του Σχήματος 10, η έννοια DW.PARTSUPP πρέπει να φορτωθεί από μία συγκεκριμένη πηγή S2. Αρκετοί μετασχηματισμοί λαμβάνουν χώρα κατά τη διάρκεια αυτής της διάδοσης δεδομένων. Για παράδειγμα, σε αυτό το σενάριο ΕΜΦ χρησιμοποιούνται, εκτός των άλλων, μία ανάθεση υποκατάστατου κλειδιού και μία συνάθροιση. Επιπλέον, υπάρχουν δύο υποψήφιοι για την έννοια S2.PARTSUPP. Από αυτούς τους δύο, μόνο ένας (Υποψήφιος 2) επιλέγεται τελικά για τη διαδικασία αυτή. Όπως κάποιος μπορεί να παρατηρήσει, οι έννοιες που λαμβάνουν μέρος σε αυτό το σενάριο είναι στιγμιότυπα της κλάσης Έννοια (που ανήκει στο επίπεδο Μετα-Μοντέλου) και συγκεκριμένα της υποκλάσης αυτής Οντότητα Ο/Σ (υποθέτοντας ότι υιοθετούμε μία επέκταση του μοντέλου Ο/Σ). Στιγμιότυπα και επικαλυπτόμενες κλάσεις σχετίζονται μέσω συνδέσμων τύπου InstanceOf. Ο ίδιος μηχανισμός εφαρμόζεται σε όλους τους μετασχηματισμούς του σεναρίου αυτού, οι οποίοι είναι (α) στιγμιότυπα της κλάσης Μετασχηματισμός και (β) στιγμιότυπα κάποιας από τις υποκλάσεις του που απεικονίζονται στο Επίπεδο Προτύπων του Σχήματος 10. Ούτε οι σχέσεις ξεφεύγουν από αυτό τον κανόνα. Για παράδειγμα, παρατηρείστε πώς οι συνδέσεις παροχής από την έννοια S2.PARTSUPP προς στην έννοια DW.PARTSUPP σχετίζονται με την κλάση Σχέση Παροχής μέσω των κατάλληλων συνδέσεων InstanceOf. Προφανώς για λόγους απλότητας στο Σχήμα 10 δεν απεικονίζονται όλες οι δυνατές συσχετίσεις μεταξύ των τριών επιπέδων μοντελοποίησης.

Όσον αφορά την κλάση Έννοια, στο Επίπεδο Προτύπων, μπορούμε να την εξειδικεύσουμε σε αρκετές υποκλάσεις, ανάλογα με το χρησιμοποιούμενο μοντέλο. Στην περίπτωση του μοντέλου Ο/Σ, έχουμε τις υποκλάσεις Οντότητα Ο/Σ και Σχέση Ο/Σ, ενώ στην περίπτωση του μοντέλου διαστάσεων, έχουμε υποκλάσεις όπως Πίνακας Γεγονότων και Πίνακας Διαστάσεων.

Παρόμοια, η κλάση Μετασχηματισμός εξειδικεύεται περισσότερο σε ένα σύνολο από επαναχρησιμοποιούμενες διεργασίες ΕΜΦ, που απεικονίζεται στο Σχήμα 11. Στο Σχήμα 11 ομαδοποιούμε τα πρότυπα διεργασιών σε έξι κύριες, λογικές κατηγορίες. Στο Σχήμα 10 δεν απεικονίζεται αυτή η ομαδοποίηση για την αποφυγή υπερφόρτωσης του σχήματος. Αντίθετα, στο εν λόγω σχήμα απεικονίζονται μόνο τέσσερις υποκλάσεις, των οποίων τα στιγμιότυπα παρουσιάζονται στο σενάριο του Επιπέδου Σχήματος.

Η πρώτη ομάδα ονομάζεται *Φίλτρα* και περιλαμβάνει ελέγχους ως προς κάποια συγκεκριμένη συνθήκη. Οι σημασίες αυτών των φίλτρων είναι οι προφανείς: ξεκινώντας από ένα γενικό φίλτρο *επιλογής* (σ) και συνεχίζοντας με ελέγχους *μη-κενής τιμής* (NN), παραβίασης *πρωτεύοντος* (PK) ή *ξένου κλειδιού* (FK), κ.λπ..

Η δεύτερη ομάδα προτύπων μετασχηματισμών, ονομάζεται *Μοναδιαίοι Μετασχηματισμοί* και εκτός από τη γενική δραστηριότητα *προώθησης δεδομένων* (push) (η οποία απλά μεταφέρει δεδομένα από τον προμηθευτή στον καταναλωτή), περιλαμβάνει επίσης τους μετασχηματισμούς: *προβολή* (π), *συνάθροιση* (γ) και *συνάρτηση* (f). Επιπλέον, περιλαμβάνει και τρεις μετασχηματισμούς εξειδικευμένους για τις αποθήκες δεδομένων: την *ανάθεση υποκατάστατου κλειδιού* (SK), την *κανονικοποίηση* (N) και την *αποκανονικοποίηση* (DN).

Η τρίτη ομάδα είναι οι *Δυαδικοί Μετασχηματισμοί* και περιλαμβάνει πρότυπα μετασχηματισμών όπως *ένωση* (U), *συνένωση* (⋈) και *διαφορά* (Δ), καθώς και μία ειδική περίπτωση για τη διαφορά που περιλαμβάνει την *αναγνώριση των ενημερώσεων* (Δ_{UPD}).

Επίσης, υπάρχει ένα σύνολο από *Σύνθετους Μετασχηματισμούς* που περιλαμβάνει πρότυπα μετασχηματισμών κατασκευασμένων από τη σύνθεση άλλων απλούστερων. Για παράδειγμα, αναφέρουμε μετασχηματισμούς όπως: ο *χειρισμός αργά μεταβαλλόμενων διαστάσεων* (SCD) που είναι μία τεχνική για τη φόρτωση διαστάσεων με καινούρια ή αλλαγμένα δεδομένα, η *ασυμβατότητα μορφής* (FM) που είναι μία ειδική περίπτωση συνάρτησης, η *εναλλαγή* (σ^*) που είναι ο συνδυασμός αρκετών επιλογών και τέλος η *εκτεταμένη ένωση* (U) που είναι ένας συνδυασμός από δυαδικές ενώσεις για την παραγωγή του αντίστοιχου ν-αδικού τελεστή.

Εκτός από τα προαναφερθέντα πρότυπα μετασχηματισμών, που αφορούν κυρίως σε λογικούς μετασχηματισμούς, αναφέρουμε και την περίπτωση προτύπων που αφορούν στην εφαρμογή φυσικών μετασχηματισμών σε ολόκληρα αρχεία ή πίνακες. Ως τέτοιους μετασχηματισμούς θεωρούμε φυσικές πράξεις μεταξύ εννοιών όπως οι *Λειτουργίες Μεταφοράς* (FTP, Z/dZ, Cr/dCr) και οι *Λειτουργίες Επεξεργασίας Αρχείων* (EB2AS, Sort).

Συνοψίζοντας, το Επίπεδο Μετα-Μοντέλου είναι ένα σύνολο από γενικές οντότητες, ικανό να αναπαραστήσει οποιοδήποτε σενάριο ΕΜΦ. Παράλληλα, η γενικότητα του Επιπέδου Μετα-Μοντέλου συμπληρώνεται με την επεκτασιμότητα του Επιπέδου Προτύπων, που είναι ένα σύνολο από προκατασκευασμένα και εξειδικευμένα πρότυπα μετασχηματισμών που είναι ειδικά σχεδιασμένα, ώστε να ανταποκρίνονται στους πιο συχνά χρησιμοποιούμενους μετασχηματισμούς των διεργασιών ΕΜΦ. Όμως, πρέπει να τονιστεί ότι η «παλέτα» μετασχηματισμών του Σχήματος 11 σε καμία περίπτωση δεν περιλαμβάνει ολόκληρο το σύνολο των μετασχηματισμών που ενδέχεται να απαιτηθούν σε ένα σενάριο ΕΜΦ. Η ιδέα γύρω από την έννοια της επεκτασιμότητας και του τρόπου σύλληψης του μοντέλου στηρίζεται στο γεγονός ότι ο σχεδιαστής μπορεί να δημιουργήσει τα δικά του πρότυπα, που δεν συμπεριλαμβάνονται μεν στο Σχήμα 11, αλλά που ενδέχεται να στηρίζονται ή όχι σε αυτά, δίχως να επηρεάζεται η λειτουργικότητα ή ο τρόπος χρήσης του προτεινόμενου μοντέλου.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Τα εργαλεία Εξαγωγής-Μετασχηματισμού-Φόρτωσης (ΕΜΦ) είναι προγράμματα υπεύθυνα για την εξαγωγή των δεδομένων από διάφορες πηγές, τον καθαρισμό, την προσαρμογή και την εισαγωγή τους σε μία αποθήκη δεδομένων. Σε αυτή την εργασία, εστιάσαμε την προσοχή μας στο πρόβλημα του ορισμού των διεργασιών ΕΜΦ και παρουσιάσαμε τα θεμέλια για την εννοιολογική τους αναπαράσταση. Συγκεκριμένα, περιγράφηκε ένα καινούριο εννοιολογικό μοντέλο προσανατολισμένο στην ταυτοποίηση των σχέσεων μεταξύ των γνωρισμάτων και των εννοιών, αλλά και στην αναγνώριση των κατάλληλων διεργασιών ΕΜΦ στα πρώτα στάδια ενός έργου σχεδίασης και ανάπτυξης μίας αποθήκης δεδομένων. Το προτεινόμενο μοντέλο κατασκευάστηκε με τρόπο προσαρμόσιμο και επεκτάσιμο, ώστε να μπορεί ο σχεδιαστής να το ενισχύσει με δικές του, επαναχρησιμοποιούμενες διεργασίες ΕΜΦ, όπως η ανάθεση υποκατάστατων κλειδιών, ο έλεγχος για παραβάσεις αναφορικής ακεραιότητας κ.λπ..

Ως μελλοντική εργασία θεωρούμε τη σύνδεση του μοντέλου που περιγράφηκε σε αυτό το άρθρο με το λογικό μοντέλο που περιγράφεται στα [28, 31], καθώς και τη βελτιστοποίηση σεναρίων ΕΜΦ σε λογικό και φυσικό επίπεδο.

7. ΑΝΑΦΟΡΕΣ

[1] M. Bouzeghoub, F. Fabret, M. Matulovic. Modeling Data Warehouse Refreshment Process as a Workflow Application. In proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg, Germany, 1999.

[2] V. Borkar, K. Deshmuk, S. Sarawagi. Automatically Extracting Structure from Free Text Addresses. Bulletin of the Technical Committee on Data Engineering, **23**(4), 2000.

[3] G. Booch, I. Jacobson, J. Rumbaugh. The Unified Modeling Language User Guide. Addison-Wesley Pub Co., ISBN: 0201571684, 1st edition, 1998.

[4] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Information integration: Conceptual modeling and reasoning support. In proceedings of the 6th International Conference On Cooperative Information Systems, pp. 280-291, 1998.

[5] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati. A principled approach to data integration and reconciliation in data warehousing. In proceedings of International Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg, Germany, 1999.

[6] H. Galhardas, D. Florescu, D. Shasha and E. Simon. Ajax: An Extensible Data Cleaning Tool. In proceedings of ACM SIGMOD International Conference On the Management of Data, pp. 590, Dallas, Texas, 2000.

[7] M. Golfarelli, D. Maio, S. Rizzi. The Dimensional Fact Model: a Conceptual Model for Data Warehouses. Invited Paper, International Journal of Cooperative Information Systems, vol. 7, n. 2&3, 1998.

[8] M. Golfarelli, S. Rizzi: Methodological Framework for Data Warehouse Design. In ACM 1st International Workshop on Data Warehousing and OLAP (DOLAP '98), pp. 3-9, November 1998, Bethesda, Maryland, USA.

[9] B. Husemann, J. Lechtenborger, G. Vossen. Conceptual data warehouse modeling. In proceedings of the 2nd International Workshop on Design and Management of Data Warehouses (DMDW), pp. 6.1-6.11, Stockholm, Sweden, 2000.

[10] M.A. Jeusfeld, C. Quix, M. Jarke: Design and Analysis of Quality Information for Data Warehouses. In proceedings of the 17th International Conference On Conceptual Modeling (ER'98), pp. 349-362, Singapore, 1998.

[11] M. Jarke, M.A. Jeusfeld, C. Quix, P. Vassiliadis: Architecture and quality in data warehouses: An extended repository approach. Information Systems, 24(3): 229-253 (1999). A previous version appeared in proceedings of the 10th Conference of Advanced Information Systems Engineering (CAiSE '98), Pisa, Italy, 1998.

[12] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis (eds.). Fundamentals of Data Warehouses. Springer, 2000.

[13] R. Kimball. A Dimensional Modeling Manifesto. DBMS Magazine. August 1997.

[14] R. Kimbal, L. Reeves, M. Ross, W. Thornthwaite. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons, February 1998.

[15] W. Labio, J.L. Wiener, H. Garcia-Molina, V. Gorelik. Efficient Resumption of Interrupted Warehouse Loads. In proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD 2000), pp. 46-57, Dallas, Texas, USA, 2000.

[16] D.L. Moody, M.A.R. Kortink: From enterprise models to dimensional models: a methodology for data warehouse and data mart design. In proceedings of the 2nd International Workshop on Design and Management of Data Warehouses, DMDW 2000, Stockholm, Sweden, 2000.

[17] A. Monge. Matching Algorithms Within a Duplicate Detection System. Bulletin of the Technical Committee on Data Engineering, **23**(4), 2000.

[18] T. B. Nguyen, A Min Tjoa, R. R. Wagner. An Object Oriented Multidimensional Data Model for OLAP. In proceedings of the 1st International Conference on Web-Age Information Management (WAIM-00), Shanghai, China, 2000.

[19] E. Rahm, H. Do. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering, **23**(4), 2000.

[20] V. Raman, J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation. Technical Report University of California at Berkeley, Computer Science Division, 2000. Available at http://www.cs.berkeley.edu/~rshankar/papers/pwheel.pdf

[21] C. Sapia, M. Blaschka, G. Höfling, B. Dinter: Extending the E/R Model for the Multidimensional Paradigm. In ER Workshops 1998, pp. 105-116. Lectures Notes in Computer Science 1552 Springer, 1999.

[22] N. Tryfona, F. Busborg, J.G.B. Christiansen. starER: A Conceptual Model for Data Warehouse Design. In ACM Second International Workshop on Data Warehousing and OLAP (DOLAP '99), pp. 3-8, Kansas City, Missouri, USA, 1999.

[23] J.C. Trujillo, M. Palomar, J. Gómez: Applying Object-Oriented Conceptual Modeling Techniques to the Design of Multidimensional Databases and OLAP Applications. In proceedings of the 1st International Conference on Web-Age Information Management (WAIM-00) pp. 83-94, Shanghai, China, 2000.

[24] A. Tsois. MAC: Conceptual data modeling for OLAP. In proceedings of the 3rd International Workshop on Design and Management of Data Warehouses (DMDW), pp. 5.1–5.11, Interlaken, Switzerland, 2001.

[25] A. Simitsis. List of ETL Tools. Available at: http://www.dbnet.ece.ntua.gr/~asimi/ETLTools.htm

[26] A. Simitsis, P. Vassiliadis. A Methodology for the Conceptual Modeling of ETL Processes. In proceedings of the Decision Systems Engineering (DSE '03), Velden, Austria, 2003.

[27] P. Vassiliadis. Gulliver in the land of data warehousing: practical experiences and observations of a researcher. In proceedings of the 2nd International Workshop on Design and Management of Data Warehouses (DMDW), pp. 12.1 – 12.16, Stockholm, Sweden, 2000.

[28] P. Vassiliadis, A. Simitsis, S. Skiadopoulos. Modeling ETL activities as graphs. In proceedings of Design and Management of Data Warehouses (DMDW'2002) 4th International Workshop in conjunction with CAiSE'02, pp. 52-61, Toronto, Canada, 2002.

[29] P. Vassiliadis, A. Simitsis, S. Skiadopoulos. Conceptual Modeling for ETL Processes. In proceedings of the 5th Data Warehousing and OLAP (DOLAP '02), McLean, VA, USA, 2002.

[30] P. Vassiliadis, C. Quix, Y. Vassiliou, M. Jarke. Data Warehouse Process Management. Information Systems, vol. 26, no.3, pp. 205-236, 2001.

[31] P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, S. Skiadopoulos. A Generic and Customizable Framework for the Design of ETL Scenarios. To appear in Information Systems Journal.

[32] P. Vassiliadis, Z. Vagena, S. Skiadopoulos, N. Karayannidis, T. Sellis. ARKTOS: Towards the modeling, design, control and execution of ETL processes. Elsevier Science Ltd, Information Systems, **26**(8), pp. 537-561, 2001.

Γλωσσάριο

Αγγλικός όρος	Μετάφραση
Active Candidate	Ενεργός Υποψήφιος
Attribute	Γνώρισμα
Candidate	Υποψήφιος
Concept	Έννοια
Data Staging Area (DSA)	Μεταβατική Περιοχή Αποθήκευσης Δεδομένων (ΜΠΑΔ)
Dimension Table	Πίνακας Διάστασης
ETL Constraint	Περιορισμός ΕΜΦ
Extract – Transform – Load (ETL)	Εξαγωγή – Μετασχηματισμός – Φόρτωση (ΕΜΦ)
Fact Table	Πίνακας Πληροφοριών
Note	Σχόλιο
On-Line Analytical Processing (OLAP)	Σύγχρονη Αναλυτική Επεξεργασία Δεδομένων (ΣΑΕΔ)
Part-Of Relationship	Σχέση Μέρους
Provider Relationship	Σχέση Παροχής
Serial Composition	Σειριακή Σύνθεση
Star Schema	Σχήμα Αστέρα
Transformation	Μετασχηματισμός