Έργο:	**«ΘΑΛΗΣ: Ενίσχυση της Διεπιστημονικής ή και Διιδρυματικής έρευνας και καινοτομίας με δυνατότητα προσέλκυσης ερευνητών υψηλού επιπέδου από το εξωτερικό μέσω της διενέργειας βασικής και εφαρμοσμένης έρευνας αριστείας»**
Τίτλος Υποέργου:	**«ΕΙΚΟΣ»: Θεωρητική και αλγοριθμική θεμελίωση για Προσωποκεντρικά Συνεργατικά Πληροφοριακά Συστήματα**

Παραδοτέο Π.5.3

Σχεδίαση οικοσυστημάτων πληροφορίας γύρω από υπερχώρους δεδομένων

Σεπτέμβριος 2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Δράση 5	**Υποστήριξη εξέλιξης της πληροφορίας και αυτορύθμισης συστημάτων**				
Ομάδα	Ερ. Ομάδα 5	**Έναρξη**	01/06/2012	**Λήξη**	30/11/2015

Συντονιστής ΕΟ5	Π. Βασιλειάδης (Παν. Ιωαννίνων)	
Υποδράση: Δ 5.3	Σχεδίαση οικοσυστημάτων πληροφορίας γύρω από υπερχώρους δεδομένων	
Συμμετέχοντες	*Μέλη ΚΕΟ*	Π. Βασιλειάδης, Ι. Βασιλείου, Α. Γούναρης, Ι. Σταύρακας
	Μέλη ΟΕΣ	Γ. Παπαστεφανάτος (ΕΚ. ΑΘΗΝΑ\ ΙΠΣΥ), Πέτρος Μανούσης, (Παν. Ιωαννίνων), Γεωργία Κούγκα (ΑΠΘ), Θεοδώρα Γαλάνη (ΕΜΠ)
Σύντομη Περιγραφή	Η Υποδράση 5.3 διερευνά τη δυνατότητα παροχής μεθόδων που να μας επιτρέπουν (α) να αποτιμήσουμε την ποιότητα της σχεδίασης ενός οικοσυστήματος δεδομένων με το βλέμμα στην εξέλιξή του και (β) να μας επιτρέπουν να συστήσουμε στο χρήστη / διαχειριστή ενός οικοσυστήματος πρότυπα και τεχνικές κατάλληλα ώστε να ελαχιστοποιήσουν το κόστος ανασυγκρότησης του οικοσυστήματος για ένα πιθανό σύνολο εξελικτικών δράσεων στο μέλλον.	
Παραδοτέο	Π.5.3 Σχεδίαση οικοσυστημάτων πληροφορίας γύρω από υπερχώρους δεδομένων	
Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 1 δημοσίευση.	
Επίτευξη στόχου	100%	

Περιεχόμενα

1 Εισαγωγή

Ο βασικός στόχος του έργου ΕΙΚΟΣ είναι να προσφέρει τη μεθοδολογία, τη θεωρητική θεμελίωση, τις αλγοριθμικές τεχνικές και την αρχιτεκτονική του λογισμικού που απαιτείται ώστε τα πληροφοριακά συστήματα να μπορούν να προσφέρουν στους χρήστες αφενός την δυνατότητα εξατομίκευσης της παρεχόμενης πληροφορίας και αφετέρου τη δυνατότητα χρήσης ενσωματωμένων ετερογενών δεδομένων, ενδεχομένως διαφορετικής προέλευσης, με διαφανή τρόπο.

Στα πλαίσια του έργου, η Δράση 5 «Υποστήριξη εξέλιξης της πληροφορίας και αυτορρύθμισης συστημάτων» προσφέρει στο έργο αλγοριθμικά αποτελέσματα για τη σχεδίαση και την προσαρμογή ενός–οικοσυστήματος δεδομένων σε ότι αφορά εξελικτικές μεταβολές που αφορούν στη σημασιολογία, δομή και περιεχόμενο της πληροφορίας. Η Δράση οργανώνεται σε τρεις θεμελιώδεις δράσεις, εκ των οποίων η πρώτη αφορά τη μοντελοποίηση των γεγονότων εξέλιξης και των αλληλεξαρτήσεών τους, η δεύτερη την ρύθμιση και αυτοματοποίηση της εξέλιξης ενός οικοσυστήματος δεδομένων και η τρίτη την a-priori σχεδίασή του, έχοντας υπόψη τις διαθέσιμες τεχνικές προσαρμογής από την προηγούμενη υποδράση.

Το παρόν Παραδοτέο Π.5.3 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ5.3. Στην ενότητα 1 παρουσιάζουμε το γενικότερο πλαίσιο του προβλήματος. Στην ενότητα 2 διερευνούμε τη διαχείριση και βελτιστοποίηση ροών εργασίας ενός οικοσυστήματος. Στην ενότητα 3 διερευνούμε γραφοθεωρητικές μετρικές που μας επιτρέπουν να δούμε ποια σημεία ενός οικοσυστήματος είναι ευαίσθητα στην εξέλιξή του. Στην ενότητα 4 ερευνούμε ποιοι μπορεί να είναι οι «νόμοι» που διέπουν την εξέλιξη του οικοσυστήματος και στην ενότητα 5 ασχολούμαστε με την οπτικοποίηση της εσωτερικής αρχιτεκτονική ενός οικοσυστήματος. Ανακεφαλαιώνουμε τη συνεισφορά μας στην ενότητα 6.

1.1 Πλαίσιο έρευνας

Στις προηγούμενες υποδράσεις ασχοληθήκαμε εκτενώς με το πώς μπορούμε να μοντελοποιήσουμε (α) την εξέλιξη των δεδομένων αυτή καθεαυτή και (β)

οικοσυστήματα δεδομένων, με ερωτήσεις επί των δεδομένων ενταγμένες στον κώδικα εφαρμογών ώστε να διαχειριστούμε την εξέλιξή τους.

Όμως, ένα σημαντικό χαρακτηριστικό των μοντέρνων οικοσυστημάτων, αυτό της δυναμικές διαχείρισης των δεδομένων μέσω ροών εργασίας που τα επεξεργάζονται, δεν είχε καλυφθεί. Στην Υποδράση 5.3, λοιπόν, καλούμαστε να αντιμετωπίσουμε το πρόβλημα αυτό. Το ερώτημα που μας απασχολεί στην παρούσα υποδράση, λοιπόν, είναι τριπλό. Κατ αρχήν, πρέπει να μοντελοποιήσουμε τις ροές εργασίας που διαχειρίζονται δεδομένα. Έπειτα, πρέπει να βρούμε μεθόδους που να μπορούν, σε λογικό επίπεδο, ανεξάρτητα από την πλατφόρμα εκτέλεσής τους, να βοηθήσουν το σχεδιαστή να βελτιστοποιήσει την εκτέλεσή τους. Τρίτον, πρέπει να εισάγουμε μεθόδους, που να λαμβάνουν υπ' όψιν την πλατφόρμα εκτέλεσης (στις μέρες μας, συχνά, ένα νέφος από ετερογενή και ανεξάρτητα μηχανήματα που εκτελούν μια ροή με παράλληλο τρόπο).

Ναι, αλλά, έχοντας αυτό ως κεκτημένο, πώς εξελίσσονται οι εν λόγω ροές επεξεργασίας των δεδομένων? Μπορούμε να αποτιμήσουμε με κάποιο τρόπο αν η σχεδίασή μας είναι ανθεκτική σε πιθανές αλλαγές του οικοσυστήματος, ή, μήπως, μια μικρή αλλαγή (π.χ., η προσθήκη ή η διαγραφή ενός πεδίου) μπορεί να οδηγήσει στην ανάγκη να συντηρήσουμε ένα μεγάλο αριθμό από δραστηριότητες (activities -- δλδ., μονάδες λογισμικού), που συνεργάζονται για να εκτελεσθεί η ροή? Όπως αναπτύξαμε διεξοδικά και στο προηγούμενο παραδοτέο Π5.2, μια αλλαγή στη δομή («σχήμα») των δεδομένων μπορεί να καταστήσει όλο το λογισμικό που τα χρησιμοποιεί συντακτικά και σημασιολογικά λάθος. Στις ροές εργασίας, όπου η μία δραστηριότητα τροφοδοτεί την επόμενη με δεδομένα, μια τέτοια αλλαγή μπορεί να έχει εξαιρετικά ευρεία διάδοση και αντίστοιχα, επιπτώσεις και ανάγκη συντήρησης του λογισμικού.

Ένα άλλο ερώτημα που μας απασχόλησε είναι η κατανόηση των νόμων που διέπουν την εξέλιξη των οικοσυστημάτων στον πραγματικό κόσμο. Υπάρχουν κανόνες / συχνά εμφανιζόμενα πρότυπα που μπορούν να μας παρέχουν μια

αρχική εικόνα για το πώς ένα οικοσύστημα μπορεί να αλλάζει στο χρόνο? Πριν ασχοληθούμε με το πρόβλημα αυτό, η σχετική εμπειρία από την πλευρά των βάσεων δεδομένων ήταν εξαιρετικά μικρή: ελάχιστες καταγραφές για την ιστορία κάποιας βάσης δεδομένων υπήρχαν. Από την πλευρά της τεχνολογίας λογισμικού, σε σχέση με το παραδοσιακό λογισμικό όμως, η κατανόηση ήταν βαθύτερη. Το βασικό υπόβαθρο της επιστήμης στο χώρο της τεχνολογίας λογισμικού για την εξέλιξη του λογισμικού συγκεντρώνεται στους νόμους του Lehamn, που είναι ένα σύνολο 8 «νόμων», το οποίο μπορεί να συνοψισθεί στην διατύπωση ότι η εξέλιξη του λογισμικού είναι ένα σύστημα ανάδρασης, όπου η επέκταση του συστήματος λόγω των απαιτήσεων των χρηστών συχνά αντισταθμίζεται από την ανάγκη «σύμπτυξης» και εσωτερικής αναδιάρθρωσης, ώστε το σύστημα να είναι συντηρήσιμο και κατανοητό.

Όλες οι περιοχές της μηχανικής (engineering), από την κατασκευή κτιρίων, τη ναυπηγική, ως και την τεχνολογία λογισμικού, χρησιμοποιούν τεχνικές αφαίρεσης και οπτικής αναπαράστασης των design artifacts τους (γνωστά και ως blueprints). Έχοντας ήδη κατακτήσει την αφαίρεση ενός οικοσυστήματος και την αναπαράστασή του με ένα γράφημα, το Γράφημα Αρχιτεκτονικής, το πρόβλημα της αφαίρεσης είχε λυθεί. Όχι όμως και το πρόβλημα της οπτικής αναπαράστασης. Ένα καίριο ερώτημα, λοιπόν, που μας απασχόλησε στα πλαίσια της υποδράσης ΥΔ5.3 ήταν: «πώς μπορούμε να αναπαραστήσουμε οπτικά ένα blueprint, ένα οπτικό χάρτη ενός οικοσυστήματος?»

1.2 Κίνητρα της έρευνας και κεντρική ιδέα

Στο πλαίσιο της Υποδράσης 5.3 σχεδιάσαμε και υλοποιήσαμε μια πληθώρα από τεχνικές που βοηθούν το σχεδιαστή και διαχειριστή ενός οικοσυστήματος να το σχεδιάσει και να το ρυθμίσει με βάση τη δυναμική του συμπεριφορά και την εξέλιξή του.

Κατ' αρχήν, σχεδιάσαμε και υλοποιήσαμε αλγοριθμικές μεθόδους που επιτρέπουν στους διαχειριστές και τους χρήστες ενός οικοσυστήματος δεδομένων να ποσοτικοποιούν και να βελτιστοποιούν τη σχεδίαση εφαρμογών του οικοσυστήματος, οι οποίες επεξεργάζονται και τροποποιούν μέρος από τα

συνολικά δεδομένα του, ώστε να τα παράσχουν στον τελικό χρήστη στην επιθυμητή τους μορφή. Για το σκοπό αυτό, η κεντρική ιδέα είναι να αντιμετωπισθεί η επεξεργασία και τροποποίηση των δεδομένων της εφαρμογής ως πολίτης 1ης κατηγορίας και να μοντελοποιηθεί η διαδικασία μετατροπής των αρχικών δεδομένων του οικοσυστήματος σε χρήσιμη πληροφορία για τον χρήστη ως ροή εργασίας με επίκεντρο τα δεδομένα (data-centric workflow). Μία τέτοια ροή εργασίας μπορεί να αναπαρασταθεί ως ένα (κατευθυνόμενο) ακυκλικό γράφημα, όπου κάθε κορυφή είναι ένα στάδιο επεξεργασίας και τροποποίησης των δεδομένων του οικοσυστήματος.

Για να μπορέσουμε να αποτιμήσουμε με ποσοτικό τρόπο την ευαισθησία μιας ροής εργασίας στις αλλαγές, χρησιμοποιήσαμε τα Γραφήματα Αρχιτεκτονικής, που με ενιαίο αλλά και αναλυτικό τρόπο αναπαριστούν δεδομένα και λογισμικό για να στηρίξουμε τη διερεύνησή μας. Εκμεταλλευόμενοι το γράφημα, χρησιμοποιήσαμε τις ροές εργασίας μιας πραγματικής περίπτωσης για να μελετήσουμε τις μετρικές του γραφήματος και να αποφανθούμε για το αν μπορούμε να τις χρησιμοποιήσουμε σαν μηχανισμό πρόβλεψης για την ευαισθησία της ροής σε αλλαγές.

Στα πλαίσια της ανάγκης να κατανοήσουμε «νόμους» που διέπουν την εξέλιξη των οικοσυστημάτων δεδομένων, ξεκινήσαμε με το αναγκαίο ερώτημα: «πώς εξελίσσονται οι βάσεις δεδομένων?» Ο πρώτος στόχος της μελέτης μας είναι να εκτιμήσει τη δυνατότητα εφαρμογής των νόμων του Lehman για την εξέλιξη λογισμικού, σε βάσεις δεδομένων λογισμικού ανοιχτού κώδικα. Ωστόσο, η ισχύς των νόμων αυτών όσον αφορά τις βάσεις δεδομένων δεν έχει μελετηθεί μέχρι σήμερα. Το ερώτημα που θέσαμε ήταν «Ισχύουν οι νόμοι του Lehman (και ποιοι) για την εξέλιξη του σχήματος μιας βάσεων δεδομένων?»

Τέλος, αλλά όχι λιγότερο σημαντικά, αντιμετωπίσαμε και το πρόβλημα της αναπαράστασης ενός οπτικού χάρτη του οικοσυστήματος. Συνεχίσαμε να στηριζόμαστε σε γραφοθεωρητική προσέγγιση και εκμεταλλευόμενοι την διαθεσιμότητα των Γραφημάτων Αρχιτεκτονικής των οικοσυστημάτων,

ερευνήσαμε εναλλακτικούς αλγορίθμους για πώς μπορούμε να κατασκευάσουμε ένα οπτικό χάρτη γι' αυτά.

2 Διαχείριση ροών επεξεργασίας των δεδομένων

Η εργασία μας αναφορικά με τη βελτιστοποίηση εφαρμογών, θεωρώντας τες ως συγκεκριμένες μορφές ροών επεξεργασίας, επικεντρώθηκε σε δύο σημεία. Το πρώτο σημείο αφορά την αναδιάταξη των επί μέρους διεργασιών, ώστε να προκύψει βελτιστοποιημένη εκτέλεση, ενώ το δεύτερο σημείο αφορά την επιλογή της πλατφόρμας εκτέλεσης κάθε διεργασίας. Τα σημεία αυτά περιγράφονται στις δύο επόμενες υποενότητες, ενώ αναλυτική περιγραφή βρίσκεται στις δημοσιεύσεις [KG13], [KG14] και [KGT15].

2.1 Δηλωτική περιγραφή και αναδιάταξη των διεργασιών

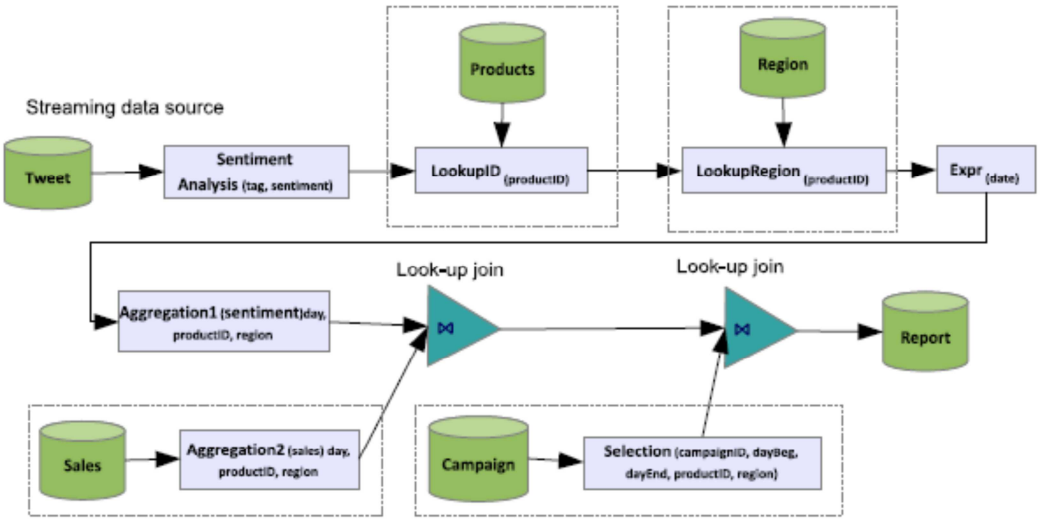
Οι μεθοδολογίες που έχουν προταθεί μέχρι σήμερα σχετικά με τη διαχείριση των ροών δεδομένων εξαρτώνται άμεσα από το χρήστη και σχεδιαστή αυτών των ροών, ο οποίος είναι υπεύθυνος για τον καθορισμό της δομής αυτών των ροών. Όμως, ο σχεδιαστής των ροών δεν κατέχει πάντοτε την κατάλληλη τεχνογνωσία και εμπειρία, ώστε να σχεδιάσει ένα βέλτιστο πλάνο εκτέλεσης για κάθε ροή δεδομένων.

Ο στόχος στην εργασία [KG13] είναι ο σχεδιαστής μίας ροή δεδομένων να μην απαιτείται να σχεδιάσει την ακριβή σειρά εκτέλεσης των δραστηριοτήτων μίας ροής, αλλά να ορίζει υψηλότερου επιπέδου πληροφορία μέσα από περιορισμούς προτεραιότητας. Με σκοπό να το επιτύχουμε αυτό, προτείνουμε έναν δηλωτικό τρόπο για να προσδιορίζουμε τις ροές δεδομένων, κάτι πολύ συνηθισμένο στον ευρύτερο χώρο των βάσεων δεδομένων. Ο δηλωτικός τρόπος που προτείνουμε βασίζεται σε μία θεώρηση των διεργασιών ως εικονικές σχέσεις και σε σύμβολα τα οποία υπάρχουν στο σχήμα εξόδου κάθε δραστηριότητας της ροής. Συγκεκριμένα, προτείναμε τα δεσμευτικάπρότυπα (binding patterns), όπου μας βοηθούν να διαχωρίσουμε τα χαρακτηριστικά που απαιτούνται ως είσοδο (δεσμευμένα-bound) και αυτά που παράγονται στην έξοδο της εκτέλεσης κάθε δραστηριότητας μίας ροής (μη δεσμευμένα/ελεύθερα-free). Για παράδειγμα, η σχέση Task1($X^{b,Task2}$, Y^{b}, Z^{f}) αντιστοιχεί σε ένα στάδιο επεξεργασίας Task1, το

οποίο απαιτεί την είσοδο των X και Y τιμών και την εκτέλεση της δραστηριότητας Task2 νωρίτερα. Με τη μεθοδολογία αυτή, λοιπόν, μπορούμε να εκφράσουμε τις ροές δεδομένων ως ερωτήματα με συμβολισμούς και να εφαρμόσουμε σε αυτά αλγορίθμους βελτιστοποίησης ερωτημάτων, οι οποίοι λαμβάνουν υπόψη τυχόν περιορισμούς προτεραιότητας μεταξύ των δραστηριοτήτων ροής.

Η εφαρμογή αυτής της προσέγγισης αποδείχθηκε ότι είναι επαρκής ώστε να καλύψει ένα ευρύ φάσμα μετασχηματισμών δεδομένων, οι οποίοι μέχρι τώρα δεν μπορούσαν να υποστηριχθούν από τους παραδοσιακούς σχεσιακούς τελεστές. Συγκεκριμένα η δηλωτική προσέγγιση μπορεί να υποστηρίξει όχι μόνο απλές ροές δεδομένων με μία είσοδο και έξοδο, αλλά και πιο σύνθετες ροές με περισσότερες εισόδους και πιο πολύπλοκα στάδια.

Ένα ενδεικτικό παράδειγμα παρουσιάζεται παρακάτω, όπου η ροή δεδομένων μετασχηματίζει tweets σε χρήσιμη πληροφορία.



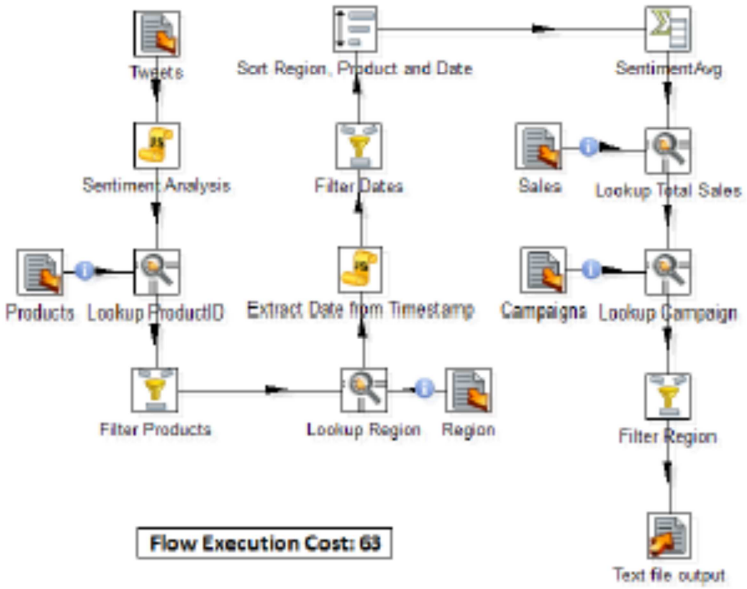
Η συγκεκριμένη ροή δεδομένων, που είναι ουσιαστικά ένα ακυκλικό κατευθυνόμενο γράφημα, αντιστοιχεί στο ακόλουθο ερώτημα, το οποίο είναι δηλωτικό:

```
Select *
From   Tweet(tag^f, timestamp^f) ⋈
       Sentiment_Anal(tag^b, sentiment^f) ⋈
       LookupID(tag^b, productID^f) ⋈
       LookupRegion(tag^b, region^f) ⋈
       Expr(tag^b, timestamp^b, date^f) ⋈
       Aggregation1(tag^b, sentiment^b, productID^b, region^b, date^b,
                    productIdGroup^f, dateGroup^f, regionGroup^f, AvgAggSentiment^f) ⋈
       Aggregation2(productID^f, region^f, date^f, totalAggSales^f) ⋈
       Selection(productID^f, campaignID^f, dayBeg^f, dayEnd^f, region^b)
Where  Tweet.tag = Sentiment_Anal.tag and
       Tweet.tag = LookupID.tag and
       Tweet.tag = LookupRegion.tag and
       Tweet.tag = Expr.tag and
       Tweet.tag = Aggregation1.tag and
       Aggregation1.productID = Aggregation2.productID and
       Aggregation1.region = Aggregation2.region and
       Aggregation1.date = Aggregation2.date and
       Aggregation1.productID = Selection.productID and
       Aggregation1.region = Selection.region
```

Η δηλωτική διατύπωση των ροών αφήνει περιθώρια στο σύστημα να επιλέξει το ίδιο τον ακριβή τρόπο εκτέλεσης. Όπως αναφέραμε και προηγουμένως, οι ροές δεδομένων δε σχεδιάζονται πάντοτε από ειδικούς και επομένως, δε διαθέτουν πάντοτε την τεχνογνωσία για τον καθορισμό ενός βέλτιστου πλάνου εκτέλεσης. Ακόμη, και εάν οι σχεδιαστές αυτών των ροών είναι μόνο ειδικοί, θα ήταν πολύ δύσκολο και γι' αυτούς να παράγουν πάντα το βέλτιστο πλάνο εκτέλεσης, καθώς οι σύγχρονες ροές δεδομένων είναι αρκετά πολύπλοκες και μεγάλες σε αριθμό δραστηριοτήτων. Επιπρόσθετα, οι σύγχρονες ροές δεδομένων εκτελούνται σε συνεχώς μεταβαλλόμενα περιβάλλοντα, όπου τα δεδομένα εισόδου μίας ροής δεδομένων αλλάζουν διαρκώς. Ακόμη και αν καθοριστεί το βέλτιστο πλάνο για ένα συγκεκριμένο σύνολο δεδομένων, αυτό δε μας εξασφαλίζει ότι το πλάνο θα συνεχίσει να είναι βέλτιστο για ένα άλλο σύνολο δεδομένων ή ακόμη και για το ίδιο σύνολο δεδομένων με χαρακτηριστικά τα οποία ανανεώνονται. Για όλους τους λόγους που αναφέραμε παραπάνω, η υιοθέτηση αυτοματοποιημένων τεχνικών βελτιστοποίησης είναι απαραίτητη, κάτι που αποτελεί το βασικό αντικείμενο της εργασίας [KG14].

Σε αυτήν την εργασία, επικεντρωνόμαστε σε εξειδικευμένες προτάσεις βελτιστοποίησης που αφορούν τη βελτιστοποίηση του λογικού πλάνου εκτέλεσης μέσα από μεθοδολογίες αναδιάταξης δραστηριοτήτων. Συγκεκριμένα, στην εργασία αυτήν αποδεικνύεται ότι ακόμη και για την πιο απλή περίπτωση μίας ροής δεδομένων με μία είσοδο και μία έξοδο το λογικό πλάνο εκτέλεσης

μπορεί να απέχει αρκετά από το βέλτιστο. Ως ένα παράδειγμα, χρησιμοποιούμε την εκτέλεση του προηγούμενου παραδείγματος στο εργαλείο Pentaho Data Integration (βλ. επόμενο σχήμα) και παρατηρούμε ότι ενώ χωρίς να επέμβουμε το κόστος εκτέλεσης είναι 63 μονάδες χρόνου, ενώ ο βέλτιστος χρόνος εκτέλεσης μπορεί να κατέβει στο συγκεκριμένο παράδειγμα στις 18.3 μονάδες.



Σε αυτήν την εργασία, δείχνουμε ότι η υπάρχουσα βιβλιογραφία που αφορά τη βελτιστοποίηση των ροών δεδομένων σε λογικό επίπεδο και πιο συγκεκριμένα, στον βελτιωμένο καθορισμό της ακολουθίας των εργασιών-σταδίων μίας ροής δεδομένων, χρειάζεται να επεκταθεί από εξειδικευμένους αλγορίθμους (που ονομάσαμε RO-I&II) οι οποίοι θα κλιμακώνονται για ροές δεδομένων με μεγάλο αριθμό δραστηριοτήτων και θα έχουν σημαντική βελτίωση στην απόδοση. Μετά από αξιολόγηση των αλγορίθμων, καταλήξαμε ότι οι προτεινόμενες λύσεις βελτιστοποίησης στη χειρότερη περίπτωση έχουν 16% καλύτερη βελτίωση απόδοσης από τον υπάρχοντα καλύτερο αλγόριθμο της βιβλιογραφίας, ενώ στην καλύτερη μέχρι και 41% καλύτερη βελτίωση απόδοσης.

2.2 Επιλογή της πλατφόρμας εκτέλεσης

Στην πάροδο του χρόνου, οι ροές δεδομένων που υιοθετούνται, ολοένα και περισσότερο διαχειρίζονται δεδομένα μεγάλου όγκου, τόσο σε επιχειρηματικά σενάρια, όσο και σε επιστημονικά. Ένα επιπλέον κοινό στοιχείο είναι ότι οι

επιμέρους διεργασίες μπορούν να εκτελεστούν κατανεμημένα και σε διαφορετικές μηχανές (ή αλλιώς πλατφόρμες) εκτέλεσης.

Στην εργασία [KGT15] επικεντρωνόμαστε στη βελτιστοποίηση του φυσικού πλάνου εκτέλεσης μίας ροής δεδομένων και συγκεκριμένα, στοχεύουμε στη λύση του προβλήματος της κατανομής των δραστηριοτήτων των ροών σε ετερογενείς και ανεξάρτητες μηχανές εκτέλεσης με σκοπό την εκτέλεση του πλάνου στο μικρότερο δυνατό χρόνο. Αποδεικνύεται ότι ακόμη και στη γενική περίπτωση το πρόβλημα χαρακτηρίζεται ως NP-hard στη γενική περίπτωση. Στην ειδική περίπτωση για ροές δεδομένων που χαρακτηρίζονται ως γραμμικές, προτείνουμε μία ψευδοπολυωνυμική λύση, που βρίσκει όμως το βέλτιστο κόστος. Επίσης, προτείνουμε πρακτικούς αλγορίθμους, οι οποίοι μπορούν να επιτύχουν την κατανομή των δραστηριοτήτων μέσα σε λίγα δευτερόλεπτα.

Η βελτιστοποίηση της κατανομής των δραστηριοτήτων μίας ροής δεδομένων βασίζεται στην επιλογή της καλύτερης μηχανής εκτέλεσης για κάθε δραστηριότητα μέσα από ένα σύνολο υποψήφιων μηχανών εκτέλεσης. Επίσης, η απόδοση των πλάνων εκτέλεσης υπολογίζεται από το άθροισμα των χρόνων εκτέλεσης όλων των δραστηριοτήτων μίας ροής δεδομένων. Μέχρι και σήμερα, για την προσέγγιση αυτού του προβλήματος είχαν προταθεί μόνο απλά ευρετικά και αλγόριθμοι που δεν κλιμακώνονται. Επίσης, κάποιες από τις ιδιαίτερες προκλήσεις αυτού του προβλήματος ήταν ο μεγάλος αριθμός των δραστηριοτήτων και των υποψήφιων μηχανών εκτέλεσης, η ετερογένεια και διαθεσιμότητα των μηχανών και το επιπρόσθετο κόστος που επιφέρει η μεταφορά δεδομένων από μηχανή εκτέλεσης σε μηχανή εκτέλεσης.

Με σκοπό να επιλύσουμε το πρόβλημα αυτό προτείναμε διάφορους αλγορίθμους, με σημαντικά διαφορετικό σκεπτικό, π.χ., branch-and-bound, random walk, set-cover, οι οποίοι οποιαδήποτε στιγμή και αν τους διακόψουμε κατά την εκτέλεση τους, μπορούν να επιστρέψουν ένα πλάνο που τηρεί του κανόνες προτεραιότητας (αλγόριθμοι anytime) και αποδίδουν αποδοτικότερα από τις απλοϊκές λύσεις που είχαν προταθεί, ακόμη και όταν εφαρμόζονται σε ροές δεδομένων με μεγάλο αριθμό δραστηριοτήτων μέσα σε λίγα δευτερόλεπτα.

Οι προτεινόμενες λύσεις αξιολογήθηκαν τόσο σε συνθετικές, όσο και σε πραγματικές επιστημονικές ροές δεδομένων. Αξίζει να σημειωθεί ότι εξαιτίας της δομής των πραγματικών σεναρίων, τα οποία αν και πολύπλοκα έχουν πολύ μεγάλα τμήματα που είναι γραμμικά, ο χρόνος εκτέλεσης των ροών που έχουν δομή που συναντάται σε πραγματικά σενάρια, μπορεί να μειωθεί μέχρι και τρεις φορές. Σημαντικές βελτιώσεις παρατηρήθηκαν και σε συνθετικά σενάρια, κάτι που περαιτέρω ενισχύει την πεποίθησή μας ότι οι συγκεκριμένες προτάσεις πραγματοποιούν μία σημαντική συνεισφορά στο πεδίο της βελτιστοποίησης ροών εργασίας.

3 Μετρικές της εξέλιξης οικοσυστημάτων

Έχοντας μοντελοποιήσει τις ροές δεδομένων και έχοντας προτείνει μεθόδους για την σχεδίαση και τη βελτιστοποίηση τους, μπορούμε επίσης να προχωρήσουμε και στη χρήση μετρικών οι οποίες βοηθούν το σχεδιαστή να αποτιμήσει την ευαισθησία μιας σχεδίασης στις αλλαγές.

Στα πλαίσια της Υποδράσης 5.3 ασχοληθήκαμε με τη μελέτη ροών δεδομένων της κατηγορίας Extract-Transform-Load (ETL) σε μία μελέτη περίπτωσης. Μοντελοποιήσαμε τις ροές εργασίας ενός οργανισμού του Ελλ. Δημοσίου με τα Γραφήματα Αρχιτεκτονικής που έχουμε εισάγει ήδη από την Υποδράση 5.2 και μελετήσαμε την εξέλιξή τους. Θυμίζουμε ότι σε ένα Γράφημα Αρχιτεκτονικής, ανάγουμε όλα τα στοιχεία και τις εξαρτήσεις του οικοσυστήματος (πίνακες, ερωτήσεις, εμπλεκόμενα πεδία, σχέσεις τροφοδοσίας με δεδομένα) σε ένα ενιαίο γράφημα αναπαράστασης. Οι δραστηριότητες μιας ροής δεδομένων, η οποία ανακτά, μετασχηματίζει και φορτώνει δεδομένα σε μια αποθήκη δεδομένων, λόγω της έντονα στραμμένης προς τα δεδομένα φύσεώς της, μπορεί να αναπαρασταθεί εύκολα με ένα γράφημα αρχιτεκτονικής.

Στη συνέχεια, προχωρήσαμε στον ορισμό γραφοθεωρητικών μετρικών ποιότητας για την αποτίμηση της εξέλιξης πολυδιάστατων οικοσυστημάτων δεδομένων (π.χ., ο έξω/έσω βαθμός ενός κόμβου, ο μεταβατικός βαθμός που λαμβάνει υπόψη τα μονοπάτια τροφοδοσίας, κλπ). Έπειτα, αξιολογήσαμε τις μετρικές αυτές στο προαναφερθέν πραγματικά συστήματα αποθηκών

δεδομένων. Συγκεκριμένα, διαπιστώσαμε ότι (α) ο αριθμός των πεδίων που εμπλέκονται σε μια δραστηριότητα και (β) η πολυπλοκότητά της μπορούν να μας «προβλέψουν» πόσο ευαίσθητη μπορεί να είναι σε πιθανές αλλαγές. Ο έξω βαθμός ενός κόμβου, από πλευράς γραφήματος, είναι η μετρική με την καλύτερη ικανότητα πρόβλεψης.

Με βάση τα ευρήματά μας, έχουμε προτείνει ένα σύνολο από «καλές πρακτικές» για την αποτίμηση και σχεδίαση οικοσυστημάτων δεδομένων με στόχο την διαχείριση της εξέλιξής τους στο χρόνο. Οι συστάσεις μας συνοψίζονται εν τάχει (α) στην υιοθέτηση «στενών» σχημάτων, όπου αυτό είναι εφικτό, (β) στην απλοποίηση της εσωτερικής δομής των διαδικασιών και (γ) στην αναδιάταξη των δραστηριοτήτων ώστε να μειωθεί ο αριθμός των εμπλεκόμενων στη ροή πεδίων, όσο το δυνατόν πιο νωρίς.

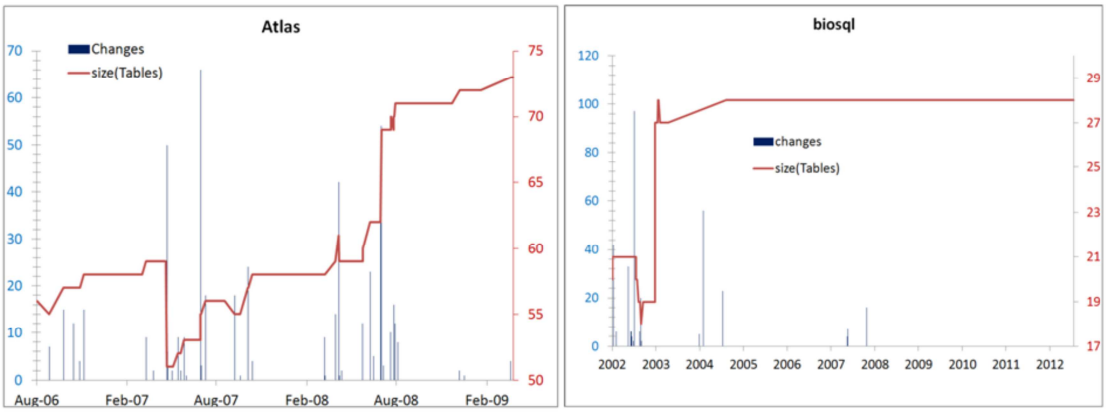
Τα αποτελέσματά μας δημοσιεύθηκαν στο περιοδικό Journal of Data Semantics, Springer [PVSV12].

4 Μελέτη των θεμελιωδών νόμων της εξέλιξης

Όπως προαναφέρθηκε, ο πρώτος στόχος της μελέτης μας είναι να εκτιμήσει τη δυνατότητα εφαρμογής των νόμων του Lehman για την εξέλιξη λογισμικού, σε βάσεις δεδομένων λογισμικού ανοιχτού κώδικα. Αυτό το σύνολο των οκτώ νόμων της εξέλιξης είναι ένα καλά εδραιωμένο σύνολο παρατηρήσεων (που έχει ωριμάσει κατά τη διάρκεια των τελευταίων σαράντα ετών) σχετικά με το πώς τα συστήματα λογισμικού εξελίσσονται.

Για το σκοπό αυτό, αξιοποιήσαμε την ύπαρξη συστημάτων ανοικτού λογισμικού που δημοσιεύουν σε δημόσια αποθετήρια τον κώδικά τους. Για 8 από αυτά ανακτήσαμε την ιστορία της εξέλιξης του σχήματος της βάσης δεδομένων που εμπεριέχουν και πραγματοποιήσαμε μια εμπεριστατωμένη μελέτη σχετικά με την εξέλιξη των διαφόρων ιδιοτήτων του σχήματος της βάσης δεδομένων (όπως μέγεθος, ανάπτυξη, και το ύψος των αλλαγών ανά έκδοση, τόσο όσον αφορά τις σχέσεις όσο και τα γνωρίσματα αυτών) και αναφέρουμε τα αποτελέσματα σχετικά με την εγκυρότητα του κάθε νόμου με βάση τις εν λόγω παρατηρήσεις.

LAW & RESEARCH QUESTIONS	MEASURES	FINDINGS & COMMENTS	DATASETS
I Continuing change The schema is continually adapted	Heartbeat	The schema is adapted in the long run, albeit not continually, but in focused periods of modification	All datasets abide by the law, with two exceptions: - BioSQL, with a "sleep" of some years - phpBB with a turbulence period
III Self-regulation - Schema size expands with recurring patterns of smooth expansion, interrupted by abrupt change - Existence of shrinking versions (negative feedback)	Size	- Patterns of change include (a) expansion, (b) shrinking and (c) stability; differently from the expression of the law - Perfective maintenance is evident	All datasets without exceptions
- Change normally distributed around an average value	Growth	- Growth is small, typically close to zero, following a Zipfian model - Oscillations of large size do exist	
VIII Feedback System We can estimate the schema size via regressive formula	Regression analysis	Size estimation can be achieved; out of the different alternatives for effort estimation, the ones with small time window work better	All datasets without exceptions
VI Continuing growth The schema size is increasing in the long run	Size	Size increases in the long run, indeed, albeit not continually, but in focused periods of modification	All datasets without exceptions
V Conservation of familiarity - The average growth between versions is slowly declining	Heartbeat, Size	- The linear interpolation of growth typically drops or stays stable; importantly, the frequency of change declines	All datasets except for Atlas and BioSQL (with some extra activity in the end of their lifetimes)
- What happens after excessive changes?	Growth	- Spikes are followed by all possible combinations (calmness, other spikes, large oscillations around zero)	All datasets exhibit various patterns
IV Invariant work rate Avg. work-rate is constant within phases of smooth growth, connected with bursts of effort	(approximate) Heartbeat & Size	There is are no phases of constant growth; albeit periods of stability connected via focused periods of modifications	All datasets without exceptions
II Increasing complexity Complexity increases over time	(approximate) Schema Complexity	Complexity drops	All datasets except for phpBB (having a turbulence period in the end)
VII Declining Quality Quality declines over time	Conjecture by logical induction	Impossible to hold as valid, as complexity (albeit approximated) seems to drop	-



Συνδυαστική αναπαράσταση του μεγέθους του σχήματος (κόκκινη γραμμή) με το heartbeat των αλλαγών στην πάροδο του χρόνου, για τα οικοσυστήματα Atlas και Biosql

Βασικές παρατηρήσεις:

- Η θεμελιώδης ουσία των νόμων της εξέλιξης φαίνεται να ισχύει. Θυμίζουμε ότι Ο βασικός νόμος του Lehman είναι ότι η εξέλιξη του λογισμικού είναι ένα σύστημα ανάδρασης, όπου η επέκταση του συστήματος λόγω των απαιτήσεων των χρηστών συχνά αντισταθμίζεται από την ανάγκη «σύμπτυξης» και εσωτερικής αναδιάρθρωσης, ώστε το σύστημα να είναι συντηρήσιμο και κατανοητό. Διαφαίνεται ότι στην περίπτωση της εξέλιξης των σχημάτων βάσεων δεδομένων, το σύστημα ανάδρασης ισχύει, καθώς το μέγεθος μίας συγκεκριμένης έκδοσης μπορεί να εκτιμηθεί μέσω μίας regression φόρμουλας που στηρίζεται κυρίως στην πολύ πρόσφατη ιστορία.
- Τα σχήματα γενικώς επεκτείνονται και αυξάνονται με την πάροδο του χρόνου. Εν αντιθέσει όμως με τους νόμου του παραδοσιακού λογισμικού, το σχήμα μιας βάσης δεδομένων εξελίσσεται σε ριπές, σε ομαδοποιημένες περιόδους εξελικτικής δραστηριότητας και όχι σε μια συνεχή διαδικασία.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [SkVZ14] που παρουσιάστηκε στο 26th International Conference on Advanced Information Systems Engineering (CAiSE 2014). Να σημειωθεί ότι τα συνέδρια ER και CAiSE είναι τα κορυφαία συνέδρια στο χώρο της ερευνητικής κοινότητας του Conceptual Modeling και του Information Systems Engineering.

Λόγω της υψηλής κατάταξης του άρθρου στην αξιολόγηση της επιτροπής προγράμματος του συνεδρίου, το άρθρο προσεκλήθη στο περιοδικό Information Systems ως ένα από τα κορυφαία άρθρα του συνεδρίου. Η εκτεταμένη εκδοχή του άρθρου διήλθε την διαδικασία κρίσης κανονικά και έγινε αποδεκτή [SkVZ15]. Το περιοδικό Information Systems είναι ένα από τα κορυφαία περιοδικά του χώρου της διαχείρισης δεδομένων και πληροφοριακών συστημάτων.

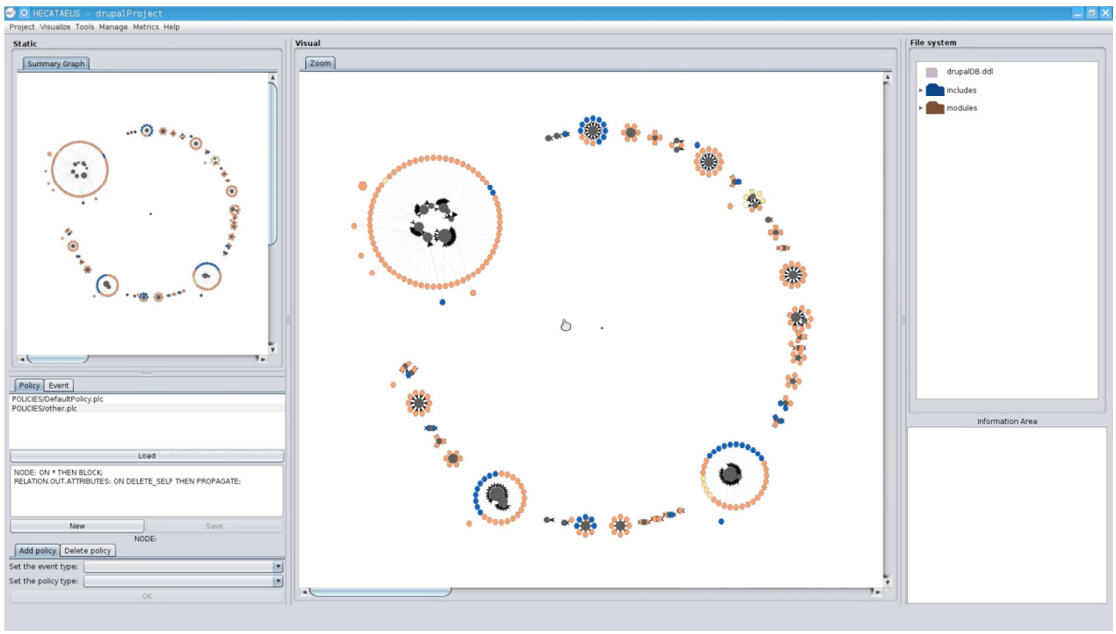
5 Οπτικοποίηση της σχεδίασης

Στα πλαίσια της forward σχεδίασης ενός οικοσυστήματος, αλλά και στο πλαίσιο του reverse engineering του, χρειαζόμαστε *οπτικούς χάρτες* που να μας καταδεικνύουν τα συστατικά και τις αλληλεξαρτήσεις εντός του. Οι οπτικοί χάρτες έχουν ποικίλες χρήσεις, καθώς επιτρέπουν αφενός να κατανοήσουμε ένα οικοσύστημα, να το τεκμηριώσουμε και να εντοπίσουμε πρότυπα δόμησης εντός του, και αφετέρου, να εκτιμήσουμε με μια ματιά την επίπτωση των αλλαγών σε ένα οικοσύστημα εντός του.

Χτίζοντας λοιπόν πάνω στα Γραφήματα Αρχιτεκτονικής των οικοσυστημάτων, παρέχουμε και μια αρχική προσέγγιση στο πώς μπορούμε να κατασκευάσουμε ένα οπτικό χάρτη γι' αυτά. Για την παροχή αυτού του οπτικού χάρτη, αναπαριστούμε την εσωτερική δομή των οικοσυστημάτων ως γράφημα που στηρίζεται σε πίνακες και ερωτήσεις ως βασικούς κόμβους οπτικής αναπαράστασης. Μέσω ενός αλγορίθμου ιεραρχικής συναθροιστικής ομαδοποίησης (hierarchical agglomerative clustering) οι κόμβοι του συστήματος επιμερίζονται σε clusters οι οποίοι και διατάσσονται στον καμβά. Στο παρακάτω σχήμα φαίνεται ένα τέτοιος χάρτης για το σύστημα Drupal, καθώς και λεπτομέρειες που το επεξηγούν.

Έχουμε αναπτύξει με τρεις κυκλικές μεθόδους διάταξης των clusters για την απεικόνιση του χάρτη (εδώ δείχνουμε μόνο τη μία που στηρίζεται σε ένα εξωτερικό κύκλο). Έχουμε επιλύσει επιπλέον και διάφορα τεχνικά προβλήματα

σε σχέση με την διάταξη των κόμβων ώστε να μην υπάρχουν επικαλύψεις και να μειωθεί ο οπτικός θόρυβος,



Χάρτης του οικοσυστήματος του Drupal. Οι κόμβοι οργανώνονται σε clusters, οι οποίοι (α) τοποθετούνται κυκλικά στον καμβά φτιάχνοντας ένα (εξωτερικό) μεγα-κύκλο, (β) με το εσωτερικό τους να οργανώνεται επίσης σε κύκλο (οι μικρότερες κυκλικές διατάξεις του σχήματος). Οι μικροί κόμβοι εσωτερικά στους clusters (α) αναπαριστούν SQL ερωτήσεις ενσωματωμένες στην php (χρωματιστοί κόμβοι), καθώς και (β) πίνακες της βάσης δεδομένων του Drupal (γκρι κόμβοι).

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [KoMP14] που παρουσιάστηκε στο 33rd International Conference on Conceptual Modeling (ER 2014). Να σημειωθεί ότι τα συνέδρια ER και CAiSE είναι τα κορυφαία συνέδρια στο χώρο της ερευνητικής κοινότητας του Conceptual Modeling και του Information Systems Engineering.

Επίσης, τα αποτελέσματα παρουσιάσθηκαν με ομιλία στο 13ο Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων, το 2015 [KoMB15].

6 Ανακεφαλαίωση

Το παρόν παραδοτέο Π5.3 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ5.3 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ5.3 ήταν να παρέχει μεθόδους που να μας επιτρέπουν (α) να αποτιμήσουμε την ποιότητα της σχεδίασης ενός

οικοσυστήματος δεδομένων με το βλέμμα στην εξέλιξή του και (β) να μας επιτρέπουν να συστήσουμε στο χρήστη / διαχειριστή ενός οικοσυστήματος πρότυπα και τεχνικές κατάλληλα ώστε να ελαχιστοποιήσουν το κόστος ανασυγκρότησης του οικοσυστήματος για ένα πιθανό σύνολο εξελικτικών δράσεων στο μέλλον.

Στα πλαίσια της έρευνάς μας, λοιπόν, επιτύχαμε να ανταποκριθούμε στο στόχο της υποδράσης με τους ακόλουθους τρόπους:

1. Συμπληρώσαμε τη μοντελοποίηση του οικοσυστήματος, επεκτείνοντάς την και σε ροές εργασίας. Η μοντελοποίηση αυτή μας επέτρεψε να δώσουμε στο σχεδιαστή του οικοσυστήματος μεθόδους που βελτιστοποιούν την εκτέλεση μιας ροής με φυσική αναδιάταξη των εσωτερικών συστατικών της και με την έξυπνη τοποθέτηση εργασιών στο σωστό σημείο, όταν η ροή εκτελείται με παράλληλο τρόπο σε ένα ετερογενές δίκτυο μηχανών.
2. Εισάγαμε τρόπους μέτρησης γραφοθεωρητικών μετρικών μιας ροής σαν μηχανισμό πρόβλεψης για την ευαισθησία της σε αλλαγές, με βάση την μελέτη μιας πραγματικής περίπτωσης.
3. Μελετήσαμε θεμελιώδεις νόμους της εξέλιξης για να μπορούμε να δούμε τους μηχανισμούς με τους οποίους το σχήμα μιας βάσης εξελίσσεται και υιοθετήσαμε ένα δημοφιλές μοντέλο εξέλιξης από το χώρο της Τεχνολογίας Λογισμικού.
4. Παρέχουμε προχωρημένες τεχνικές οπτικοποίησης στο σχεδιαστή και στον χρήστη ενός οικοσυστήματος διατάσσοντας στο δισδιάστατο χώρο το Γράφημα Αρχιτεκτονικής του.

Δημοσιεύσεις

[KG13] Georgia Kougka, Anastasios Gounaris. Declarative expression and optimization of data-intensive flows. In Proceedings 15th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2013), Prague, Czech Republic, August 26-29, 2013.

[KG14] Georgia Kougka, Anastasios Gounaris. Optimization of Data-

intensive Flows: Is it Needed? Is it Solved?. In Proceedings 17th International Workshop on Data Warehousing and OLAP (DOLAP 2014), Shanghai, China, November 3-7, 2014.

[KGT15] Georgia Kougka, Anastasios Gounaris, Kostas Tsichlas. Practical algorithms for execution engine selection in data flows. Future Generation Computer Systems 45 (2015).

[PVSV12] George Papastefanatos, Panos Vassiliadis, Alkis Simitsis, Yannis Vassiliou. Metrics for the Prediction of Evolution Impact in ETL Ecosystems: A Case Study. Journal on Data Semantics, volume 1, number 2, pp. 75-97, 2012, Springer.

[SkVZ14] Ioannis Skoulis, Panos Vassiliadis, Apostolos Zarras. Open-Source Databases: Within, Outside, or Beyond Lehman's Laws of Software Evolution?. In Proceedings 26th International Conference on Advanced Information Systems Engineering (CAiSE 2014), Thessaloniki, Greece, June 16-20, 2014.

[SkVZ15] Ioannis Skoulis, Panos Vassiliadis, Apostolos V. Zarras. Growing up with stability: How open-source relational databases evolve. Information Systems, Volume 53, October–November 2015, Pages 363 - 385. doi:10.1016/j.is.2015.03.009

[KoMP14] Efthymia Kontogiannopoulou, Petros Manousis, Panos Vassiliadis. Visual Maps for Data-Intensive Ecosystems. In Proceedings 33rd International Conference on Conceptual Modeling (ER 2014), Atlanta, U.S.A., 27-29 October, 2014.

[KoMB15] Ευθυμία Κοντογιαννοπούλου, Πέτρος Μανούσης, Παναγιώτης Βασιλειάδης. Οπτικοποίηση Πληροφοριακών Οικοσυστημάτων Παρουσιάσθηκε στο 13ο Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων (ΕΣΔΔ 2015), Αθήνα, 30-31 Ιουλίου, 2015.

Παράρτημα