Μοτίβα Εξέλιξης Πινάκων σε Σχεσιακές Βάσεις Δεδομένων

Η Μεταπτυχιακή Διπλωματική Εργασία

υποβάλλεται στην ορισθείσα
από την Συνέλευση
του Τμήματος Μηχανικών Η/Υ και Πληροφορικής
Εξεταστική Επιτροπή

από την

Θεολογία Καλάκου

ως μέρος των υποχρεώσεων για την απόκτηση του

ΔΙΠΛΩΜΑΤΟΣ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΜΕ ΕΙΔΙΚΕΥΣΗ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΚΑΙ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ

Πανεπιστήμιο Ιωαννίνων

Ιανουάριος 2021

Εξεταστική επιτροπή:

- **Παναγιώτης Βασιλειάδης**, Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων (Επιβλέπων)
- **Νικόλαος Μαμουλής**, Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
- **Ευαγγελία Πιτουρά**, Καθηγήτρια, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων

ΑΦΙΕΡΩΣΗ

Στον Γιώργο

Ευχαριστιες

Ευχαριστώ την Κλειώ και την Κατερίνα για το υπέροχο χιούμορ με το οποίο υποστήριξαν αυτή την προσπάθεια της μαμάς τους.

Ευχαριστώ όλους όσους γνώρισα στο Τμήμα Πληροφορικής (τις κυρίες στη γραμματεία, τους καθηγητές μου, τους συμφοιτητές μου) για την ευγένειά τους.

Ένα μεγάλο ευχαριστώ από καρδιάς στον Π. Βασιλειάδη.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Περιληψη

Θεολογία Καλάκου, Δ.Μ.Σ. στη Μηχανική Δεδομένων και Υπολογιστικών Συστημάτων, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων, Ιανουάριος 2021

Μοτίβα Εξέλιξης Πινάκων σε Σχεσιακές Βάσεις Δεδομένων

Επιβλέπων : Παναγιώτης Βασιλειάδης - Καθηγητής

Το λογισμικό που είναι σχεδιασμένο για να κάνει τη διαχείριση μιας σχεσιακής βάσης δεδομένων εξελίσσεται με την πάροδο του χρόνου. Η εξέλιξη αυτή μπορεί να είναι απόρροια λαθών στην αρχική σχεδίαση ή οι χρήστες να ανακαλύπτουν πώς θα ήθελαν επιπλέον δυνατότητες από το λογισμικό. Συνεπώς, για να συνεχίσει το λογισμικό να είναι λειτουργικό και βιώσιμο, θα πρέπει να προσαρμόζεται στις νέες απαιτήσεις. Οι νέες πληροφορίες που αποθηκεύονται καθώς και η εξέλιξη του λογισμικού για την διαχείριση αυτών των νέων πληροφοριών είναι ζητήματα μεγάλης σημασίας. Σε κάθε αλλαγή τα δεδομένα και ο κώδικας εφαρμογής πρέπει να συγχρονιστούν με το νέο σχήμα. Συνεπώς, η κατανόηση της εξέλιξης του σχήματος είναι πολύ σημαντική ώστε να συνεχίζεται ο κύκλος ζωής της εφαρμογής.

Το βασικό ερώτημα που θέλουμε να απαντήσουμε στην παρούσα εργασία είναι αν η εξέλιξη των βάσεων δεδομένων ακολουθεί κάποια πρότυπα. Τα πρότυπα αυτά είναι (α) το πρότυπο Γάμμα που συνδυάζει το μέγεθος ενός πίνακα κατά τη γέννηση με την διάρκειας ζωής του και (β) το πρότυπο του αντίστροφο Γ που συνδυάζει την διάρκεια ζωής ενός πίνακα με το πλήθος των αλλαγών.

Το πρώτο μέρος της εργασίας (κεφάλαιο 2 και 3) ασχολείται με το ακόλουθο ερώτημα: *Τι έρευνα έχει γίνει πάνω στην εξέλιξη του σχήματος;* Καταγράφονται αποτελέσματα 8 ερευνών και τι απαντήσεις έδωσαν σε 3 βασικές κατηγορίες ερωτημάτων: (α) σχετικά με την μακροσκοπική περιγραφή της εξέλιξης στο χρόνο (πόσο συχνά και σε τι βαθμό αλλάζουν τα σχήματα από έκδοση σε έκδοση), (β) σχετικά με την εσωτερική δομή και ανάλυση των αλλαγών (ποσοτικές μετρήσεις) και (γ) σχετικά με την εξωτερική σχέση των αλλαγών του σχήματος με τον περιβάλλοντα κώδικα. Μελετήθηκαν επίσης τα πρότυπα που εξάγονται κατά τη μελέ-

τη πληροφοριών σχετικών με τις ”γεννήσεις”, του “θανάτους” και τις ενημερώσεις πινάκων και πεδίων, μαζί με την εξέλιξη του μεγέθους του σχήματος.

Το δεύτερο μέρος της εργασίας αποτελείται από τα κεφάλαια 4 και 5. Στο κεφάλαιο 4 τίθεται το ακόλουθο ερώτημα: *ποια είναι η συμπεριφορά όσον αφορά την επιβίωση των wide πινάκων (δηλαδή των πινάκων με πάνω από 10 attributes);* Η κατανόηση του κύκλου ζωής των wide πινάκων βοηθά στην έρευνα του προτύπου Γάμμα καθώς αυτό το πρότυπο βασίζεται στην διάρκεια ζωής των πινάκων. Η έρευνα έγινε σε 35 σύνολα δεδομένων. Τα βασικά ερωτήματα που απαντήθηκαν είναι (α) Ποια είναι η πιθανότητα να ζήσουν οι wide πίνακες; (β) Ποια είναι η πιθανότητα να επιβιώσουν οι wide πίνακες σε σχέση με τους not wide ; και (γ) αν επιβιώνουν οι wide πίνακες με μεγαλύτερη πιθανότητα όταν είναι λίγοι; Στο κεφάλαιο 5 το βασικό ερώτημα αφορά στο να εντοπιστούν τα χαρακτηριστικά των συνόλων δεδομένων στα οποία μπορεί να μελετηθεί το πρότυπο του αντίστροφου Γ. Η μελέτη έγινε σε 195 σύνολα δεδομένων και η βασική έρευνα αφορά του πίνακες με πολλές αλλαγές και πως αυτές οι αλλαγές συνδυάζονται με την διάρκεια ζωής των πινάκων.

Η έρευνα των παραπάνω ερωτημάτων έγινε στο σύνολο των συνόλων δεδομένων αλλά και ανά κατηγορία. Τα σύνολα δεδομένων κατηγοριοποιήθηκαν με βάση τις ουσιαστικές αλλαγές πάνω στο σχήμα (όχι σε αλλαγές που αφορούν σχολιασμό, δημιουργία δεικτών ή οτιδήποτε δεν διαφοροποιεί το σχήμα) και με βάση το πλήθος των *attributes* που αλλάχθηκαν. Οι κατηγορίες που μελετήθηκαν είναι : 0_FROZEN (totalActivity = 0), 1_ALMOST_FROZEN (totalActivity <= 10 updated attributes), 1_FocusedShot_n_FROZEN (totalActivity > 10 updated attributes), 2_MODERATE (totalActivity <= 90 updated attributes), 3_FocusedShot_n_LOW, 5_ACTIVE (totalActivity > 90 updated attributes).

EXTENDED ABSTRACT

Theologia Kalakou, MSc in Computer Science, Department of Computer Science and Engineering, University of Ioannina, Greece, January 2021.

Patterns of table evolution in relational databases

Advisor: Panos Vassiliadis, Professor.

Software designed to operate on top of a relational database evolves over time. This development may take place to fix errors in the original design or the demand of the users to have additional features in the software. Therefore, in order for the software to continue to be functional and viable, it will have to update to the new requirements. The new information stored and the updates of the software for managing this are very important issues. At each change the data and the software must be synchronized with the new format. Therefore, understanding the evolution of the schema is very important in order to continue the life cycle of the application.

In this Thesis, we want to answer the key question if the evolution of databases follows certain patterns. These patterns are (a) the model that combines the size of a table at birth with the duration of its life. (Gamma pattern) and (b) the inverse Gamma pattern that combines the duration of the life of a table with the number of updates.

The first part of the Thesis (chapters 2 and 3), deals with the following question: What research has been done on the evolution of the schema? We record the results of 8 papers and the answers they gave to 3 main categories of questions: (a) about the macroscopic description of evolution over time (how often and to what extent the patterns change from version to version), (b) about the internal structure and analysis of changes (quantitative measurements) and (c) on the external relationship of schema changes to the surrounding code. We studied also to the patterns extracted during the study of information related to "births", "deaths" and table updates along with the evolution of the database schema`.

The second part of the work consists of chapters 4 and 5. In chapter 4, the following question is asked: what is the behavior regarding the survival of wide tables (i.e., tables with more than 10 attributes)? Understanding the life cycle of wide

tables helps in researching the Gamma pattern as this pattern is based on the circle of tables life. The research was conducted on 35 datasets. The questions answered are (a) what is the probability of wide tables survival? (b) What is the probability of wide tables to survive in relation of the probability of the not wide ones? and (c) are wide tables more likely to survive when they are few? In Chapter 5, the main question is to identify the characteristics of the data sets in which the inverse C model can be studied. The study was conducted on 195 data sets and the basic research concerns the tables with many changes and how these changes are combined with the lifespan of this tables.

The research of the above questions was done in the whole of data sets but also separated by category. Datasets were categorized based on substantial changes to the schema (not changes to annotation, indexing, or anything that does not differentiate the schema) and on the number of attributes changed. The categories studied are: 0_FROZEN (totalActivity = 0), 1_ALMOST_FROZEN (totalActivity <= 10 updated attributes), 1_FocusedShot_n_FROZEN (totalActivity > 10 updated attributes), 2_MODERATE (totalActivity <= 90 updated attributes), 3_FocusedShot_n_LOW, 5_ACTIVE (totalActivity > 90 updated attributes).

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Στόχοι

Ένα πληροφοριακό σύστημα που στηρίζεται σε μια βάση δεδομένων είναι ένα σύστημα λογισμικού που συλλέγει, διαχειρίζεται και ανακτά δεδομένα, τα οποία αποθηκεύονται σε μια βάση δεδομένων και διαχειρίζονται από ένα σύστημα διαχείρισης βάσεων δεδομένων **(DBMS)**. Μια σχεσιακή βάση δεδομένων αποτελείται από πίνακες (σχέσεις). Ένας πίνακας έχει πεδία (columns) και εγγραφές (rows). Ως σχήμα πίνακα θεωρείται το όνομα του και το σύνολο των χαρακτηριστικών του και ως σχήμα ΒΔ είναι η ένωση όλων των σχημάτων των πινάκων στη ΒΔ. Επίσης όλα τα σημαντικά DBMS επιτρέπουν στον διαχειριστή να αλλάξει πίνακες και πεδία σε ένα σχήμα βάσης δεδομένων (προσθήκη, διαγραφή και μετονομασία). Το σύνολο των πινάκων μια βάσης δεδομένων και των περιορισμών ορθότητας που τους συνοδεύουν είναι το **σχήμα** της.

Τα συστήματα λογισμικού εξελίσσονται, εξαιτίας των αλλαγών στις απαιτήσεις και αυτές οι αλλαγές μπορεί να επιφέρουν τροποποιήσεις και στις βάσεις δεδομένων. Σε κάθε αλλαγή τα δεδομένα και ο κωδικός εφαρμογής πρέπει να συγχρονιστούν με το νέο σχήμα. Έτσι, μια βάση δεδομένων, στο πέρασμα του χρόνου, αλλάζει εσωτερική δομή καθώς νέοι πίνακες δημιουργούνται, παλαιοί κα-

ταστρέφονται, πεδία προστίθενται, διαγράφονται, μετονομάζονται κλπ. Η διαδικασία αυτή ονομάζεται **εξέλιξη του σχήματος της βάσης δεδομένων** (schema evolution).

Ποιο Είναι το Πρόβλημα

Ως εξέλιξη του λογισμικού θεωρούνται οι αλλαγές που υφίσταται ένα σύστημα λογισμικού με την πάροδο του χρόνου - συνήθως μέσω μιας εξαιρετικά δύσκολης, περίπλοκης και χρονοβόρας διαδικασίας, συντήρησης λογισμικού.

Όπως κάθε συντήρηση του λογισμικού, η εξέλιξη του σχήματος είναι μια διαδικασία που μπορεί να επηρεάσει σοβαρά τον κύκλος ζωής του λογισμικού, καθώς οι ενημερώσεις του σχήματος μπορούν να οδηγήσουν σε 'συντριβή' των εφαρμογών ή να παρέχουν λανθασμένα δεδομένα στους τελικούς χρήστες. Συγκεκριμένα, μια αλλαγή στο σχήμα μιας βάσης δεδομένων μπορεί να οδηγήσει άμεσα σε μια αποτυχία στο περιβάλλον της εφαρμογής (σε περίπτωση διαγραφής ή μετονομασίας) ή σε σημασιολογικά ελαττωματική ή ανακριβή λειτουργία (στην περίπτωση προσθήκης πληροφοριών ή αναδιάρθρωσης). Συνολικά, η εξέλιξη του σχήματος απειλεί τη συντακτική και σημασιολογική εγκυρότητα του περιβάλλοντος εφαρμογών και επηρεάζει σοβαρά προγραμματιστές και τελικούς χρήστες.

Το μεγαλύτερο πρόβλημα κατά την διαδικασία της αλλαγής εντοπίζεται στο ότι οι προγραμματιστές μπορεί να έχουν διαφορετικό επίπεδο γνώσης της βάσης δεδομένων. Όταν τροποποιείται το σχήμα, εάν οι προγραμματιστές εφαρμογών δεν καταλαβαίνουν γιατί και πώς αλλάζει το σχήμα, μπορούν να αντιμετωπίζουν δυσκολίες στην ορθή εξέλιξη του κώδικα των εφαρμογών που περιβάλλουν την ΒΔ.

Το αποτέλεσμα είναι ότι οι εφαρμογές και το σχήμα δεν είναι συγχρονισμένα, γεγονός που έχει ως συνέπεια (α) οι εφαρμογές ενίοτε να αποτυγχάνουν, λόγω της έλλειψης συντακτικής ορθότητας κάποιων ερωτήσεων που υποβάλλονται στη βάση δεδομένων, και, (β) οι εφαρμογές να πάσχουν από ζητήματα εννοιολογικής ορθότητας, αν ρωτούν λάθος πίνακες ή πεδία, ούσες μη ενημερωμένες για τις αλλαγές που έχουν -εν τω μεταξύ- επέλθει στο σχήμα της βάσης δεδομένων.

Το να κατανοήσουμε το πώς οι βάσεις δεδομένων εξελίσσονται στην πράξη, θα έχει ως αποτέλεσμα αφενός να μπορούμε να επιστήσουμε την προσοχή στους προγραμματιστές εφαρμογών και τους διαχειριστές για τους τρόπους με τους οποίους μπορεί μια βάση δεδομένων να είναι αναντίστοιχη με τις εφαρμογές που την προσπελάζουν, και αφετέρου, να μπορούμε να προγραμματίζουμε χρόνο και πόρους, για μελλοντική συντήρηση ώστε να μειωθεί τόσο η προσπάθεια όσο και το κόστος που συνδέονται με μια αλλαγή και να γίνονται οι αλλαγές οργανωμένα και ομαλά.

Γιατί το Πρόβλημα Είναι Σημαντικό και Ενδιαφέρον

Ιστορικά, η εξέλιξη της βάσης δεδομένων έχει μελετηθεί σε πολύ μικρό εύρος. Αυτό το έλλειμμα στη γνώση είναι δυσανάλογο με τη σοβαρότητα των συνεπειών της εξέλιξης της βάσης.

Πολύ σημαντική είναι η γνώση των προτύπων στον κύκλο ζωής και στην διαγραφή των πινάκων. Η αφαίρεση ή η ενημέρωση ενός πίνακα απαιτεί συντήρηση του κώδικα των εφαρμογών που περιβάλλουν την ΒΔ. Ως εκ τούτου, η κατανόηση της πιθανότητας της αλλαγής ή της αφαίρεσης ενός πίνακα μπορεί να βοηθήσει την ομάδα ανάπτυξης στην αποφυγή πάρα πολλών προσπαθειών στην συντήρηση του κώδικα.

Συνεπώς, η κατανόηση των θεμελιωδών μηχανισμών και των πρότυπων πίσω από την εξέλιξη του σχήματος είναι μεγάλης σημασίας, καθώς μπορούν να δείξουν προβλήματα σχετικά με τον τρόπο χρήσης των βάσεων δεδομένων, να προβλέψουν την αλλαγή των μελλοντικών πινάκων και να προσαρμόσουν την ανάπτυξη των εφαρμογών, τη συντήρηση και τους πόρους της διαχείρισης, στις επικείμενες τάσεις.

Στόχος αυτής της εργασίας είναι να παρουσιασθούν τα πρότυπα που αφορούν την εξέλιξη των πινάκων σχεσιακών βάσεων δεδομένων και να επιβεβαιωθούν ή όχι σε πραγματικές βάσεις δεδομένων. Επίσης, με βάση μια πρόσφατη κατηγοριοποίηση των βάσεων δεδομένων ως προς τον τρόπο με τον οποίο εξελίσσονται συνολικά [Vass21], στόχο είναι να μελετηθεί η εξέλιξη πινάκων σε σχέση με τα εν λόγω πρότυπα, στις επιμέρους κατηγορίες που προτείνονται στη βιβλιογραφία.

1.2 Δομή της Διατριβής

Τα κεφάλαια αυτής της εργασίας είναι :

Κεφάλαιο 2 : παρουσίαση των μελετών για την εξέλιξη των Βάσεων Δεδομένων και σε ποια βασικά ερωτήματα απαντούν αυτές οι μελέτες.

Κεφάλαιο 3 : Συνοπτική αναφορά στις μελέτες ταξινομημένες με βάση τα επιμέρους θέματα που ερευνούν.

Κεφάλαιο 4 : Μελέτη του προτύπου Γ στις κατηγορίες στις οποίες ανήκουν 35 σύνολα δεδομένων, με βάση την δραστηριότητα των αλλαγών σε κάθε κατηγορία. Μελέτη της συμπεριφοράς των μεγάλων (wide) πινάκων σε κάθε κατηγορία.

Κεφάλαιο 5 : Μελέτη του προτύπου αντίστροφο Γ στις κατηγορίες στις οποίες ανήκουν 195 σύνολα δεδομένων με βάση την δραστηριότητα των αλλαγών σε κάθε κατηγορία. Μελέτη της συμπεριφοράς των πινάκων με πολλές αλλαγές σε κάθε κατηγορία.

Κεφάλαιο 6 : Συνοπτική παρουσίαση των συμπερασμάτων και προτάσεις για μελλοντική εργασία.

ΚΕΦΑΛΑΙΟ 2

ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΜΕΛΕΤΩΝ

Στη συγκεκριμένη εργασία μελετήθηκαν και παρουσιάζονται συνοπτικά 8 μελέτες που αφορούν το θέμα της εξέλιξης των Βάσεων Δεδομένων (ΒΔ).
Οι 8 εργασίες είναι:

Πίνακας 2-1 Οι 8 μελέτες για την Εξέλιξη των Βάσεων Δεδομένων

Collateral Evolution of Applications and Databases	[LiNe09]	2009
Schema Evolution Analysis for Embedded Databases	[WuNe11]	2011
An Empirical Analysis of the Co-evolution of Schema and Code in Database Applications	[QiLS13]	2013
Understanding database schema evolution: A case study	[CGMM13]	2013
Open-Source Databases: Within, Outside, or Beyond Lehman's Laws of Software Evolution?	[SkVZ14]	2014
Growing up with stability: How open-source relational databases evolve	[SkVZ15]	2015
Gravitating to Rigidity: Patterns of Schema Evolution – and its Absence – in the Lives of Tables	[VaZS16]	2016
Schema Evolution Survival Guide for Tables: Avoid Rigid Childhood and You 're En Route to a Quiet Life	[VaZa17]	2017

Οι μελέτες [LiNe09] και [QiLS13] ερευνούν την *ταυτόχρονη εξέλιξη εφαρμογών και σχήματος* και αποκάλυψαν ότι τα σχήματα και ο κώδικας δεν εξελίσσονται πάντα συγχρονισμένα.

Στην μελέτη [WuNe11] οι συγγραφείς εργάστηκαν σε 4 Βάσεις Δεδομένων και έδειξαν ότι τα σχήματα των βάσεων δεδομένων τείνουν να σταθεροποιηθούν στο χρόνο, με τις περισσότερες αλλαγές να συμβαίνουν στην αρχή της ζωής του σχήματος και να οδηγούνται σε ένα σταθερό σχήμα αργότερα..

Στις μελέτες [SkVZ14] και [SkVZ15] ερευνήθηκε_από τους συγγραφείς η δυνατότητα εφαρμογής των νόμων του Lehman για την εξέλιξη του λογισμικού στην περίπτωση των σχημάτων βάσης δεδομένων. Αναφέρονται στοιχεία που αποδεικνύουν ότι τα σχήματα αναπτύσσονται με τέτοιο τρόπο ώστε να ικανοποιούν νέες απαιτήσεις. Ωστόσο, αυτή η ανάπτυξη δεν εξελίσσεται γραμμικά ή μονοτονικά, αλλά με περιόδους ηρεμίας και σύντομες περιόδους εστίασης στη συντήρηση. Μετά από αρκετές αλλαγές, το μέγεθος του σχήματος φτάνει σε ένα πιο ώριμο επίπεδο σταθερότητας.

[Στην [CGMM13] ερευνήθηκε μια βάση δεδομένων με μεγάλους πίνακες (πάνω από 200 στήλες). Μια ενδιαφέρουσα ιδιότητα που εντοπίστηκε είναι η *σταθερότητα* των πινάκων. Ένας πίνακας που έχει δημιουργηθεί εδώ και πολύ καιρό, και δεν υπόκειται σε συχνές τροποποιήσεις μπορεί να θεωρηθεί ***σταθερός***. Στην μελέτη αυτή έγινε μια αντιστοίχηση των προγραμματιστών με την εξέλιξη του σχήματος. Οι συγγραφείς υπολόγισαν για κάθε προγραμματιστή, τον αριθμό των πινάκων στην εξέλιξη των οποίων συμμετείχε (by creating, updating or deleting tables).

Στην εργασία **[**VaZS16**]** η προσοχή στρέφεται στη ζωή των πινάκων. Μελετήθηκαν οι σχέσεις της διάρκειας του πίνακα, της επιβίωσης και των ενημερώσεων του πίνακα με χαρακτηριστικά του πίνακα όπως το μέγεθος του σχήματος, ο χρόνο δημιουργίας (γέννησης) του πίνακα κ.λπ. Σε αυτή τη μελέτη, εξήχθησαν τα παρακάτω μοντέλα που αφορούν την εξέλιξη:

- Το *πρότυπο Γ* μελετά την αλληλεξάρτηση του μεγέθους ενός πίνακα με τη συνολική διάρκεια ζωής του και δείχνει ότι πίνακες με μεγάλα σχήματα τείνουν να έχουν μακρές διάρκειες και να αποφεύγουν την διαγραφή.

- Το πρότυπο *Κομήτη* (*Comet)* μελετά την αλληλεπίδραση του μεγέθους ενός πίνακα κατά τη γέννησή του με το συνολικό ποσό των ενημερώσεων και αποδει-

κνύει ότι οι πίνακες με τις περισσότερες αναβαθμίσεις είναι συχνά αυτοί με μεσαίο μέγεθος σχήματος.

- Το *πρότυπο αντίστροφου Γ* μελετά την αλληλεπίδραση του αριθμού των ενημερώσεων με τη διάρκειά και δείχνει ότι οι πίνακες με μεσαίες ή μικρές διάρκειες έχουν ποσότητες ενημερώσεων χαμηλότερες από τις αναμενόμενες, ενώ οι πίνακες με μεγάλες διάρκειες παρουσιάζουν όλα τα είδη της συμπεριφοράς κατά την διαδικασία των ενημερώσεων.

- Το *πρότυπο άδειου Τριγώνου (Empty Triangle) εξετάζει* την αλληλεπίδραση της γέννησης ενός πίνακα με τη συνολική διάρκεια της και αποδεικνύει σημαντική απουσία διαγραφών πινάκων μεσαίας ή μεγάλης διάρκειας, που σημαίνει κυρίως σύντομες ζωές για διαγραμμένους πίνακες και χαμηλή πιθανότητα διαγραφής για παλιούς πίνακες.

Στην μελέτη [VaZa17] οι συγγραφείς ερευνούν κατά πόσο η δραστηριότητα σχετικά με τις ενημερώσεις και η διάρκεια σχετίζονται με την επιβίωση. Τα ευρήματα σε ένα πρότυπο, το οποίο ονομάστηκε από τους συγγραφείς *ηλεκτρόλυση*, δηλώνοντας ότι οι διαγραμμένοι και επιζώντες πίνακες ζουν πολύ διαφορετικές ζωές. Εντόπισαν μια αντίθεση στις ζωές νεκρών και επιζώντων πινάκων: ενώ οι νεκροί πίνακες έχουν οι ζωές μικρής ή μεσαίας διάρκειας, οι επιζώντες βρίσκονται κυρίως σε μεσαίες ή υψηλές διάρκειες και όσο πιο δραστήριοι είναι, τόσο ισχυρότερα προσελκύονται από υψηλές διάρκειες.

Εξίσου σημαντική είναι η απόδειξη ότι η εξέλιξη του σχήματος "πάσχει" από τον ανταγωνισμό της βαρύτητας με την ακαμψία, δηλαδή την τάση να ελαχιστοποιήσει την εξέλιξη όσο το δυνατόν περισσότερο, προκειμένου να ελαχιστοποιηθεί η προκύπτουσα επίπτωση στον περιβάλλοντα κώδικα.

2.1 Χρονική Εξέλιξη της Έρευνας στο υπό Μελέτη Πρόβλημα

Την έρευνα πάνω στον τομέα της εξέλιξης του σχήματος βοήθησε σημαντικά η εμφάνιση του ελεύθερου λογισμικού ανοιχτού κώδικα. Πριν από τη διαθεσιμότητα του ιστορικού του σχήματος στα αποθετήρια (svn/sourceforge/github), οι ερευνητές δεν μπορούσαν να εργαστούν με πραγματικά δεδομένα στον τομέα της εξέλιξης του σχήματος. Αξιοσημείωτο είναι ότι μέχρι τα τέλη του περασμένου αιώνα, υπήρξε μόνο μια μεμονωμένη μελέτη περίπτωσης. Όπως φαίνεται και στην χρονο-

λογική εξέλιξη των μελετών υπήρξε ένα κενό 15 χρόνων στην έρευνα της συγκεκριμένης επιστημονικής περιοχής. Οι 8 μελέτες που παρουσιάζονται στην εργασία διενεργήθηκαν από το 2009 έως το 2017.

Σχήμα 2-1 : Χρονολογική εξέλιξη των μελετών

2.2 Ερωτήματα που Απαντήθηκαν στις Μελέτες

Κάποια από τα βασικά ερωτήματα που απαντήθηκαν στις 8 μελέτες κατατάχθηκαν σε 3 βασικές κατηγορίες και παρουσιάζονται στον πίνακας 1. Όλες οι μελέτες ασχολούνται με τις ερωτήσεις της πρώτης κατηγορίας που είναι η μακροσκοπική εξέλιξη του σχήματος στον χρόνο. Ερευνούν πόσο συχνά και σε τι βαθμό αλλάζουν τα σχήματα από έκδοση σε έκδοση για να κατανοηθεί πως εξελίσσονται στις φάσεις της ανάπτυξης και της συντήρησης.

Μια άλλη παρατήρηση είναι ότι οι πρώτες μελέτες που έγιναν στο συγκεκριμένο πεδίο επικεντρώθηκαν σε αρκετές ποσοτικές μετρήσεις (π.χ. πόσα αρχεία προστέθηκαν, πόσα διαγραφήκαν, πόσες αλλαγές έγιναν σε πεδία). Οι μελέτες που ακολούθησαν χρονικά προσπάθησαν να εντοπίσουν πρότυπα συμπεριφοράς που ακολουθούν οι πίνακες στην διαδικασία της εξέλιξης τους.

Πίνακας 2-2 Ερωτήματα που απαντήθηκαν στις 8 μελέτες

			QiLS13	LiNe09	WuNe11	SkVZ14	SkVZ15	CGMM13	VaZS16	VaZa17
	Αριθμός Βάσεων δεδομένων		*10*	*2*	*4*	*8*	*8*	*1*	*8*	*8*
	Διάρκεια έρευνας (σε μήνες)		*38 - 122*	*43-74*	*23-87*	*31-160*	*31-160*	*120*	*31-160*	*31-160*
1	**Μακροσκοπική περιγραφή της εξέλιξης στον χρόνο**									
1.1	Ποσο συχνά αναπτύσσονται τα σχήματα		√	√	√	√	√	√	√	√
1.2	Πως εξελίσσεται το μέγεθος του σχήματος στον χρόνο	*Σε πίνακες*	√	√	√	√	√	√	√	
		Σε στήλες	√	√	√	√	√	√	√	
1.3	Αλλαγές κατά χρονική περιοδο	release	√		√					
		version	√		√	√	√	√	√	√
		έτος	√		√					
1.4	Πόσο αλλάζουν τα χαρακτηριστικά της ΒΔ σε σχέση με την φάση της ανάπτυξης		√	√	√	√	√	√	√	√
2	**Εσωτερική δομή και ανάλυση των αλλαγών**									
2.1	Κατανομή των αλλαγών ανά τύπο αλλαγής		√	√	√			√		
2.2	Ποιο ποσοστό των πινάκων αλλάζει σπάνια στη διάρκεια του κύκλου ζωής		40%	41%				20%		63%
2.3	Πόσο συχνά αλλάζουν τα foreign keys; (*πώς εξελίσσονται τα ξένα κλειδιά με την πάροδο του χρόνου;*)		√	√				√		
2.4	Ύπαρξη ενός μηχανισμού ανατροφοδότησης που ρυθμίζει την εξέλιξη (Νομος Lehman - VIII) -					√	√			
2.5	Ποιοι παράγοντες επηρεάζουν τελικά την διατήρηση ή την διαγραφή ενός πίνακα;								√	√
2.6	Συσχέτιση προγραμματιστών με αλλαγές στους πίνακες του σχήματος							√		
3	**Εξωτεική σχέση των αλλαγών του σχήματος με τον περιβάλλοντα κώδικα**									
3.1	Ποιες κατηγορίες αλλαγών του σχήματος έχουν τον πιο σημαντικό αντίκτυπο στον κώδικα εφαρμογής ;		√	√						
3.2	Κατά πόσο οι αλλαγές στο σχήμα προκαλούν αλλαγές στον κώδικα; (μέτρηση σε αριθμό γραμμών κώδικα)		100 -1000							

2.2.1 Μακροσκοπική Περιγραφή της Εξέλιξης στο Χρόνο

Όλες οι μελέτες που παρουσιάζονται σε αυτή την εργασία ερευνούν το πόσο συχνά αλλάζουν τα σχήματα ώστε να κατανοηθεί αν εξελίσσονται έντονα στις φάσεις της ανάπτυξης και της συντήρησης. Στην μελέτη [QiLS13] επισημαίνεται ότι τα σχήματα βάσης δεδομένων εξελίσσονται σημαντικά κατά τη φάση της ανάπτυξη των εφαρμογών της ΒΔ. Στην [WuNe11] παρατηρήθηκε από τους συγγραφείς, ότι τα σχήματα τείνουν να αλλάζουν περισσότερο στην αρχή του κύκλου ζωής τους και με την πάροδο του χρόνου η δομή σταθεροποιείται (οι μεταγενέστερες εκδόσεις έχουν λιγότερες αλλαγές). Στις μελέτες [SkVZ14] και [SkVZ15] όλα τα σύνολα δεδομένων δείχνουν την τάση να αυξηθούν με την πάροδο του χρόνου. Η αλλαγή όμως έρχεται μετά από συχνές (και μερικές φορές μεγάλες) περιόδους

ηρεμίας, όπου το μέγεθος του σχήματος δεν αλλάζει (ή αλλάζει ελάχιστα). Οι αλλαγές συμβαίνουν κατά διαστήματα και δεν ακολουθούν πρότυπα σταθερής συμπεριφοράς. Επίσης η ηλικία του σχήματος έχει ως αποτέλεσμα τη μείωση της πυκνότητας των αλλαγών. Σε γενικές γραμμές όλες οι μελέτες που απαντούν σε αυτό ερώτημα, εντοπίζουν τις περισσότερες αλλαγές στις αρχικές εκδόσεις του σχήμα τος (ή στην αρχή του κύκλου ζωής του).

2.2.2 Εσωτερική Δομή και Ανάλυση των Αλλαγών

Σε 4 μελέτες [LiNe09], [WuNe11], [QiLS13], [CGMM13] καταμετρήθηκαν οι αλλαγές που συμβαίνουν στην βάση δεδομένων ανά είδος αλλαγών (*Ερώτηση 2.1*). Οι αλλαγές αυτές χωρίστηκαν σε 2 μεγάλες κατηγορίες, τις αλλαγές που σχετίζονται με τους πίνακες και τις αλλαγές που σχετίζονται με τα πεδία. Οι 2 βασικές αλλαγές που εμφανίζονται στους πίνακες κατά την διάρκεια ζωής τους σχήματος είναι η προσθήκη πίνακα και η διαγραφή πίνακα. Σε ορισμένες περιπτώσεις μάλιστα όπως φαίνεται και στην Εικόνα 2-2[1], υπάρχει εντυπωσιακά μεγάλος αριθμός προσθήκης πινάκων (279 πίνακες προστέθηκαν στη ΒΔ Tikiwiki στη μελέτη [QiLS13]). Όσον αφορά τις αλλαγές στα πεδία (Εικόνα 2-3), μελετήθηκαν αρκετά είδη αλλαγών που σχετίζονται με αυτά. Η αλλαγή που αφορά την πρόσθεση νέων πεδίων είναι η πολυπληθέστερη. Αυτό δείχνει την ανάγκη του σχήματος να καλύψει νέες απαιτήσεις στην διάρκεια του κύκλου ζωής του.

Σε 2 μελέτες [SkVZ14] και [SkVZ15] εξετάσθηκε κατά πόσο μπορούν να εφαρμοστούν οι νόμοι του Lehman για την εξέλιξη του λογισμικού στην περίπτωση των σχημάτων βάσης δεδομένων (*Ερώτηση 2.4*). Οι συγγραφείς επιβεβαίωσαν την ύπαρξη ενός μηχανισμού ανατροφοδότησης που περιορίζει την ανεξέλεγκτη εξέλιξη των βάσεων δεδομένων. Η ανάπτυξη γίνεται για να καλύψει τις ανάγκες των χρηστών αλλά και για την συντήρηση των ΒΔ ώστε να διατηρηθεί το σχήμα "καθαρό" (π.χ. χωρίς πίνακες που δεν χρησιμοποιούνται, περιττά πεδία κ.τ.λ). Έτσι η ανάπτυξη συχνά ταλαντεύεται μεταξύ των θετικών και αρνητικών τιμών.

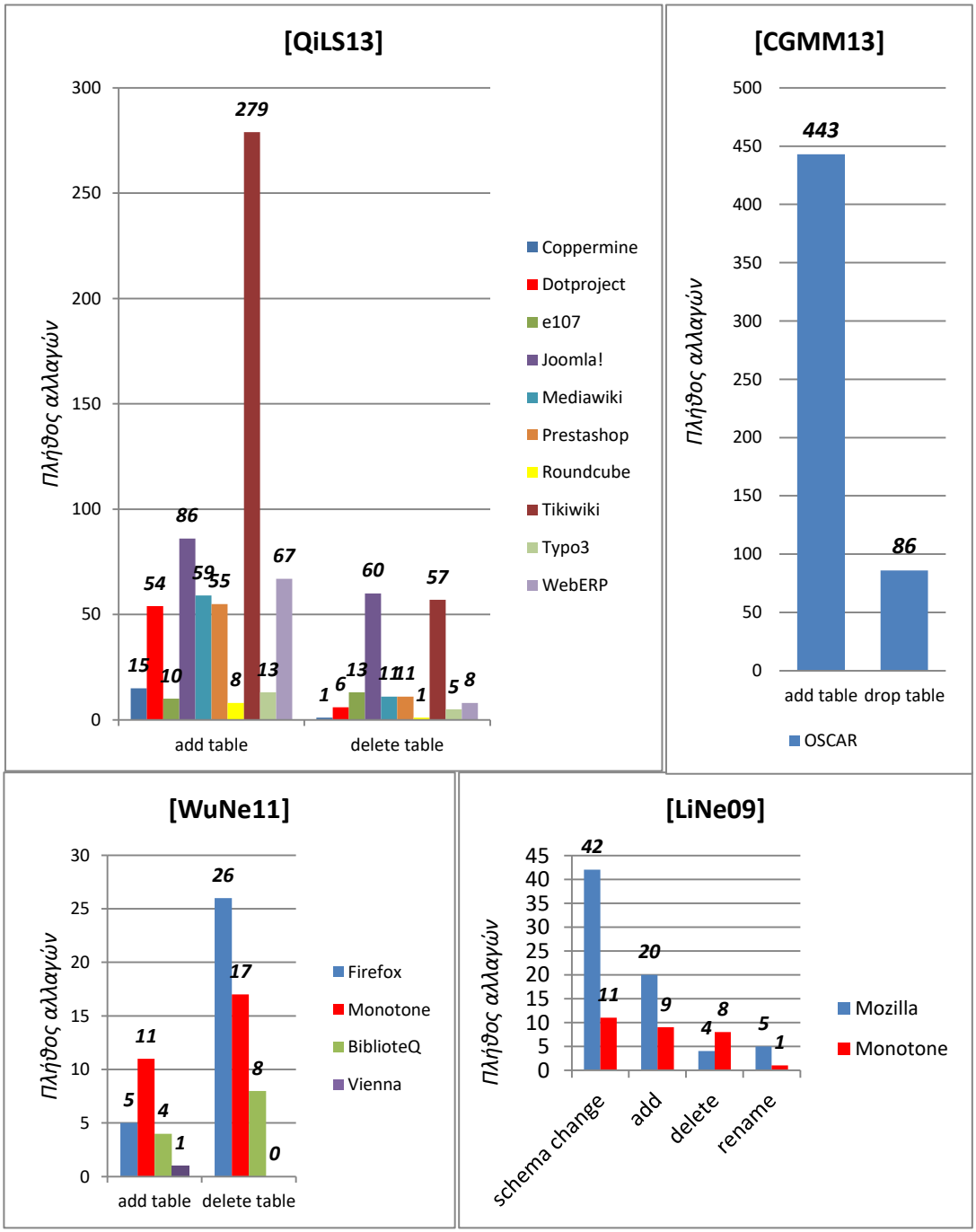
[1] Στα σχήματα 2-2 και 2-3, έχουμε προσφέρει μια *δική μας* διαγραμματική απεικόνιση των συλλεγέντων στοιχείων που παραθέτουν οι εργασίες που αναφέρονται στα εν λόγω διαγράμματα .

Στ ερώτημα "2.2 *Ποιο ποσοστό των πινάκων αλλάζει σπάνια στη διάρκεια του κύκλου ζωής;*", απαντούν 4 μελέτες και αν εξαιρέσουμε την [CGMM13] η οποία μελετά μόνο μία ΒΔ και υπολογίζει το ποσοστό σε ***20%***, οι άλλες μελέτες (που έγιναν σε περισσότερες από μία βάσεις δεδομένων) υπολογίζουν ποσοστά που κυμαίνονται από 40% έως 63%.

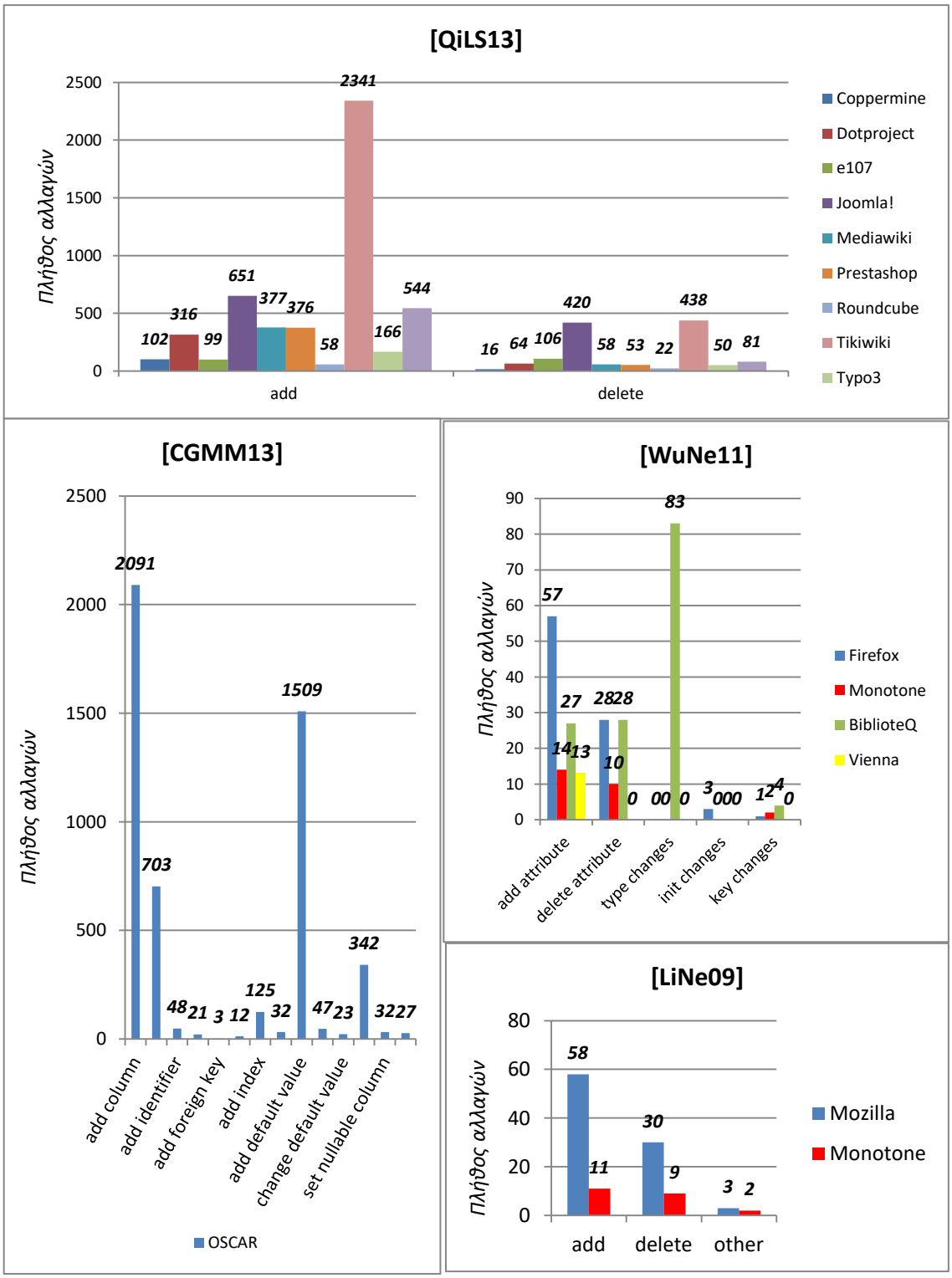
Σε δύο μελέτες, στη [VaZS16] και στη [VaZa17], ερευνήθηκε το ερώτημα (2.6) που αφορά τους παράγοντες που επηρεάζουν την διατήρηση ή την διαγραφή ενός πίνακα. Αποδείχθηκε ότι η διατήρηση ενός πίνακα εξαρτάται από παράγοντες όπως η διάρκεια ζωής του, το μέγεθος κατά την ημερομηνία δημιουργίας του, τον αριθμό των αλλαγών που έγιναν στον συγκεκριμένο πίνακα. Εξίσου σημαντική είναι και η απόδειξη ότι η εξέλιξη του σχήματος παρουσιάζει μια έλξη προς την ακαμψία, δηλαδή την τάση να ελαχιστοποιήσει την εξέλιξη όσο το δυνατόν περισσότερο, προκειμένου να ελαχιστοποιηθούν οι επιπτώσεις στον περιβάλλοντα κώδικα.

2.2.3 Εξωτερική Σχέση των Αλλαγών του Σχήματος με τον Περιβάλλοντα Κώδικα

Σε αυτό το ερώτημα απαντούν οι μελέτες [QiLS13], [LiNe09] που εξετάζουν τη σχέση των αλλαγών του σχήματος με τον περιβάλλοντα κώδικα. Οι αλλαγές του σχήματος που προκαλούν τις περισσότερες αλλαγές στον κώδικα είναι η προσθήκη και διαγραφή πίνακα και πεδίου. Στην [QiLS13] μεγάλο ποσοστό αλλαγών στον περιβάλλοντα κώδικα προκαλούν και οι αλλαγές στις default value ενός πεδίου καθώς και η αλλαγή της τιμής ενός πεδίου από "not null" σε "null" ή το αντίστροφο. Στην [QiLS13] καταμετρήθηκαν και οι γραμμές κώδικα που αλλάζουν μετά από μια ατομική αλλαγή σχήματος και μετά από μια έγκυρη αναθεώρηση βάσης δεδομένων (αυτή η μέτρηση έγινε μόνο σε αυτή την μελέτη από τις 8 που εξετάστηκαν). Ως συμπέρασμα οι συγγραφείς καταλήγουν στο ότι αν οι αλλαγές στο σχήμα της ΒΔ δεν γίνουν ταυτόχρονα με αλλαγές στον κώδικα της εφαρμογής μπορεί να οδηγήσουν σε απώλεια δεδομένων, application crash ή μείωση της απόδοσης.



Σχήμα 2-2 : Γραφική αναπαράσταση των αλλαγών σε αρχεία όπως καταμετρήθηκαν σε 4 μελέτες



Σχήμα 2-3 : Γραφική αναπαράσταση των αλλαγών σε πεδία όπως καταμετρήθηκαν σε 4 μελέτες

ΚΕΦΑΛΑΙΟ 3

Σύνοψη των Εργασιών που Μελετήθηκαν

3.1 Ταυτόχρονη Εξέλιξη Εφαρμογών και Σχήματος

Η εξέλιξη της ΒΔ και του λογισμικού είναι απαραίτητη στην διάρκεια ζωής μιας εφαρμογής έτσι ώστε να παραμείνει ανταγωνιστική και οι πάροχοι του λογισμικού βρίσκονται κάτω από συνεχώς αυξανόμενη πίεση για νέες εκδόσεις του λογισμικού, διόρθωση σφαλμάτων και νέες λειτουργίες. Η πιο κοινή τακτική είναι να σταματήσει η λειτουργία της εφαρμογής να γίνει ενημέρωση και στην συνέχεια επανεκκίνηση. Αυτό είναι πολύ ενοχλητικό για τον χρήστη και έτσι πολλές εφαρμογές για κινητά, υπολογιστές και servers αντί για αποθήκευση δεδομένων σε

προσαρμοσμένες μορφές αρχείων οδηγούνται προς την αποθήκευση δεδομένων χρησιμοποιώντας ένα σύστημα διαχείρισης βάσεων δεδομένων που περιέχονται μέσα στην εφαρμογή, κρυμμένα από τον χρήστη και απαιτούν ελάχιστη ή καθόλου συντήρηση. Αυτά τα συστήματα ονομάζονται ***Embedded Databases (EDs)***. Μία ED δίνει τη δυνατότητα διαχείρισης μιας εφαρμογής δεδομένων με ασφαλέστερο και πιο ευέλικτο τρόπο, ενώ συγχρόνως καθιστά ευκολότερα τα ερωτήματα στη ΒΔ.

3.1.1 Παράλληλη Εξέλιξη [LiNe09]

Στην εργασία [LiNe09] χρησιμοποιείται ο όρος *collateral evolution (παράλληλη εξέλιξη)* για να υποδηλώσει τις πιθανές ασυνέπειες που προκύπτουν όταν μια ΒΔ και τα προγράμματα εφαρμογών σε αυτή τη ΒΔ, δεν εξελίσσονται συγχρονισμένα. Γίνεται μια εμπειρική μελέτη όπου οι συγγραφείς μελετάνε ποιες αλλαγές είναι ***συχνές*** στη βάση δεδομένων (προσθήκη/διαγραφή/μετονομασίαα πίνακα ή αντίστοιχες αλλαγές σε πεδία). Η μελέτη τονίζει ότι ο πηγαίος κώδικας των εφαρμογών δεν εξελίσσεται πάντα συγχρονισμένα με τις αλλαγές στο σχήμα βάσης δεδομένων.

Η μελέτη έγινε σε ***2*** εφαρμογές ανοικτού κώδικα (Mozilla για 43 μήνες, Monotone για 74 μήνες) που έχουν μεγάλη διάρκεια και χρησιμοποιούν μεγάλα σχήματα με δεκάδες πίνακες. Για κάθε revision της εφαρμογής Mozilla και για κάθε version της Monotone εξήγαγαν τους πίνακες και τα χαρακτηριστικά τους manually από τον πηγαίο κώδικα C ++. Μελέτησαν τα επίσημα εγχειρίδια και οδηγούς μετεγκατάστασης που συνοδεύουν τα συστήματα διαχείρισης ΒΔ (το SQLite, MySQL και PostgreSQL). Βάσει αυτών των πληροφοριών, οι συγγραφείς χαρακτήρισαν την εξέλιξη των αλλαγών μορφοποίησης σε όλη τη διάρκεια ζωής για από κάθε ένα από τα τρία DBMS, καθώς και τους μηχανισμούς που χρησιμοποιούν τα DBMS για την πρόσβαση και τη μετεγκατάσταση όταν γίνονται αλλαγές.

Στην μελέτη αυτή οι συγγραφείς κατέταξαν τις παρακάτω αλλαγές ως προς την συχνότητα εμφάνισης:

- Αλλαγή στο σχήμα πίνακα (***42*** στη Mozilla και ***11*** στη Monotone)

- Προσθήκη πίνακα (***20*** στη Mozilla και ***9*** στη Monotone)
- Διαγραφή πίνακα (***4*** στη Mozilla και ***8*** στη Monotone)
- Μετονομασία πίνακα (***5*** στη Mozilla και ***1*** στη Monotone)

Όσον αφορά τα χαρακτηριστικά (attributes) η πλειοψηφία των αλλαγών αφορά προσθήκες (***58*** στη Mozilla και ***11*** στη Monotone) και διαγραφές (***30*** στη Mozilla και ***9*** στη Monotone). Και στις 2 ΒΔ εντόπισαν μόνο ***1*** αλλαγή τύπου, μόνο ***1*** αλλαγή κλειδιού, ***3*** μετονομασίες και ***2*** αλλαγές αρχικής τιμής.

Από τον Table 4 της μελέτης (Schema changes details for Mozilla and Monotone) εξήχθησαν τα εξής: η ΒΔ Mozilla έχει ***26*** πίνακες από τους οποίους οι 10 δεν είχαν καμία αλλαγή (***38%***) και οι 4 παρουσίασαν μεγάλο αριθμό αλλαγών (***15%***) ενώ η ΒΔ Monotone έχει ***23*** πίνακες από τους οποίους οι 10 δεν είχαν καμία αλλαγή (***43%***) και οι 3 παρουσίασαν μεγάλο αριθμό αλλαγών (***13%***). Συνεπώς ο μέσος όρος των πινάκων χωρίς αλλαγές είναι **41%** ενώ ο μέσος όρος των πινάκων με πολλές αλλαγές είναι **14%.**

3.1.2 Ανάλυση και Ποσοτικοποίηση των Επιπτώσεων της Εξέλιξης [QiLS13]

Στην εργασία [QiLS13] οι Qiu, Li και Su αναφέρουν τα ευρήματα από μιας μεγάλης κλίμακας εμπειρική μελέτη σε δέκα δημοφιλείς εφαρμογές βάσεων δεδομένων, ώστε να εξαχθούν πληροφορίες για το πώς εξελίσσονται τα σχήματα και ο κώδικας ταυτόχρονα. Οι συγγραφείς επικεντρώθηκαν στο γεγονός ότι οι αλλαγές στο σχήμα της βάσης δεδομένων και στον κώδικα των εφαρμογών που περιβάλλουν την ΒΔ, πραγματοποιούνται σε διαφορετικές χρονικές στιγμές. Αυτό είναι πολύ πιο δύσκολο καθώς οι αλλαγές του σχήματος δεν επιδρούν άμεσα στον κώδικα.

Συγκεκριμένα, οι συγγραφείς μελετάνε

1. την ιστορική εξέλιξη των εφαρμογών για να κατανοήσουν εάν η βάση δεδομένων εξελίσσεται συχνά (*πόσο*) και σημαντικά (*πως*),
2. πώς εξελίσσονται τα σχήματα και πώς επηρεάζουν τον κώδικα των εφαρμογών που περιβάλλουν την ΒΔ.

Η μελέτη έγινε σε ***10*** εφαρμογές βάσεων δεδομένων ανοιχτού κώδικα από διάφορους τομείς. Ως σύστημα ελέγχου εκδόσεων επιλέχθηκε το Subversion

(SVN). Οι εφαρμογές εξετάσθηκαν σε διάρκεια από 38 έως 122 μήνες (μέσος όρος διάρκειας 87 μήνες). Σε 9 εφαρμογές οι πίνακες παρουσίασαν αύξηση του αριθμού των πινάκων από ***7*** έως ***222*** πίνακες και μόνο 1 εφαρμογή μείωση πινάκων (-3 πίνακες). Οι γραμμές κώδικα - LoC αυξήθηκαν σε όλες τις εφαρμογές με μέσο όρο ***373*** γραμμές.

Οι Ερευνητικές Ερωτήσεις που έθεσαν οι συγγραφείς στοχεύουν να απαντήσουν στον τρόπο που τα σχήματα και ο κώδικας εξελίσσονται σε εφαρμογές βάσεων δεδομένων.

1: Πόσο συχνά και εκτενώς εξελίσσονται τα σχήματα; Πόσο συχνά και σε τι βαθμό αλλάζουν τα σχήματα από έκδοση σε έκδοση για να κατανοηθεί αν εξελίσσονται έντονα στις φάσεις της ανάπτυξης και της συντήρησης.

2: Πώς εξελίσσονται τα σχήματα βάσης δεδομένων; Αναλύονται όλες οι πιθανές αλλαγές σε ένα σχήμα για να κατανοηθεί τι τύποι αλλαγών εμφανίζονται συχνότερα στην πράξη και αν κάποιες αλλαγές εμφανίζονται συχνότερα σε σχέση με άλλες.

3: Πόσος κώδικας της εφαρμογής αλλάζει ταυτόχρονα με τις αλλαγές στο σχήμα; Αυτό το ερώτημα βοηθάει να γίνει ποσοτικοποίηση των επιπτώσεων των αλλαγών του σχήματος στον κώδικα της εφαρμογής που περιβάλλει τη ΒΔ. Χρησιμοποιώντας το ιστορικό των αλλαγών από το αποθετήριο αντιστοιχήθηκαν οι αλλαγές στις γραμμές του κώδικα με τις αλλαγές στο σχήμα, για να εκτιμηθεί η επιρροή των αλλαγών του σχήματος στον κώδικα. Το ενδιαφέρον των συγγραφέων επικεντρώθηκες επίσης, στο κατά πόσο ορισμένοι τύποι αλλαγών σχήματος τείνουν να έχουν μεγαλύτερο αντίκτυπο στον κώδικα σε σχέση με άλλους.

Διαδικασία ανάλυσης (Βήματα για την εξαγωγή των απαραίτητων πληροφοριών)

- Locate schema files – Οι συγγραφείς εντόπισαν τα αρχεία σχήματος και όπου απαιτήθηκε έκαναν ακόμη και χειροκίνητο εντοπισμό των αρχείων του σχήματος, στην περίπτωση που οι τοποθεσίες ή τα ονόματα είχαν τροποποιηθεί.
- Extract DB revisions – Έγινε εξαγωγή των αναθεωρήσεων της ΒΔ που περιείχαν τροποποιήσεις στο σχήμα. Στο SVN, έγινε εύκολα η ανάκτηση των διαδρομών όλων των αλλαγμένων αρχείων σε οποιαδήποτε αναθεώρηση -

εάν υπάρχει αλλαγή αρχείου στην συγκεκριμένη αναθεώρηση τότε θεωρείται αναθεώρηση της DB.

- Extract valid DB revisions – Έγινε εξαγωγή των έγκυρων αναθεωρήσεων της DB.
- Extract atomic changes – Οι συγγραφείς προχώρησαν στην εξαγωγή ατομικών αλλαγών. Μετά την εξαγωγή των έγκυρων αναθεωρήσεων της DB, εξήχθησαν όλες τις αλλαγές του σχήματος συγκρίνοντας χειροκίνητα συνεχόμενες έγκυρες εκδόσεις της ΒΔ.
- Co-change analysis – Οι συγγραφείς ανέλυσαν τις επιπτώσεις που προκλήθηκαν από τις ατομικές αλλαγές στο σχήμα εξάγοντας το ιστορικό των εκδόσεων.

Πόσο συχνά και εκτενώς εξελίσσονται τα σχήματος;

Για να βγάλουν οι συγγραφείς κάποια συμπεράσματα για το πόσο συχνά και πόσο εκτεταμένα αλλάζει ένα σχήμα, υπολογίσθηκαν 2 ποσοστά το ***GR*** (Growth Rate) και το ***CR*** (Change Rate) για τους πίνακες και τις στήλες αντίστοιχα. Αν θεωρήσουμε ως ***AE*** το πλήθος των Added Elements το ***DE*** το πλήθος των Deleted Elements και ως ***IE*** το πλήθος των Initial Elements, το GR είναι ο λόγος ***(AE -DE / IE***) και το CR είναι ο λόγος ***(AE + DE / IE).***

Τα αποτελέσματα στα οποία κατέληξαν οι συγγραφείς είναι: (1) τα σχήματα αναπτύσσονται ***συχνά***: κατά μέσο όρο 90 αλλαγές σχήματος πραγματοποιούνται ετησίως στα δέκα έργα, (2) το μέγεθος των σχημάτων στα περισσότερα έργα αυξήθηκε σημαντικά (το ***GR*** των πινάκων στο 60% των εφαρμογών αυξήθηκε κατά 100% και το CR στο 90% των εφαρμογών αυξήθηκε πάνω από 100%), (3) παρόμοια είναι η τάση στο GR και CR των στηλών και τέλος, (4) σε 7 έργα το μέγεθος των σχημάτων άγγιξε το 60% των μέγιστων τιμών στο 20% του κύκλου ζωής κατά το οποίο μελετήθηκε που σημαίνει ότι τα περισσότερα χαρακτηριστικά των ΒΔ εισήχθησαν στην αρχική φάση της ανάπτυξης του έργου. Συνοπτικά, τα σχήματα βάσης δεδομένων εξελίσσονται σημαντικά κατά τη φάση της ανάπτυξη των εφαρμογών της ΒΔ.

Πώς εξελίσσονται τα σχήματα βάσης δεδομένων;

Για να απαντήσουν σε αυτό το ερώτημα οι συγγραφείς εξέτασαν 6 κατηγορίες αλλαγών σχήματος:

- Transformations (***Trans***): add table, add column, add view
- Structure Refactoring (***SR***): drop table, rename table, drop column, rename column, change column datatype, drop view, add key, drop key
- Referential Integrity Refactoring (***RIR***): add foreign key, drop foreign key, add trigger, drop trigger
- Architectural Refactoring (***AR***): add index, drop index
- Data Quality Refactoring (***DQR***): Add column default value, drop column default value, change column default value, make column not null, make column not null
- Method Refactoring (***MR***): add stored procedure, drop stored procedure

Οι 3 αλλαγές Trans, SR και DQR αποτελούν το ***80%*** των αλλαγών και στις 10 εφαρμογές ενώ αποτελούν το ***95%*** των αλλαγών σε 7 εφαρμογές. Η κατηγορία ***AR*** βρέθηκε σε 8 έργα, η ***RIR*** μόνο σε 3 με πολύ μικρό ποσοστό και τέλος η ***MR*** δεν πραγματοποιήθηκε σε καμία από τις 10 εφαρμογές.

Συγκεκριμένα οι πιο συχνές αλλαγές είναι η προσθήκη πίνακα, η προσθήκη στήλης και η αλλαγή του τύπου (data type) της στήλης. Αντιθέτως αλλαγές που αφορούν αλλαγή σε foreign keys και triggers συμβαίνουν σπάνια στην πράξη.

Πόσος κώδικας της εφαρμογής αλλάζει ταυτόχρονα με τις αλλαγές στο σχήμα;

Για να απαντήσουν στις ερωτήσεις, επέλεξαν ομοιόμορφα και τυχαία το 10% (146) των έγκυρων αναθεωρήσεων των ΒΔ από το σύνολο των 1.464 έγκυρων αναθεωρήσεων των ΒΔ και ανέλυσαν χειρωνακτικά τις γραμμές πηγαίου κώδικα και ταυτόχρονα και ταυτόχρονα ανέλυσαν τις αλλαγές που προκάλεσαν στο σχήμα.

Πρώτα αναλύεται και κατανοείται πως οι αλλαγές παρουσιάζονται στην εξέλιξης της εφαρμογής. Χρησιμοποιούνται οι παρακάτω συμβολισμοί:

R → το σύνολο όλων των έγκυρων αναθεωρήσεων της DB

r → η τρέχουσα αναθεώρηση

Cr → όλες οι αλλαγές που έγιναν σε αυτήν την αναθεώρηση

SCr → οι αλλαγές σχήματος

CCr → οι αλλαγές κώδικα

RCr → οι αλλαγές κώδικα που προκαλούνται από το SCr

Υπάρχουν τέσσερις πιθανές καταστάσεις σχετικές με την ταυτόχρονη αλλαγή σχήματος και κώδικα:

(S1) CCr = 0 και RCr ≠ 0. Όχι στην ίδια αναθεώρηση.

(S2) CCr ≠ 0 και RCr = 0. Αυτό δείχνει ότι οι αλλαγές σχήματος δεν επηρεάζουν τον κώδικα.

(S3) CCr≠0 και CCr RCr ≠0. Αλλαγές κώδικα μαζί με αλλαγές σχήματος. Εάν CCr=RCr, όλος ο αλλαγμένος κώδικας προκλήθηκε από αλλαγές στο σχήμα. (Διαφορετικά, περιλάμβανε άλλες αλλαγές και μπορεί να οδηγήσει σε ανακριβή πληροφορίες)

(S4) CCr6 ≠ 0 και CCr RCr = 0. Οι αλλαγές κώδικα δεν ήταν σχετικές με τις αλλαγές του σχήματος.

Στην ανάλυση βρέθηκε ότι το ***72%*** των έγκυρων αναθεωρήσεων παρείχαν χρήσιμες πληροφορίες σχετικά με την αλλαγή σχήματος-κώδικα. Στο ***22%*** των έγκυρων αναθεωρήσεων της ΒΔ δεν έγινε αλλαγή κώδικα.

Οι αλλαγές στο σχήμα προκαλούν αλλαγές στον κώδικα σε μεγάλο βαθμό. Μια ατομική αλλαγή σχήματος μπορεί να αλλάξει κατά μέσο όρο περίπου ***10 ~ 100*** LoC και σε μια έγκυρη αναθεώρηση βάσης δεδομένων, οι προγραμματιστές αλλάζουν περίπου ***100 ~ 1000*** LoC.

Ανάλυση εξέλιξης σχήματος για ενσωματωμένες Βάσεις δεδομένων

Οι συγγραφείς του [WuNe11] έκαναν μια μακροχρόνια μελέτη εξέλιξης του σχήματος σε τέσσερα δημοφιλή προγράμματα ανοιχτού κώδικα που χρησιμοποιούν ενσωματωμένες βάσεις δεδομένων: Firefox, Monotone, BiblioteQ and Vienna.

Ο γενικός στόχος στη συγκεκριμένη μελέτη είναι να εξαχθούν συμπεράσματα έτσι ώστε όταν αλλάζει ο κώδικας, η ΒΔ να μην παραμένει στη παλιά έκδοση για να μην προκύψει ασυμβατότητα του σχήματος και ενδεχομένως αποτυχία ενημέρωσης. Για να γίνει αυτό πρέπει να επιτρέπονται ασφαλείς και δυναμικές ενημερώσεις σχήματος. Το πρώτο βήμα προς αυτό το στόχο είναι να κατανοηθεί το πώς εξελίσσονται τα σχήματα.

Η μελέτη διεξήχθη με χρήση του εργαλείου ***SCVD*** (*Schema extraCtion και eVolution analysis for embedded Databases)*, που κατασκεύασαν οι συγγραφείς και στοχεύει στην κατανόηση και την ποσοτικοποίηση της εξέλιξης του σχήματος της

ΒΔ. Μελετά το ιστορικό των releases μιας εφαρμογής, κάνει εξαγωγή ***ED***, συγκρίνει τα σχήματα και παράγει ένα σύνολο με τις εξελίξεις του σχήματος.

Στο WuNe11, οι συγγραφείς εντόπισαν αλλαγές στους πίνακες (δημιουργία και διαγραφή) τόσο σε απόλυτες τιμές όσο και ποσοστιαία. Επίσης εντόπισαν αλλαγές στα attributes που αφορούν Προσθήκη στήλης, Αλλαγή Τύπου, Αλλαγή στις αρχικοποιήσεις και Αλλαγή Κλειδιού. Επίσης οι αλλαγές στα attributes μελετήθηκαν σε απόλυτες τιμές και ποσοστιαία για κάθε εφαρμογή.

Όσον αφορά τις αλλαγές στους πίνακες οι συγγραφείς εντόπισαν ότι δημιουργήθηκαν συνολικά και στις 4 εφαρμογές ***21*** νέοι πίνακες και διαγράφηκαν ***51***, αποτελώντας το ***6,1%*** και το ***14,9%*** των συνολικών αλλαγών που έγιναν στη διάρκεια μελέτης των εφαρμογών.

Όσον αφορά τις αλλαγές που σχετίζονται με τα attributes η πιο συνηθισμένη αλλαγή ήταν η αλλαγή τύπου, ***83*** αλλαγές (όλες στην BiblioteQ) αποτελώντας το ***24,3%*** των συνολικών αλλαγών και στις 4 εφαρμογές που μελετήθηκαν. Ακολουθεί η διαγραφή στηλών με ***66*** αλλαγές (ποσοστό ***19,3%)***. Πραγματοποιήθηκαν μόνο *7* αλλαγές κλειδιού (ποσοστό ***2%*** επί των συνολικών αλλαγών) και μόνο ***3*** αλλαγές στις αρχικοποιήσεις των attributes (***0.9%)***

Τα ευρήματα της έρευνας εξετάσθηκαν ως προς τις ακόλουθες ***2*** κατευθύνσεις:

1. Φύση των αλλαγών – Ποιές αλλαγές του σχήματος είναι περισσότερο συχνές σε EDs σε σύγκριση με τις ΒΔ επιχειρήσεων;

Οι πιο συχνά εμφανιζόμενες αλλαγές είναι: *ADD COLUMN*, *DROP COLUMN*, *DROP TABLE και CREATE TABLE*. Η αλλαγή τύπου πεδίου εμφανίστηκε μόνο στη ΒΔ BiblioteQ (83 φορές) και η πλειοψηφία αφορούσε αλλαγή του LONGTEXT σε TEXT και INTEGER στο BIGINT. Οι συγγραφείς υπολόγισαν ποσοστά για τις πιο συχνές αλλαγές (αγνοώντας τις αλλαγές τύπου στο BiblioteQ):

Προσθήκη Στήλης	42,9%
Διαγραφή Στήλης	25,5%
Διαγραφή Πίνακα	19,7%
Δημιουργία Πίνακα	8,1%

Ο μεγάλος αριθμός διαγραφών σε Columns και Tables οδηγούν στο συμπέρασμα ότι οι ενσωματωμένες βάσεις δεδομένων που αναλύθηκαν τείνουν να υπο-

στούν περισσότερες αναδιαρθρώσεις, παρά να παρουσιάζουν συνεχή ανάπτυξη. Ο μικρός αριθμός αλλαγών τύπου στήλης (εκτός από το Biblioteq) οδηγούν στο συμπέρασμα ότι οι προσθήκες/διαγραφές είναι πιο σημαντικές από τις αλλαγές σε τύπους στηλών, αρχικοποιήσεις τιμών και αλλαγές σε κλειδιά αρχείων.
2. Συχνότητα και χρόνος των αλλαγών: πότε και πόσο συχνά, αλλάζουν τα σχήματα των ενσωματωμένων βάσεων δεδομένων

Στις fig. 2,3,4 του [WuNe11] σελ. 154 οι συγγραφείς παρουσιάζουν την συχνότητα των αλλαγών στην διάρκεια του κύκλου ζωής των Βάσεων Δεδομένων. Παρατηρήθηκε ότι τα σχήματα τείνουν να αλλάζουν περισσότερο στην αρχή του κύκλου ζωής τους και με την πάροδο του χρόνου η δομή σταθεροποιείται (οι μεταγενέστερες εκδόσεις έχουν λιγότερες αλλαγές). Αυτό υποδηλώνει ότι οι ενημερώσεις σχήματος on-the-fly είναι απαραίτητο, ειδικά στην αρχή της ζωής του προγράμματος.

3.2 Δυνατότητα Εφαρμογής των Νόμων του Lehman για την Εξέλιξη του Λογισμικού στην Περίπτωση των Σχημάτων ΒΔ.

Οι Νόμοι του ***Lehman*** αφορούν ένα σύνολο κανόνων που εισήχθησαν στα μέσα της δεκαετίας του ’70, γνωστοί και ως Νόμοι για την Εξέλιξη Λογισμικού.

Ο Meir M. Lehman και οι συνάδελφοί του, διατύπωσαν και στη συνέχεια τροποποίησαν, εμπλούτισαν και διόρθωσαν ένα σύνολο κανόνων σχετικά, με τη συμπεριφορά του λογισμικού καθώς εξελίσσεται με την πάροδο του χρόνου. Οι νόμοι του Lehman επικεντρώνονται στα συστήματα τύπου ***E*** (Το λογισμικό που επιλύει ένα πρόβλημα ή απευθύνεται σε μια εφαρμογή στο πραγματικό κόσμο).

Η κύρια ιδέα πίσω από τους νόμους της εξέλιξη για συστήματα λογισμικού τύπου Ε είναι ότι η *ανάπτυξή τους είναι μια διαδικασία ανατροφοδότησης*. Σε ένα σύστημα που βασίζεται στην ανατροφοδότηση, η διαδικασία εξέλιξης του, πρέπει να εξισορροπεί ανάμεσα (α) *στη θετική ανάδραση*, δηλαδή στην ανάγκη να προσαρμοστούν σε ένα μεταβαλλόμενο περιβάλλον και να εξελιχθούν έτσι ώστε να προσφέρουν μεγαλύτερη λειτουργικότητα και (β) στην *αρνητική ανάδραση*, δηλαδή στην ανάγκη ελέγχου, περιορισμού και άμεσης αλλαγής με τέτοιο

τρόπο ώστε να αποτρέπεται η υποβάθμιση της συντήρησης και της διαχείρισης του λογισμικού. Οι νόμοι συνοψίζονται ως εξής:

(I) *Νόμος Συνεχιζόμενης Αλλαγής*: Ένα σύστημα τύπου Ε πρέπει να προσαρμόζεται συνεχώς ή αλλιώς γίνεται προοδευτικά λιγότερο ικανοποιητικό κατά τη χρήση.

(II) *Νόμος αύξησης της πολυπλοκότητας*: Καθώς ένα σύστημα τύπου Ε αλλάζει, η πολυπλοκότητά του αυξάνει και γίνεται πιο δύσκολο να εξελιχθεί εκτός αν γίνει δουλειά για τη συντήρησή του ή τη μείωση της πολυπλοκότητας του.

(III) *Νόμος Αυτορρύθμισης*: Η εξέλιξη ενός συστήματος τύπου Ε βασίζεται στη διαδικασία της ανατροφοδότησης

(IV) *Νόμος διατήρησης της οργανωτικής σταθερότητας*: Το ποσοστό των αλλαγών που συμβαίνουν σε ένα σύστημα τείνει να είναι σταθερό στην διάρκεια ζωής αυτού του συστήματος ή σε κάποιες από τις φάσεις της ζωής του.

(V) *Νόμος διατήρησης της εξοικείωσης*: Γενικά, η τάση αύξησης περιορίζεται από την ανάγκη των συστημάτων να διατηρήσουν την εξοικείωση.

(VI) *Νόμος της Συνεχιζόμενης Ανάπτυξης*: η λειτουργικότητα των συστημάτων τύπου Ε πρέπει να βελτιώνονται συνεχώς, για να διατηρηθεί ο βαθμός της ικανοποίησης των χρηστών σε όλη την διάρκεια ζωής του συστήματος.

(VII) *Νόμος της Πτώσης Ποιότητας:* η ποιότητα ενός συστήματος τύπου Ε μειώνεται καθώς εξελίσσεται.

(VIII) *Νόμος ανατροφοδότησης του συστήματος:* Οι διαδικασίες εξέλιξης των συστημάτων βασίζονται στην ανατροφοδότηση.

Στις εργασίες [SkVZ14], [SkVZ15] πραγματοποιήθηκε μια διεξοδική μελέτη μεγάλης κλίμακας σχετικά με την εξέλιξη των βάσεων δεδομένων ανοιχτού κώδικα, που διατίθενται μέσω αποθετηρίων κώδικα και ερευνήθηκε η εγκυρότητα των νόμων του ***Lehman*** σε ιδιότητες όπως το μέγεθος, η ανάπτυξη και το ποσό αλλαγής ανά έκδοση.

Συγκεκριμένα, μελετήθηκε η εξέλιξη του σχήματος ***8*** βάσεων δεδομένων, που αποτελούν τμήματα δημόσιων έργων λογισμικού ανοιχτού κώδικα.

Συλλογή και επεξεργασία δεδομένων.

Για κάθε σύνολο δεδομένων που αφορούσε και τις 8 ΒΔ συγκεντρώθηκαν όσες εκδόσεις σχήματος (αρχεία) ήταν δυνατόν να εξαχθούν από τους δημόσιους χώ-

ρους αποθήκευσης πηγαίου κώδικα (cvs, svn, git). Προκειμένου να μεγιστοποιηθεί η εγκυρότητα των αποτελεσμάτων οι συγγραφείς επέλεξαν μόνο τις αλλαγές του τμήματος της βάσης δεδομένων του έργου όπως αυτές ενσωματώνονται στον κορμό του αποθετηρίου. Τα αρχεία συλλέχθηκαν τον *Ιούνιο 2013*. Τα αρχεία επεξεργάστηκαν από το εργαλείο ***Hecate***, που αναπτύχθηκε από τους συγγραφείς, και ανιχνεύει τις αλλαγές τόσο σε επίπεδο χαρακτηριστικών όσο και σχέσεων. Το συγκεκριμένο εργαλείο δίνει τις διαφορές μεταξύ δύο εκδόσεων και τον συνολικό αριθμό των αλλαγών (***παλμό***) για κάθε μετάβαση από μια έκδοση στην επόμενη.
Για την μελέτη υπολογίσθηκαν τα παρακάτω:
Μέγεθος σχήματος μιας έκδοσης: Ο αριθμός των πινάκων μιας έκδοσης σχήματος.
Ανάπτυξη Σχήματος: Η διαφορά του μέγεθος του σχήματος δύο (συνήθως συνεχόμενων) εκδόσεων (νέα - παλιά).
Heartbeat (Παλμός): Σύνολο, ένα ανά μετάβαση, των γεγονότων που συνέβησαν κατά τη διάρκεια αυτής της μετάβασης. Στο πλαίσιο του παρόντος εγγράφου, για κάθε μετάβαση μεταξύ δύο μεταγενέστερων εκδόσεων, παράγεται ένα σύνολο μετρήσεων συμπεριλαμβανομένων:

- Table Insertions - Εισαγωγές Πινάκων
- Table Deletions – Διαγραφές Πινάκων
- Attributes Insertions - Εισαγωγές Χαρακτηριστικών
- Attributes Deletions - Διαγραφές Χαρακτηριστικών
- Attributes Alternations - Αλλαγή τύπου δεδομένων
- Attributes Inserted at Table Formation - Εισαγωγές Χαρακτηριστικών στο σχηματισμό πίνακα
- Attributes deletions at Table Removal - Διαγραφές Χαρακτηριστικών στην Αφαίρεση πίνακα.

Attributes Insertions αφορούν σε έναν υπάρχοντα πίνακα *Attributes Inserted at Table Formation* αφορούν το αριθμός χαρακτηριστικών που παράγεται κάθε φορά που δημιουργείται ένας νέος πίνακας. *Attributes Deletions* αφορούν διαγραφές από πίνακα που εξακολουθεί να υπάρχει, ενώ *Attributes deletions at Table Removal* αφορούν τα χαρακτηριστικά που αφαιρούνται όποτε αφαιρείται ο πίνακας στον οποίο ανήκουν. Το σύνολο αυτών των μέτρων ανά μετάβαση, παράγει τον παλμό της διάρκειας ζωής του συνόλου δεδομένων.

Αξιολόγηση των νόμων για την εξέλιξη του σχήματος

Οι νόμοι της εξέλιξης του λογισμικού αναπτύχθηκαν και εξελίχθηκαν πάνω από ***40*** χρόνια. Οι συγγραφείς του [SkVZ15] ισχυρίζονται ότι εξηγώντας κάθε νόμο μεμονωμένα από τους άλλους τα αποτελέσματα είναι επισφαλή, καθώς κινδυνεύει να χαθεί η ουσία που προκύπτει από τις αλληλεξαρτήσεις των νόμων. Οι νόμοι οργανωθήκαν σε τρεις θεματικούς ομάδες που καλύπτουν το συνολικό μηχανισμό διαχείρισης της εξέλιξης:

- Η πρώτη ομάδα νόμων αναφέρεται στην ύπαρξη ενός μηχανισμού ανάδρασης που περιορίζει την ανεξέλεγκτη εξέλιξη του λογισμικού. (*Is There a Feedback-Based System for Schema Evolution?*)
- Η δεύτερη ομάδα διαπραγματεύεται τις ιδιότητες της *ανάπτυξης* του συστήματος, δηλαδή το μέρος του μηχανισμού εξέλιξης που θεωρείται θετική ανατροφοδότηση. *(Properties of Growth for Schema Evolution)*
- Η τρίτη ομάδα νόμων διαπραγματεύεται τις ιδιότητες που πρέπει να έχει η *συντήρηση* έτσι ώστε να περιορίζει την ανεξέλεγκτη ανάπτυξη, δηλαδή το τμήμα του μηχανισμού εξέλιξης που θεωρείται αρνητική ανατροφοδότηση. *(Perfective Maintenance for Schema Evolution).*

Στη συνέχεια αναλύονται τα αποτελέσματα για τις 3 αυτές ομάδες νόμων, όπως αυτές παρουσιάζονται στο [SkVZ14] και στο [SkVZ15]

3.2.1 Ισχύουν οι Νόμοι της Εξέλιξης Λογισμικού στην Εξέλιξη Σχήματος ([SkVZ15] και [SkVZ14])?

1η ομάδα νόμων - Υπάρχει μηχανισμός ανατροφοδότηση κατά την εξέλιξη του σχήματος;

Νόμος Ι - Νόμος συνεχούς αλλαγής

Ο πρώτος νόμος υποστηρίζει ότι το σύστημα αλλάζει συνεχώς με την πάροδο του χρόνου. Ένα σύστημα τύπου Ε πρέπει να προσαρμόζεται συνεχώς ή αλλιώς γίνεται προοδευτικά λιγότερο ικανοποιητικό κατά τη χρήση. Η βασική ιδέα πίσω από αυτόν τον νόμο είναι απλή: όπως το περιβάλλον του πραγματικού κόσμου εξελίσσεται, έτσι και το λογισμικό που προορίζεται να αντιμετωπίσει τα προβλήμα-

τά αυτού του κόσμου, πρέπει να ακλουθεί αυτή την εξέλιξη. Αν αυτό δεν συμβεί, το σύστημα καθίσταται λιγότερο ικανοποιητικό.

Για να ισχύει ο νόμος, πρέπει να δείξουμε ότι το λογισμικό εμφανίζει σημάδια της εξέλιξης καθώς περνά ο καιρός. Πιθανές μετρήσεις από το πεδίο της τεχνολογίας λογισμικού περιλαμβάνουν ***(α)*** το συνόλου του αριθμού των αλλαγών και ***(β)*** την ανάλυση των αλλαγών στο πέρασμα του χρόνου.

Για να επικυρωθεί η υπόθεση ότι ο νόμος της συνεχούς αλλαγής ισχύει, μελετήθηκε ο παλμός της ζωής του σχήματος σε συνδυασμό με το μέγεθος του σχήματος. (fig 2 - 4 του [SkVZ15]–σελ. 368-270). Από την οπτική θεώρηση των δεδομένων αλλαγής-χρόνου, εξάγεται το συμπέρασμα ότι ο όρος ***συνεχώς*** στον ορισμό του νόμου αμφισβητείται: παρατηρήθηκε ότι η εξέλιξη του σχήματος της βάσης δεδομένων συμβαίνει σε ***ριπές***, κατά περιόδους της εξελικτικής δραστηριότητας, και όχι ως μια συνεχής διαδικασία.

Οι *εκδόσεις με μηδενικές αλλαγές* είναι εκδόσεις στις οποίες δεν γίνονται αλλαγές στο σχήμα (σχέσεις, ιδιότητες και περιορισμοί), αλλά αντιθέτως, αφορούν πτυχές απόδοσης της βάσης δεδομένων (ευρετήρια, μηχανές αποθήκευσης κ.λπ.).

Γενικά, αν επιμένουμε στην ακριβή διατύπωση του νόμου, συμπεραίνεται ότι ο νόμος ισχύει εν μέρει.

Νόμος III - Νόμος αυτορρύθμισης

Ο τρίτος νόμος για την εξέλιξη του λογισμικού είναι γνωστός ως νόμος της "Αυτορρύθμισης" δηλαδή η εξέλιξη ενός συστήματος τύπου E ρυθμίζεται από την ανατροφοδότηση.

Καθώς οι χρήστες του συστήματος απαιτούν περισσότερη λειτουργικότητα, το σύστημα μεγαλώνει σε μέγεθος για να ανταποκριθεί σε αυτή την ζήτηση αλλά ταυτόχρονα πραγματοποιείται διορθωτική και αποδοτική συντήρηση, ώστε να αφαιρεθούν τα σφάλματα και να βελτιωθεί η ποιότητα του λογισμικού (μειωμένη πολυπλοκότητα, αυξημένη κατανόηση). Έτσι, η ανάπτυξη του συστήματος δεν μπορεί να γίνεται συνεχώς με τον ίδιο ρυθμό αλλά αντίθετα, αυτό που αναμένεται είναι μια βασική ανάπτυξη, που διακόπτεται με releases άριστης συντήρησης.

Μετρήσεις για την ισχύ του Νόμου

Οι κυματισμοί στο μέγεθος του συστήματος θεωρείται ότι υποδηλώνουν την ύπαρξη ανατροφοδότησης στο σύστημα: θετική ανατροφοδότηση είναι η επέκταση του συστήματος με την προσθήκη στοιχείων π.χ. πινάκων και πεδίων. Αρνητική

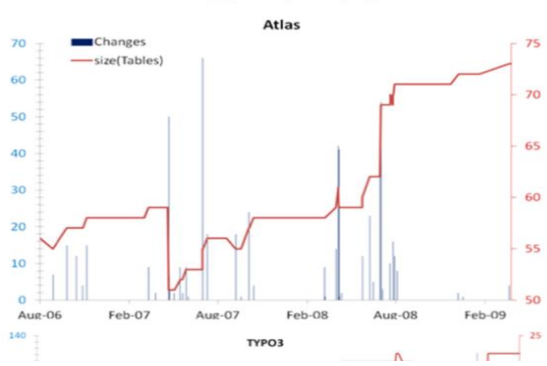
ανατροφοδότηση σημαίνει συντήρηση με μειωμένο ρυθμό ανάπτυξη η οποία δεν οφείλεται στη λειτουργική ανάπτυξη της εφαρμογής αλλά στην προσπάθεια για καλύτερη ποιότητα κώδικα π.χ. συρρίκνωση του συστήματος λόγω αφαίρεσης περιττών στοιχείων ή τη συγχώνευσή τους με άλλα).

Σχετικά με το μέγεθος παρατηρήθηκαν οι παρακάτω περίοδοι:

Περίοδοι αύξησης, στην αρχή της ζωής ή μετά από μια μεγάλη μείωση του μεγέθους του σχήματος. Αυτό είναι ένδειξη θετικής ανατροφοδότησης, δηλαδή της ανάγκης επέκτασης του σχήματος για την κάλυψη των αναγκών των χρηστών.

Μειώσεις στο μέγεθος του σχήματος. Αυτές οι μειώσεις είναι συνήθως αιφνίδιες και απότομες και συνήθως πραγματοποιούνται σε σύντομες χρονικές περιόδους. Στην πραγματικότητα, αυτές οι μειώσεις έχουν σημαντικά μεγαλύτερο μέγεθος από μια τυπική αλλαγή και δείχνουν την ύπαρξη αρνητικού μηχανισμό ανατροφοδότησης στη διαδικασία εξέλιξης.

Περίοδοι ηρεμίας δηλαδή περίοδοι μη τροποποίησης της λογικής δομής του σχήματος. Αυτό είναι ιδιαίτερα εμφανές αν παρατηρήσει κανείς τον παλμό, όπου οι αλλαγές ομαδοποιούνται σε πολύ συγκεκριμένες στιγμές. (fig.3 του [SkVZ15] – σελ. 369)

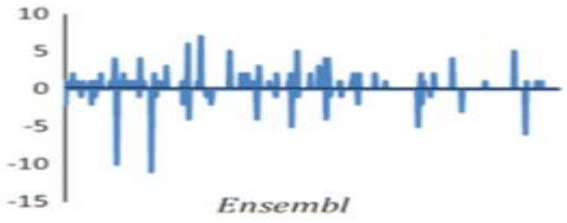


Σχήμα 3-1 : Combined demonstration of heartbeat. - από το SkVZ15 με άδεια των συγγραφέων – σελ 369

Η ανάπτυξη και οι ταλαντώσεις της.

Η Ανάπτυξη (δηλαδή, η διαφορά στο μέγεθος μεταξύ δύο συνεχόμενων εκδόσεων) παρουσίασε κοινά χαρακτηριστικά σε όλα τα σύνολα δεδομένων. Στις περισσότερες περιπτώσεις, η ανάπτυξη είναι μικρή (συνήθως κυμαίνεται μεταξύ 0 και 1). Λόγω της συντήρησης που γίνεται για καλύτερη ποιότητα κώδικα, υπάρχουν

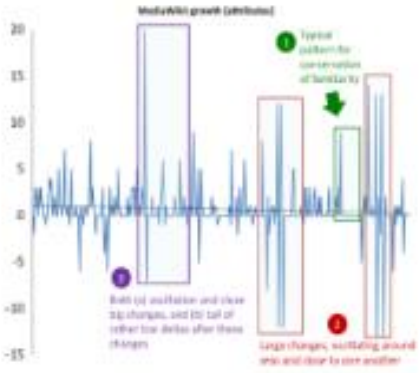
και αρνητικές τιμές ανάπτυξης (λιγότερες από τις θετικές). (fig. 5 του [SkVZ15] – σελ. 371).



Σχήμα 3-2 : Growth (tables) over version id for all the datasets- από το SkVZ15 με άδεια των συγγραφέων – σελ 371

Δεν παρατηρήθηκε σταθερή ροή εκδόσεων που το μέγεθος του σχήματος να αλλάζει συνεχώς, αλλά μικρές αιχμές μεταξύ ενός και μηδέν. Έτσι, προκύπτει ότι η ανάπτυξη έρχεται με ένα μοτίβο ακίδων. Εξαιτίας αυτού του χαρακτηριστικού, η μέση τιμή είναι συνήθως πολύ κοντά στο μηδέν (στη θετική πλευρά) σε όλα τα σύνολα δεδομένων, τόσο για πίνακες όσο και για πεδία.

Οι ταλαντώσεις της ανάπτυξης, ειδικά στο επίπεδο των χαρακτηριστικών, δείχνουν ότι είναι αρκετά συνηθισμένο, να εμφανιστούν *ακολουθίες ταλαντώσεων μεγάλου μεγέθους*: δηλαδή υπερβολικά θετικές ακολουθίες ακολουθούμενες αμέσως από υπερβολικά αρνητικές ακολουθίες. (fig.9 του [SkVZ15]– σελ. 378)

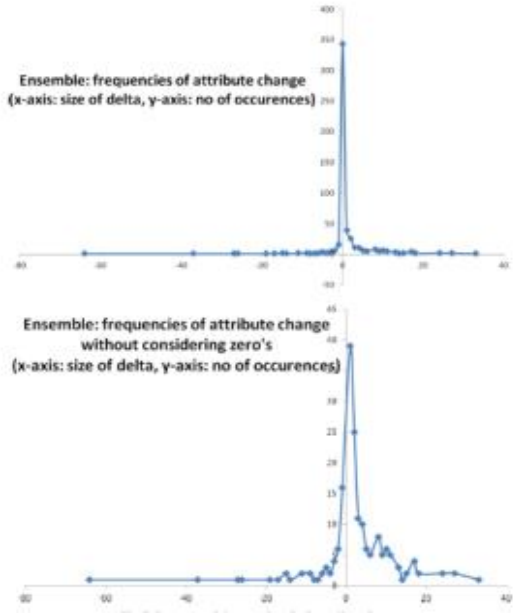


Σχήμα 3-3 : Different patterns of change in attribute growth of MediaWiki - από το SkVZ15 με άδεια των συγγραφέων – σελ 378

- Όσον αφορά τους πίνακες, η ανάπτυξη ως επί το πλείστον περιορίζεται σε μικρές τιμές. Όλοι οι αριθμοί βρίσκονται στην περιοχή [-2..2] και κυρίως στο εύρος [0..2]. Λίγες απότομες αλλαγές συμβαίνουν.

- Όσον αφορά τα χαρακτηριστικά οι αριθμοί είναι υψηλότεροι, και εξαρτώνται

Παρατηρήθηκε ένα πρότυπο κοινό σε όλα τα σύνολα δεδομένων: υπάρχει ένα μοντέλο Zipfian για τη διανομή συχνοτήτων. Φαίνεται στην εικόνα 7 του [SkVZ15] – σελ 374, όπου απεικονίζεται πόσο συχνά μια τιμή ανάπτυξης εμφανίζεται στα attribute του Ensemble.



Σχήμα 3-4 : Frequency of change values for Ensembl attributes - από το SkVZ15 με άδεια των συγγραφέων – σελ 374

Στην πραγματικότητα, δύο φαινόμενα παρατηρούνται: (α) η μικρή αλλαγή γύρω από το μηδέν, (β) Zipfian διανομή συχνοτήτων σε όλα τα σύνολα δεδομένων χωρίς εξαιρέσεις.

Παρά το γεγονός ότι η αλλαγή δεν ακολουθεί το πρότυπο της βασικής ομαλής ανάπτυξης του Lehman και το γεγονός ότι η αλλαγή υπακούει σε κατανομή Zipfian με κορυφή στο μηδέν, οι συγγραφείς καταλήγουν ότι η παρουσία ανατροφοδότησης στη διαδικασία εξέλιξης είναι σαφής. Συνεπώς *ισχύει ο νόμος.*

Νόμος VIII - Νόμος ανατροφοδότησης του συστήματος

Οι συγγραφείς επικεντρωθήκαν στο κατά πόσο πράγματι υπάρχει μηχανισμός που σταθεροποιεί την τάση για αδιάλειπτη λειτουργία ανάπτυξης του σχήματος. Για να είναι έγκυρος ο νόμος, πρέπει να διαπιστωθεί αν είναι δυνατόν να προσομοιωθεί η εξέλιξη του μεγέθους του σχήματος μέσω ενός ακριβή τύπου.

Εκτελείται παλινδρομική ανάλυση για την εκτίμηση του αριθμού των σχέσεων για κάθε έκδοση του σχήματος. Εντοπίζονται τύποι σχετικά με τη σχέση του νέου μεγέθους του συστήματος ως συνάρτηση του προηγούμενου μεγέθους του.

E: παράμετρος μοντέλου που προσεγγίζει την προσπάθεια. (μέση τιμή του E_i)

E_i: για κάθε transition

Κάνουν χρήση τύπων που χρησιμοποιούν παραμέτρους

s_i: αναφέρεται στο πραγματικό μέγεθος του σχήματος στην έκδοση *i*

α: αναφέρεται στην έκδοση από την οποία αρχίζει η καταμέτρηση

και η ουσία του τύπου είναι ότι, για να υπολογίσουμε το ***E_i***, χρησιμοποιούν ***α*** προηγούμενες εκδόσεις για την εκτίμηση της προσπάθειας.

Η αξιολόγηση ξεκινά με τον υπολογισμό του μέσου ***E*** από τα ατομικό ***E_i*** σε ολόκληρο το σύνολο δεδομένων με 4 διαφορετικές τιμές για το ***α***, συγκεκριμένα:

- ***i -1*** (τελευταία έκδοση)
- ***1*** για όλες τις εκδόσεις
- ***5*** πρόσφατες εκδόσεις
- ***10*** πρόσφατες εκδόσεις

Τα αποτελέσματα δείχνουν ότι η προσέγγιση έχει μέτρια επιτυχία στο να προβλέψει μια συνολική τάση αύξησης και για τις τέσσερις περιπτώσεις και στην πραγματικότητα και οι 4 προσεγγίσεις στοχεύουν στο να προβλέψουν μια αυξανόμενη τάση που δεν επιβεβαιώνεται στο πραγματικό σχήμα. Συγχρόνως, και οι τέσσερις προσεγγίσεις αποτυγχάνουν στο να καταγράψουν τις μεμονωμένες διακυμάνσεις στη διάρκεια ζωής του σχήματος.

Με την προσέγγιση ότι η παράμετρος ***E*** δεν ήταν σταθερή καθ ’όλη τη διάρκεια ζωής του έργου (διότι το έργο χωρίστηκε σε φάσεις) και έτσι για κάθε εκδοχή *i*, υπολογίστηκε το ***E*** ως μέσο των τελευταίων ***τ*** *Ej* (τ = 1/5/10) - σε αντίθεση με προηγούμενες προσπάθειες όπου το ***E*** υπολογίστηκε ως συνολικός μέσος όρος για ολόκληρο το σύνολο δεδομένων (δηλαδή σταθερό για όλους εκδόσεις) ή ως ένας συνολικός μέσος όρος από την αρχή του έργου μέχρι τις τρέχουσες εκδόσεις.

Επίσης αποφασίστηκε από τους συγγραφείς να χρησιμοποιηθούν οι τελευταίες 5 ή 10 εκδόσεις ***E_i***, δηλαδή, ***α*** ίσον με 5 ή 10. Αυτό έχει ήδη χρησιμοποιηθεί σε προηγούμενα πειράματα.

Η ιδέα του υπολογισμού του μέσου όρου Ε με σύντομη μνήμη 5 ή 10 παραλλαγών, παράγει εξαιρετικά ακριβή αποτελέσματα. Αυτό ισχύει για όλα τα σύνολα δεδομένων. (Fig 8 του [SkVZ15] – σελ. 376)

Συνολικά, εμφανίζεται η εξέλιξη του σχήματος βάσης δεδομένων να υπακούει στη συμπεριφορά ενός μηχανισμού που βασίζεται σε ανατροφοδότηση. Το μέγεθος του σχήματος μιας συγκεκριμένης έκδοσης της βάσης δεδομένων μπορεί να εκτιμάται με ακρίβεια μέσω ενός παλινδρομικού τύπου που εκμεταλλεύεται το ποσό των αλλαγών στις πρόσφατες, προηγούμενες εκδόσεις.

2η ομάδα νόμων - Ιδιότητες ανάπτυξης για την εξέλιξη του σχήματος

Νόμος VI - Νόμος συνεχούς ανάπτυξης

Ο έκτος νόμος μοιάζει με τον πρώτο νόμο (συνεχής αλλαγή) με μια πρώτη ματιά. Όμως καλύπτουν διαφορετικά φαινόμενα. Ο πρώτος νόμος αναφέρεται στην ανάγκη ενός συστήματος λογισμικού να προσαρμοστεί σε ένα μεταβαλλόμενο κόσμο. Ο έκτος νόμος αναφέρεται στο γεγονός ότι το σύστημα δεν μπορεί να περιλαμβάνει όλες τις απαραίτητες λειτουργίες σε μια ενιαία έκδοση καθώς, λόγω έλλειψης χρόνου και πόρων, αρκετές επιθυμητές λειτουργίες του συστήματος αποκλείονται από μια έκδοση. Με το πέρασμα του χρόνου, αυτές οι λειτουργίες αναμειγνύονται στο σύστημα με τις νέες απαιτήσεις που απορρέουν από τον πρώτο νόμο στο πλαίσιο ενός εξελισσόμενου κόσμου.

Οι μετρήσεις για την εκτίμηση του Νόμου προέρχονται από το λογισμικό, συμπεριλαμβανομένου του LOC (Lines Of Code), του αριθμού των ορισμών (τύπων, λειτουργιών και συνολικών μεταβλητών) και του αριθμού των modules. Ωστόσο, οι συγγραφείς αναγνωρίζουν ότι οι μετρήσεις για την εκτίμηση του νόμου παρουσιάζουν αρκετές δυσκολίες, και χρησιμοποιούν *το μέγεθος του σχήματος* ως ασφαλές μέτρο παρατηρώντας "προσθήκες στο υπάρχον σύστημα".

Αξιολόγηση: ‘Ολα τα σύνολα δεδομένων δείχνουν την τάση να αυξηθούν με την πάροδο του χρόνου. Η αλλαγή όμως έρχεται μετά από συχνές (και μερικές φορές μεγάλες) περιόδους ηρεμίας, όπου το μέγεθος του σχήματος δεν αλλάζει (ή αλλάζει ελάχιστα). Ηρεμία είναι ένα φαινόμενο που δεν συναντήθηκε στη μελέτη του Lehman πάνω στα παραδοσιακά συστήματα λογισμικού και αποκτά επιπλέον σημασία, εάν θεωρηθεί ότι στην παρούσα μελέτη οι συγγραφείς έχουν επικε-

ντρωθεί μόνο στο σχήμα και άρα υπάρχουν εκδόσεις του συστήματος, για τις οποίες το σχήμα παραμένει σταθερό ενώ ο κώδικας άλλαξε.

Ως εκ τούτου μπορούμε να συμπεράνουμε ότι ο νόμος ισχύει (η χωρητικότητα του σχήματος της βάσης δεδομένων βελτιώνεται μακροπρόθεσμα), αν και τροποποιήθηκε για να ικανοποιήσει τις ιδιαιτερότητες των σχημάτων βάσης δεδομένων (οι αλλαγές δεν είναι συνεχείς αλλά, έρχονται μετά από μεγάλες περιόδους ηρεμίας).

Νόμος V - Νόμος διατήρησης της εξοικείωσης

Ο πέμπτος νόμος για την εξέλιξη του λογισμικού είναι γνωστός ως νόμος της "Διατήρησης της εξοικείωσης". Γενικά, η αυξητική τάση (τάση αναλογίας ανάπτυξης) των συστημάτων τύπου Ε περιορίζεται από την ανάγκη της εξοικείωσης. Καθώς το σύστημα εξελίσσεται, όσοι σχετίζονται με αυτό (προγραμματιστές, χρήστες, διαχειριστές κ.λπ.) προσπαθούν να καταλάβουν και στην πραγματικότητα, να ελέγξουν το περιεχόμενο και τη λειτουργικότητα του. Κάθε φορά που υπάρχει υπερβολική ανάπτυξη σε μια έκδοση, ο μηχανισμός ανάδρασης τείνει να μειώσει τη ανάπτυξη σε μεταγενέστερες εκδόσεις, έτσι ώστε η αλλαγή να μπορεί να αφομοιωθεί από τους ανθρώπους. Μέσα από έρευνες, φαίνεται ότι τις εκδόσεις με μεγάλο όγκο αλλαγών, ακολουθούν εκδόσεις που εκτελούν διορθωτική ή τελειοποιημένη συντήρηση.

Οι συγγραφείς αξιολογούν τον νόμο μέσω μετρήσεων όπως (i) η ανάπτυξη του συστήματος, (ii) ο ρυθμός ανάπτυξης του συστήματος και (iii) ο αριθμό των αλλαγών σε κάθε έκδοση και αξιολογούν την ισχύ του ελέγχοντας αν:

- Η ανάπτυξη του σχήματος δεν αυξάνεται με την πάροδο του χρόνου. Στην πραγματικότητα, είναι -στην καλύτερη περίπτωση- σταθερή ή μειώνεται με την πάροδο του χρόνου/έκδοση.
- Τι συμβαίνει μετά από υπερβολικές αλλαγές. Παρατηρήθηκαν μικρές διακυμάνσεις της αλλαγής, που δείχνουν την απορρόφηση των επιπτώσεων της αλλαγής όσον αφορά τη διορθωτική συντήρηση και την γνωριμία του προγραμματιστή με τη νέα έκδοση του σχήματος.

Αξιολόγηση του Νόμου:

Όσον αφορά την επίδραση της ηλικίας, δεν εντοπίστηκε μια μειωμένη τάση της ανάπτυξης. Ωστόσο, η ηλικία οδηγεί σε μείωση της πυκνότητας των αλλαγών

και της συχνότητας των μη-μηδενικών τιμών στις αιχμές. Αυτό εξηγεί την μείωση της ανάπτυξης σε όλα σχεδόν τα σύνολα δεδομένων που ερευνηθήκαν.

Ο παλμός των συστημάτων δείχνει ότι η αλλαγή είναι πολύ συχνότερη στην αρχή, παρά το γεγονός ότι μεγάλες αλλαγές μπορεί να εμφανιστούν σε οποιαδήποτε περίοδο της ζωής του σχήματος. (fig. 2 -4 του [SkVZ15] – σελ. 368-370)

Όσον αφορά την εγκυρότητα του νόμου, οι συγγραφείς πιστεύουν ότι είναι δυνατή αλλά δεν επιβεβαιώνεται. Ο νόμος αναφέρει ότι η ανάπτυξη περιορίζεται από την ανάγκη διατήρησης της εξοικείωσης, υπάρχουν όμως και άλλοι σοβαροί λόγοι για να περιοριστεί η ανάπτυξη, όπως ο υψηλός βαθμός εξάρτησης άλλων κομματιών του συστήματος από τη βάση δεδομένων. Συνεπώς, η διατήρηση της εξοικείωσης αν και σημαντική, δεν μπορεί να δικαιολογηθεί μόνο από την περιορισμένη ανάπτυξη.

Νόμος IV - Νόμος Διατήρησης της Οργανωτικής Σταθερότητας

Ο τέταρτος νόμος είναι επίσης γνωστός και ως νόμος του "αμετάβλητου ρυθμού εργασίας". Το ποσοστό εργασίας που καταβάλλεται για την ανάπτυξη ενός συστήματος, τείνει να είναι σταθερό στην διάρκεια ζωής αυτού του συστήματος ή των φάσεων αυτής της διάρκειας ζωής.

Οι πιθανές μετρήσεις για την εκτίμηση του νόμου περιλαμβάνουν (i) τον αριθμό των αλλαγών ανά έκδοση, (ii) το μέσο αριθμό αλλαγών ανά ημέρα, και (iii) το λόγο μεταβολής και ανάπτυξης.

Για την επικύρωση του νόμου, πρέπει να διαπιστωθεί ότι η διάρκεια ζωής του έργου είναι διαιρούμενη σε φάσεις, καθεμιά από τις οποίες (α) καταδεικνύει σταθερή ανάπτυξη και (β) συνδέεται με την επόμενη φάση με μια απότομη αλλαγή. Οι απότομες αλλαγές, όμως, συμβαίνουν κατά διαστήματα και όχι καθ ᾽όλη τη διάρκεια του χρόνου (σε εξαιρετικά σύντομες φάσεις).

Οι παλμοί (fig. 2-4 του [SkVZ15] – σελ 368-370) με την αυθαίρετη αλληλουχία των αιχμών και της ηρεμίας (fig. 5, 9 του [SkVZ15] -σελ 371, 378) καθιστούν αδύνατο να επιβεβαιωθεί σταθερότητα ανάπτυξης, ακόμη και σε φάσεις και άρα μπορεί να ειπωθεί *με ασφάλεια ότι δεν ισχύει ο νόμος.*

3η ομάδα νόμων - 2η ομάδα νόμων - Ιδιότητες ανάπτυξης για την εξέλιξη του σχήματος

Ο Lehman έδειξε τη μάχη μεταξύ δύο ανταγωνιστικών διεργασιών:

1. ενός καθορισμένου ποσού πόρων για τη συντήρηση του λογισμικού (ανάγκη ανάπτυξης του συστήματος)
2. προσπάθειας να μειωθεί η αυξανόμενη πολυπλοκότητα του συστήματος (για να επιτευχθεί αυτό πρέπει να εκτελείται από καιρό σε καιρό συντήρηση, προκειμένου να απομακρυνθεί ο περιττός κώδικας, να αναδιοργανωθεί ο κώδικας για καλύτερη συντήρηση και κατανόηση, να τεκμηριωθεί κλπ)

Νόμος ΙΙ - Νόμος αύξησης της πολυπλοκότητας

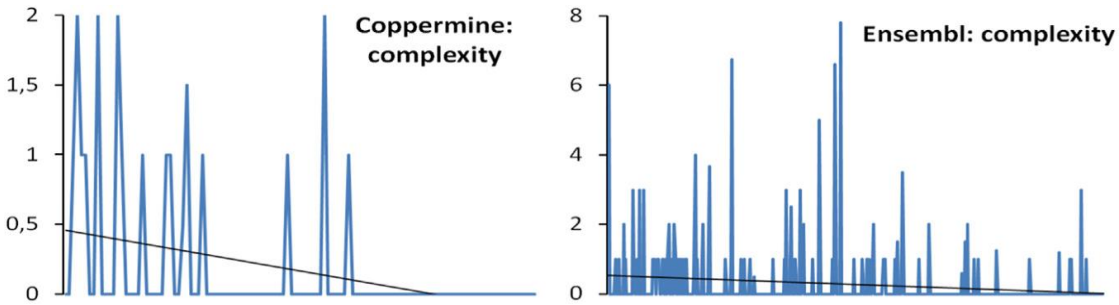
Καθώς αλλάζει το σύστημα τύπου Ε, η πολυπλοκότητά του αυξάνεται και γίνεται πιο δύσκολο να εξελιχθεί, αν δεν γίνει δουλειά για τη διατήρηση ή τη μείωση της πολυπλοκότητας. Ο νόμος δηλώνει ότι η πολυπλοκότητα αυξάνεται με την ηλικία, εκτός εάν καταβάλλεται προσπάθεια να αποφευχθεί αυτό.

Επειδή ο ορισμός της έννοιας της πολυπλοκότητας είναι δύσκολό να προσδιοριστεί αξιολογείται η ισχύς του νόμου με βάση τον συνδυασμό των ακόλουθων παρατηρήσεων:

1. Ανίχνευση των *εκδόσεων με πτώσεις (drops) στο μέγεθος και την ανάπτυξη του συστήματος*. Η γενική τάση του συστήματος είναι να αναπτυχθεί, άρα η ύπαρξη πτώσεων από καιρό σε καιρό θα δώσει μια ισχυρή ένδειξη ισχύος του νόμου.
2. Η ισχύς του νόμου θα σήμαινε έντονα την ύπαρξη ενός συστήματος βασισμένου σε ανατροφοδότηση και κατά συνέπεια, την ύπαρξη αρνητικής ανάδρασης, όπως συζητήθηκε στο δεύτερο νόμο της εξέλιξης.
3. Η προσέγγιση της μέτρηση της πολυπλοκότητας (από τον Lehman) ορίζεται ως το κλάσμα των σχέσεων εξέλιξης-επηρεαζόμενων (δηλαδή, ο αριθμός των σχέσεων που έχουν τροποποιηθεί ή προστεθεί στο σχήμα μεταξύ δύο διαδοχικών εκδόσεις του σχήματος) προς τη διαφορά του αριθμού των σχέσεων των εμπλεκόμενων εκδόσεων. Αυτός ο τύπος προσεγγίζει την προσπάθεια που έχει γίνει για την επέκταση συστήματος σε σχέση με την πραγματική διαφορά που επιτυγχάνεται (μεγάλη τιμή κλάσματος καταδεικνύει υπερβολική προσπάθεια για πολύ μικρές αλλαγές). Για καλύτερη κατανόηση, μπορεί να γίνει η υπόθεση ότι συγκρίνονται δύο μεταβάσεις με τον ίδιο παρονομαστή (δηλαδή, διαφορά στο αριθμό σχέσεων). Αν μια μετάβαση είχε περισσότερες σχέσεις από μια άλλη, σημαίνει ότι καταβάλλονται περισσότερες προσπάθειες γι ’αυτή τη μετάβαση και έτσι, υποτίθεται ότι η αρχική πολυπλοκότητα είναι υψηλότερη. Πιο συγκεκριμένα, διαιρείται η προσπά-

θεια (αριθμός των σχέσεις που τροποποιήθηκαν με οποιονδήποτε τρόπο σε μια αναθεώρηση) με την ανάπτυξη (μέγεθος του αποτελέσματος αυτής της αναθεώρησης). Σε περίπτωση που ο παρονομαστής είναι μηδέν, ορίζεται η πολυπλοκότητα ως μηδέν.

Στην παρούσα μελέτη, *η πολυπλοκότητα, όπως ορίζεται στο προηγούμενη παράγραφο, δεν αυξάνεται* (γραμμική παρεμβολή του σχήματος - fig. 10 του [SkVZ15] -σελ. 380).



Σχήμα 3-5 : Complexity for Coppermineand Ensembl - από το SkVZ15 με άδεια των συγγραφέων – σελ 380

Το φαινόμενο πρέπει να συνδυαστεί με την πτώση της αλλαγής πυκνότητας (νόμος V) με τη συνέργεια δύο αιτιών: (α) αύξηση της εξάρτησης του κώδικα από την ΒΔ, η οποία κάνει τους προγραμματιστές πιο προσεκτικούς στις αλλαγές του σχήματος καθώς αυτό συνεπάγεται υψηλότερο κόστος συντήρησης και (β) η συντήρηση, η οποία έχει ως αποτέλεσμα ένα καθαρό σχήμα, απαιτώντας λιγότερη διορθωτική συντήρηση στο μέλλον.

Η ισχύς του νόμου δεν μπορεί να επιβεβαιωθεί ή να διαψευσθεί βάσει αντικειμενικών μετρήσεων, υπάρχουν όμως ενδείξεις ότι ο νόμος ισχύει εν μέρει, αν και για διαφορετικούς λόγους από αυτούς που ανέφερε ο Lehman στα τυπικά συστήματα λογισμικού.

Νόμος VII - Νόμος μείωσης της ποιότητας

Η ποιότητα ενός συστήματος τύπου Ε, φαίνεται να μειώνεται καθώς εξελίσσεται, προσπαθώντας να ενσωματώσει τις λειτουργικές αλλαγές ώστε να καλύπτονται οι ανάγκες των χρηστών. Καθώς οι ανάγκες ακυρώνονται, η γήρανση του συστήματος και η αύξηση της πολυπλοκότητας, οδηγούν στο να πραγματοποιηθούν ενέρ-

γειες ώστε να συντηρηθούν τα τμήματα του λογισμικού που επηρεάζονται, για να αντικατοπτρίζουν πλέον, τις πραγματικές ανάγκες των χρηστών.

Μετρήσεις για την εκτίμηση του Νόμου: Πιθανές μετρήσεις για την εσωτερική ποιότητα του λογισμικού συστήματος περιλαμβάνουν (i) τον αριθμό των γνωστών προβλημάτων σε κάθε version, (ii) πυκνότητα προβλημάτων για κάθε version (iii) ποσοστό των ενοτήτων των οποίων το σώμα έχει αλλάξει

Η μείωση της ποιότητας του λογισμικού με την ηλικία εμφανίζεται να σχετίζεται με μια αύξηση της πολυπλοκότητας που πρέπει να συσχετιστεί με τη γήρανση όμως αυτή η αύξηση της πολυπλοκότητας δεν υποστηρίζεται από τις παρατηρήσεις των συγγραφέων. Επίσης δεν μπορεί να αξιολογηθεί η ποιότητα του σχήματος με αδιαμφισβήτητες μετρήσεις. Επομένως, ο νόμος δεν μπορεί να επιβεβαιωθεί ή να διαψευστεί.

3.3 Μια τυπική μελέτη περίπτωσης για την κατανόηση της εξέλιξης του σχήματος μιας βάσης δεδομένων

Στην μελέτη [CGMM13] οι συγγραφείς παρουσιάζουν μια μέθοδο που υποστηρίζεται από εργαλεία για την ανάλυση του ιστορικού ενός σχήματος βάσης δεδομένων, χρησιμοποιεί τεχνικές εξόρυξης λογισμικού (MSR) και εφαρμόζουν τη μέθοδο σε μια μελέτη περίπτωσης μεγάλης κλίμακας.

Σε αυτό το άρθρο, αναφέρονται οι εμπειρίες από την μελέτη του σχήματος ενός πολύπλοκου συστήματος ιατρικών πληροφοριών το ***OSCAR*** το οποίο είναι ένα πλήρες λειτουργικό σύστημα ηλεκτρονικής ιατρικής εγγραφής (EMR) σε κλινικές. Έχει αναπτυχθεί ως open source project και διαθέτει μια ευρεία και ενεργή κοινότητα χρηστών και προγραμματιστών.

Χαρακτηριστικά του Πληροφοριακού Συστήματος Oscar:

- Ο πηγαίος κώδικας περιλαμβάνει περίπου δύο εκατομμύρια γραμμές κώδικα.
- Χρησιμοποιεί MySQL ως μηχανή σχεσιακής βάσης δεδομένων και έναν συνδυασμό διαφορετικών τρόπων πρόσβασης σε αυτήν (Hibernate, την αρχιτεκτονική Persistence Java (JPA) και τη δυναμική SQL (μέσω του JDBC).
- Το σχήμα βάσης δεδομένων OSCAR έχει πάνω από ***440*** πίνακες και πολλές χιλιάδες ιδιότητες.

- Η κοινότητα προγραμματιστών του OSCAR χρησιμοποιεί μια σειρά από αποθετήρια λογισμικού και εργαλεία, συμπεριλαμβανομένου ενός συστήματος εντοπισμού αιτήσεων και εντοπισμού σφαλμάτων (που παρέχεται από το Sourceforge), ενός συστήματος υποβολής και ανασκόπησης πηγαίου κώδικα (κώδικας Gerrit Review), ενός συστήματος διαχείρισης βασισμένο σε git, ένα κοινό Wiki (με βάση το Plone) και τρεις λίστες αλληλογραφίας (μία για προγραμματιστές και δύο για χρήστες διαφορετικών επιπέδων τεχνικής εμπειρίας).
- Το σχήμα βάσης δεδομένων OSCAR προέκυψε από τη συμμετοχή των συγγραφέων σε ένα έργο με στόχο την ανάπτυξη λογισμικού για ένα δίκτυο έρευνας πρωτοβάθμιας περίθαλψης (PCRN).

Προβλήματα κατά την έρευνα:

- στο σχήμα βάσης δεδομένων δεν υπήρχε τεκμηρίωση.
- Στο σχήμα δεν έχουν δηλωθεί οι σχέσεις (ξένα κλειδιά).
- Στο σχήμα εντοπίστηκαν δύο φαινομενικά άσχετες δομές σχήματος που καλύπτουν το ίδιο σημασιολογικό ζήτημα. Διαπιστώθηκε ότι η μία έχει αντικατασταθεί, συνεχίζει όμως να υπάρχει για διαχείριση ιστορικών δεδομένων.

Λόγω των παραπάνω, διαπιστώθηκε ότι η ανάκτηση μιας ακριβούς γνώσης της ιστορίας της εξέλιξης του σχήματος ήταν μια σημαντική προϋπόθεση για την κατανόηση της βάσης δεδομένων OSCAR. Αρχικά θα έπρεπε να εντοπιστούν οι *βασικοί* πίνακες του σχήματος OSCAR – για να ξεκινήσει η μετάβαση από τη βάση δεδομένων συναλλαγών OSCAR στη μη σχεσιακή ερευνητική βάση δεδομένων. Θα μπορούσε επίσης να επιτρέψει την ανακάλυψη "*νεκρών* " πινάκων στο σχήμα (που δεν χρησιμοποιούνται πια, αλλά διατηρούνται για συμβατότητα προς τα πίσω και για ιατρικούς λόγους)

Η αντίστροφη μηχανική επί της βάσης δεδομένων (Database reverse engineering - DRE) πραγματοποιήθηκε με την εξέταση τριών κύριων πηγών πληροφοριών: (1) το σχήμα βάσης δεδομένων, (2) τα αποθηκευμένα δεδομένα, και (3) τα προγράμματα εφαρμογών.

Η διαδικασία για την κατασκευή του ιστορικού του σχήματος βάσης δεδομένων

ενός συστήματος αποτελείται από διάφορα βήματα:

1. *Εξαγωγή κώδικα SQL*: Εξαγωγή όλων των αρχείων SQL που αντιστοιχούν σε κάθε έκδοση του συστήματος.

2. *Εξαγωγή σχήματος*: Εξαγωγή του λογικού σχήματος που αντιστοιχεί σε κάθε αρχείο SQL, μέσω ενός αποκλειστικού αναλυτή SQL.

3. *Σύγκριση σχήματος*: Σύγκριση των διαδοχικών λογικών σχημάτων έτσι ώστε σταδιακά να προκύπτει η ιστορικότητα του σχήματος.

4. *Οπτικοποίηση και εκμετάλλευση*: Το ιστορικό σχήμα μπορεί στη συνέχεια να απεικονιστεί και να αναλυθεί περαιτέρω, ανάλογα με τις ανάγκες του project.

Κατά το βήμα σύγκρισης του σχήματος ένας πίνακας μπορεί να υπήρχε στην έκδοση 1, να αφαιρεθεί στην έκδοση 2 και να εμφανιστεί ξανά στην έκδοση 3. Οι συγγραφείς αναφέρονται σε αυτό το φαινόμενο λέγοντας ότι ένα αντικείμενο σχήματος μπορεί να έχει πολλές ζωές.

Ως ***ιστορικό σχήματος*** θεωρείται η αναπαράσταση όλων των προηγούμενων εκδόσεων ενός σχήματος βάσης δεδομένων, που περιέχει όλα τα αντικείμενα που έχουν υπάρξει σε ολόκληρη την ιστορία του σχήματος.

Κάθε αντικείμενο (***o***) στο ιστορικό του σχήματος (πίνακας ή στήλη) σχολιάζεται με τα ακόλουθα χαρακτηριστικά:

- ***isDead***: true αν το αντικείμενο δεν υπάρχει στην πιο πρόσφατη (τρέχουσα) έκδοση του σχήματος, αλλιώς ψευδές.
- ***creationDate:*** η ημερομηνία της παλαιότερης έκδοσης σχήματος όπου εμφανίζεται το αντικείμενο, δηλαδή η ημερομηνία δημιουργίας του αντικειμένου.
- ***lastAppearanceDate***: Η ημερομηνία του πιο πρόσφατου σχήματος όπου εμφανίζεται το αντικείμενο.
- ***listOfPresence***: η λίστα των ημερομηνιών έκδοσης σχήματος όπου υπάρχει το αντικείμενο.
- ***listOfDeletion***: η λίστα των ημερομηνιών έκδοσης σχήματος όπου το αντικείμενο έχει διαγραφεί.

Ο *αλγόριθμος παράγωγης σχήματος* βασίζεται σε σύγκριση ανά ζεύγη όλων των εκδόσεων σχήματος.

- Δημιουργία ενός ***κενού*** σχήματος
- Επανάλαβε για όλα τα σχήματα με χρονολογική σειρά, συγκρίνοντας το τρέχον σχήμα ***Si*** με το τρέχον ιστορικό σχήματος ***SH***.

  Σύγκρινε κάθε αντικείμενο - ***o*** (πίνακα ή στήλη) του σχήματος στο ***Si*** και στο ***SH***.

  1. Αν το ***o*** ανήκει στο *Si* αλλά δεν ανήκει σε *SH*. (*Αυτό σημαίνει ότι το **o** δημιουργήθηκε στην έκδοση i)*

  Πρόσθεσε το ***o*** στο *SH* και όρισε την ημερομηνία δημιουργίας μέχρι *σήμερα (Si)*.

  2. Αν το ***o*** ανήκει στο SI και (τώρα) ανήκει στην SH.

  Ενημέρωσε το *listOfPresence* και *lastAppearanceDate* και θέσε το χαρακτηριστικό *isDead = ψευδές*.

  3. Αν το ***o*** ανήκει στο *SH* αλλά δεν ανήκει στο *Si*.

  Αν το *isDead* = ψευδές τότε *το **o*** υπήρχε στο *Si* - 1 αλλά διαγράφηκε στο *Si*. Ενημέρωσε το χαρακτηριστικό *listOfDeletion* και όρισε *το* χαρακτηριστικό *isDead = true*.

Οι μελετητές ανέπτυξαν τρία βασικά εργαλεία:

- *Schema Exrtactor:* Το εργαλείο αυτό επιτρέπει την εξαγωγή όλων των διαδοχικών εκδόσεων του σχήματος βάσης δεδομένων από το αποθετηρίου του έργου
- *Historical schema derivator:* Το εργαλείο αυτό παίρνει ως είσοδο το σύνολο των εξαχθέντων σχημάτων και παράγει το αντίστοιχο ιστορικό σχήματος. Η τρέχουσα έκδοση του εργαλείου είναι σε θέση να εντοπίσει 16 ξεχωριστούς τύπους αλλαγών σχήματος βάσης δεδομένων:
  - προσθήκη / διαγραφή πίνακα
  - προσθήκη / διαγραφή μιας στήλης
  - προσθήκη / διαγραφή κλειδιού
  - προσθήκη / διαγραφή ξένου κλειδιού
  - προσθήκη / διαγραφή ενός δείκτη

- Προσθήκη / διαγραφή / ενημέρωση μιας default τιμή στήλης
- Αλλαγή του τύπου μιας στήλης
- Αλλαγή μιας υποχρεωτικής στήλης σε προαιρετική (δηλαδή, μηδενική) και αντιστρόφως.

- *History visualizer*: Το εργαλείο αυτό παρέχει στον χρήστη μια οπτική αναπαράσταση της εξέλιξης του σχήματος βάσης δεδομένων με την πάροδο του χρόνου, παίρνει το ιστορικό σχήμα σαν είσοδο και (1) συγκρίνει δύο εκδόσεις αυθαίρετων σχημάτων, (2) εξαγάγει το σχήμα της βάσης δεδομένων σε μια δεδομένη ημερομηνία, (3) εξαγάγει το πλήρες ιστορικό ενός συγκεκριμένου αντικειμένου σχήματος (στήλη/πίνακαα), (4) εξάγει στατιστικά για την εξέλιξη του σχήματος και (5) αναλύει τη συμμετοχή κάθε προγραμματιστή σε αυτή την εξέλιξη.

Οι συγγραφείς ανέλυσαν το ιστορικό του σχήματος της ΒΔ OSCAR για περίοδο σχεδόν ***δέκα*** ετών (22/07/2003 - 27/06/2013) στην οποία βρέθηκαν, ***670*** συνολικά διαφορετικές εκδόσεις σχήματος. Στην πρώτη έκδοση υπήρχαν ***88*** πίνακες, ενώ στη τελευταία έκδοση του σχήματος που εξετάστηκε (27/06/2013) υπήρχαν ***445*** πίνακες.
Κάποια από συμπεράσματα στα οποία κατέληξαν οι μελετητές:

- Ο αριθμός των πινάκων παρουσιάζει συνεχή αύξηση. Οι προγραμματιστές εμφανίζονται πολύ απρόθυμοι να αφαιρέσουν πίνακες (προκειμένου να αποφευχθεί ο σημαντικός αντίκτυπος ενός refactoring του σχήματος αφενός στα δεδομένα και αφετέρου στα προγράμματα εφαρμογής). Επίσης εύκολα εντοπίζονται αυτές οι εκδοχές σχήματος που θα μπορούσαν να θεωρηθούν ως "σημαντικές εκδόσεις", δηλαδή, εκείνες οι εκδόσεις όπου έχει προστεθεί ή / και διαγραφεί ένας σημαντικός αριθμός πινάκων. (fig.3 C του [CGMM13] – σελ. 118)
- Ο συνολικός αριθμός των στηλών έχει αυξηθεί από ***2.443*** σε ***13.364*** στήλες σε περίπου 10 χρόνια. Ο αριθμός αυτός ακολουθεί μια παρόμοια τάση με την εξέλιξη των πινάκων, διατηρώντας τον μέσο αριθμό των στηλών ανά πίνακα αρκετά σταθερό με την πάροδο του χρόνου (περίπου 25).
- Οι πίνακες σπάνια αφαιρούνται. Η εξέλιξη του σχήματος συνίσταται (κατά το πλείστον) από την προσθήκη πινάκων και όχι από την αντικατάσταση ή τον δια-

χωρισμό τους. Ο συνολικός αριθμός των διαγραμμένων πινάκων είναι περίπου ***30***. Παρόμοια είναι και η αναλογία δημιουργημένων και διαγραμμένων στηλών. Ο αριθμός των νέων στηλών είναι, συχνά μεγαλύτερος από τον αριθμό των διαγραμμένων στηλών. Συνολικά αφαιρέθηκαν ***3.872*** στήλες, ενώ δημιουργήθηκαν ***14.793*** στήλες. Ωστόσο, ο αριθμός των στηλών ήταν πιο ασταθής: ***964*** στήλες δημιουργήθηκαν στην έκδοση 452, ***954*** στήλες διαγράφηκαν έκδοση 453. Η εξήγηση είναι η ακόλουθη: κατά την έκδοση 452, προστεθήκαν τεράστιοι πίνακες (κάθε ένας από τους οποίους περιλαμβάνει εκατοντάδες στήλες) και η κορυφή που παρατηρούμε προέρχεται από τη διαγραφή ορισμένων από αυτούς τους πίνακες. (fig.3 D [CGMM13] – σελ. 118)

Καταμετρήθηκαν τα παρακάτω ποσά στους 16 τύπους αλλαγών:

add table	*443*	add identifier	*48*	add index	*125*	change default value	*23*
drop table	*86*	drop identifier	*21*	drop index	*32*	change column type	*342*
add column	*2 091*	add foreign key		add default value	*15 09*	set nullable column	*3 2*
drop column	*703*	drop foreign key	*12*	drop default value	*47*	set non-nullable column	*27*

Οι μεγάλοι πίνακες (πάνω από 200 στήλες) δημιουργούνται καθ ’όλη τη διάρκεια ζωής του συστήματος. Αυτό σημαίνει ότι οι πίνακες αυτοί δεν αντικατοπτρίζουν αρχικά προβλήματα σχεδιασμού. Οι μεγάλοι πίνακες συσχετίστηκαν με έναν μηχανισμό επέκτασης του OSCAR, ο οποίος παρέχει στους "power-users" εργαλεία για την προσθήκη πινάκων ώστε να γίνει καταγραφή ειδικών κλινικών δεδομένων.

Μια άλλη ενδιαφέρουσα ιδιότητα, είναι η *σταθερότητα* των πινάκων. Ένας πίνακας που έχει δημιουργηθεί εδώ και πολύ καιρό, και δεν υπόκειται σε συχνές τροποποιήσεις μπορεί να θεωρηθεί ***σταθερός***. Εξήχθη το συμπέρασμα ότι το σχήμα είναι σταθερό, με τους περισσότερους πίνακες να έχουν λιγότερες από **4** τροποποιήσεις.

Στην συγκεκριμένη μελέτη έγινε μια σύνδεση των προγραμματιστών του OSCAR με την εξέλιξη του σχήματος. Οι συγγραφείς υπολόγισαν για κάθε προγραμματιστή, τον αριθμό των πινάκων στην εξέλιξη των οποίων συμμετείχε (by creating, updating or deleting tables).

3.4 Βαρύτητα στην Ακαμψία: Μοτίβα της Εξέλιξης του Σχήματος - και της Απουσίας της - στη Ζωή των Πινάκων

Στο [VaZS16], μελετήθηκε η εξέλιξη οκτώ βάσεων δεδομένων που αποτελούν μέρος μεγαλύτερων έργων ανοιχτού κώδικα, που διατίθενται δημόσια μέσω αποθετηρίων ανοιχτού κώδικα. Συγκεκριμένα, το επίκεντρο της έρευνας ήταν η κατανόηση της εξέλιξης των ***πινάκων*** και πώς αυτή γίνεται. Σε αυτό το έργο διερευνήθηκε η συσχέτιση της εξέλιξης ενός πίνακα με σχετικές ιδιότητες όπως η δυνατότητα διαγραφής, η διάρκεια ζωής, ο αριθμός των χαρακτηριστικών ή η έκδοση της γέννησης ενός πίνακα.

Από την έρευνα προέκυψαν **4** πρότυπα σχετικά με την εξέλιξη του σχήματος. Το *πρότυπο Γ* δείχνει ότι οι πίνακες με μεγάλα σχήματα έχουν την τάση να έχουν μακρές διάρκειες και αποφυγή διαγραφής, το *πρότυπο Κομήτη (Comet)* δείχνει ότι οι πίνακες με τις περισσότερες ενημερώσεις είναι αυτοί με μέγεθος μεσαίου σχήματος, το *πρότυπο αντίστροφου Γ*, που δείχνει ότι πίνακες με μεσαίες ή μικρές διάρκειες παράγουν ποσότητες ενημερώσεων χαμηλότερες από τις αναμενόμενες και το πρότυπο *Κενού Τριγώνου (Empty Triangle)* που δείχνει ότι οι διαγραφές περιλαμβάνουν, κυρίως, νωρίς γεννημένους, ήσυχους πίνακες με μικρή διάρκεια ζωής, ενώ οι παλαιότεροι πίνακες είναι απίθανο να αφαιρεθούν. Συνολικά, οι παρατηρούμενες ενδείξεις δείχνουν έντονα ότι οι βάσεις δεδομένων είναι επιρρεπείς στην ακαμψία, παρά επιρρεπείς στην εξέλιξη. Οι συγγραφείς ονόμασαν το *φαινόμενο έλξη προς την ακαμψία*.

Οι συγγραφείς μελέτησαν 8 σχεσιακά σχήματα ως προς την εξέλιξή τους σε λογικό επίπεδο. Αυτά τα σχήματα είναι (α) ενσωματωμένα σε ανοικτό - source λογισμικό, και (β) έχουν σημαντικό αριθμό δεσμευμένων εκδόσεων, μαζί με μια αρκετά μακρά ιστορία. Για κάθε σύνολο δεδομένων συγκεντρώθηκαν όσες εκδόσεις σχήματος (αρχεία DDL) ήταν δυνατόν να εξαχθούν από τα δημόσια αποθετήρια αποθήκευσης πηγαίου κώδικα (cvs, svn, git).

Προκειμένου να μεγιστοποιηθεί η εγκυρότητα του έργου, οι συγγραφείς επέλεξαν μόνο τις αλλαγές του τμήματος της βάσης δεδομένων του έργου όπως αυτές ενσωματώνονται στον κορμό του έργου. Ως εκ τούτου, συνέλεξαν μόνο τις εκδόσεις των ΒΔ, που αφορούν στον κορμό ή στον κύριο κλάδο. Τα αρχεία συλλέχθηκαν τον Ιούνιο 2013. Τα αρχεία επεξεργάστηκαν από το εργαλείο ***Hecate***, που

αναπτύχθηκε από τους συγγραφείς, και ανιχνεύει τις αλλαγές τόσο σε επίπεδο χαρακτηριστικών όσο και σχέσεων. Για κάθε μετάβαση μεταξύ συνεχόμενων εκδόσεων του σχήματος, ανιχνεύει:

(α) ποιοι πίνακες εισήχθησαν και ποιοι διαγράφησαν και

(β) ποιοι πίνακες ενημερώθηκαν σε επίπεδο χαρακτηριστικών και ειδικά, για κάθε επηρεαζόμενο πίνακα, τα χαρακτηριστικά που εισάγονται, διαγράφονται, τα χαρακτηριστικά στα οποία αλλάχθηκε ο τύπος δεδομένων ή αλλάχθηκε το πρωτεύον κλειδί.

Λαμβάνοντας υπόψη όλα αυτά τα εξαγόμενα δεδομένα, εντοπίστηκαν για κάθε πίνακα τα εξής χαρακτηριστικά:

- Διάρκεια (σε αριθμό εκδόσεων).
- Γέννηση και θάνατος (ως αριθμός έκδοσης).
- Το μέγεθος του σχήματος του πίνακα σε τρεις παραλλαγές:
    - ***(i)*** κατά τη γέννησή του
    - ***(ii)*** στο τέλος της διάρκειας ζωής του (ή στο τέλος του συνόλου δεδομένων, εάν ο πίνακας επιβιώνει μέχρι την τελευταία μελετημένη έκδοση)
    - ***(iii)*** ως μέσο μέγεθος σχήματος κατά τη διάρκεια της ζωής του.
- Μέσος αριθμός ενημερώσεων ανά μετάβαση - ***ATU*** που είναι ο λόγος του συνόλου των ενημερώσεων (sum (updates)) προς την διάρκεια (duration)

Για την μελέτη των πινάκων εκτός από τον αριθμό των updates (είναι θεωρητικά πιθανό ότι ένα πίνακας με μικρή διάρκεια ζωής να μην έχει τη δυνατότητα να λάβει ένα μεγάλο αριθμό updates, αν και η δραστηριότητα αλλαγής του είναι έντονη σε αυτή μικρή διάρκεια ζωής) ορίστηκε ο *Μέσος αριθμός ενημερώσεων ανά μετάβαση – **ATU*** ως το κλάσμα του αριθμού των ενημερώσεων που γίνονται σε έναν πίνακα καθ ’όλη τη διάρκεια της ζωής διά την διάρκεια. Έτσι, η μέση μεταβατική ενημέρωση ενός πίνακα πρακτικά μετρά πόσο δραστήρια ή άκαμπτη είναι η ζωή του με την κανονικοποίηση του αριθμού των ενημερώσεων διά τη διάρκειά του.

<u>επιζών</u> είναι ένας πίνακας που υπήρχε στην τελευταία έκδοση της ΒΔ

μη - επιζών είναι έναν πίνακα που έχει εξαλειφθεί από τη βάση δεδομένων.
Οι πίνακες διακρίνονται σε σχέση με την εικόνα ενημερώσεων ως εξής:
Ενεργοί πίνακες ή top changers είναι οι πίνακες με (α) υψηλό ATU, υπερβαίνοντας εμπειρικά το όριο του 0,1 και, (β) ένταση ενημερώσεων που υπερβαίνει το όριο των 5 ενημερώσεων.
Άκαμπτοι πίνακες είναι αυτοί που δεν αλλάζουν καθόλου και αυτοί που δεν επιβίωσαν, καθώς απομακρύνθηκαν χωρίς αλλαγή στο σχήμα τους.
Ήσυχοι πίνακες οι υπόλοιποι πίνακες, με πολύ λίγες ενημερώσεις ή ATU μικρότεροο από ***0,1***. Οι ήσυχοι πίνακες κυριαρχούν στο τοπίο.

Οι συγγραφείς εντόπισαν ότι υπάρχουν " *οικογένειες* " πινάκων των οποίων η επιβίωση ή η διαγραφή σχετίζεται με κάποια χαρακτηριστικά:

Wide survivors (Ευρεία επιζώντες): Η σχέση του μεγέθους του σχήματος ενός πίνακα με τη διάρκεια αποκάλυψε ένα άλλο ενδιαφέρον πρότυπο, το οποίο ονομάστηκε *πρότυπο* Γ: "λεπτοί" πίνακες, με μικρά μεγέθη σχημάτων, μπορούν έχουν αυθαίρετες διάρκειες, ενώ *οι "πλατιοί" πίνακες, με μεγαλύτερα μεγέθη σχήματος, έχουν μεγάλες πιθανότητες επιβίωσης.*

Entry level removals (Καταργήσεις επιπέδου εισόδου): Η συντριπτική πλειοψηφία των διαγραμμένων πινάκων είναι: (1) νεογέννητοι, (2) με λίγες ή καθόλου ενημερώσεις, (3) με μικρή διάρκεια ζωής και αρκετά συχνά ισχύουν και τα τρία

Old Timers: ο χρόνος είναι στο πλευρό τους: Είναι αρκετά σπάνιο να καταργούνται πίνακες όταν έχουν καταφέρει να επιβιώσουν για μακρύ χρονικό διάστημα

3.4.1 Σχέσεις μεταξύ του Μέγεθος, της Διάρκειας και των Ενημερώσεων του Πίνακα.

Διάρκεια στον χρόνο - Μέγεθος πίνακα

Για κάθε πίνακα, υπολογίστηκε ένα μέτρο διάρκειας, διαιρώντας τη διάρκεια του πίνακα με την μέγιστη διάρκεια πίνακα της ΒΔ. Αυτό έχει ως αποτέλεσμα όλοι οι πίνακες να έχουν μία κανονικοποιημένη διάρκεια στην περιοχή του (0... 1] και ταξινομήθηκαν σε 3 κατηγορίες: (i) short-lived (συμπεριλαμβανομένων των πινάκων που απομακρύνθηκαν από το σύστημα και των πινάκων των οποίων η σύ-

ντομη διάρκεια οφείλεται στην καθυστερημένη εμφάνισή τους), (ii) πίνακες μέσης διάρκειας και (iii) πίνακες μεγάλης διάρκειας ζωής (long lived tables).

Ένα μεγάλο ποσοστό των πινάκων ζει ένα σύντομο χρονικό διάστημα. Πάνω από το ***30%*** έχουν μικρή διάρκεια ζωής.

Ένα μεγάλο ποσοστό των πινάκων έχει μεγάλη διάρκεια ζωής: στα μισά σύνολα δεδομένων το ποσοστό αυτών των πινάκων κυμαίνεται μεταξύ ***30-40%***, και σε άλλα τρία σύνολα δεδομένων, οι πίνακες μεγάλης διάρκειας υπερβαίνουν το ***50%***.

Το μέσο μέγεθος του πίνακα (δηλαδή ο μέσος αριθμός των χαρακτηριστικών κατά τη διάρκεια της ζωής του) είναι πολύ κοντά στο αντίστοιχο μέγιστο και ελάχιστο μέγεθος καθώς παρατηρήθηκε ότι οι πίνακες δεν αλλάζουν σε μεγάλο βαθμό.

Το πλήθος των πινάκων αναλύεται ως εξής:

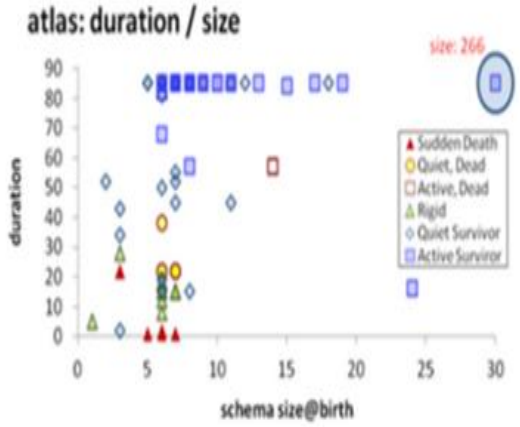
47% είναι μικροί πίνακες με λιγότερο από 5 χαρακτηριστικά.

36% είναι οι πίνακες με χαρακτηριστικά 5 - 10

17% οι πίνακες με περισσότερα από 10 χαρακτηριστικά

Μέγεθος του σχήματος κατά τη γέννηση - Διάρκεια του πίνακα

Μελετήθηκε ο συνδυασμός: μέγεθος σχήματος κατά τη γέννηση - διάρκεια του πίνακα μέσα από μια οπτική απεικόνιση (fig. 4-5 του [VaZS16] σελ. 28-29) όπου στον άξονα x απεικονίστηκε ο αριθμός των χαρακτηριστικών του σχήματος στην γέννηση του και στο άξονα y η διάρκεια ζωής.



Σχήμα 3-6 - The Γ pattern -- από το [VaZS16] με άδεια των συγγραφέων – σελ 29

Παρατηρήθηκε ένα φαινόμενο που ονομάζουμε ***πρότυπο Γ***: πίνακες με μικρά μεγέθη σχημάτων μπορεί να έχουν αυθαίρετες διάρκειες, ενώ οι πίνακες με μεγαλύτερα μεγέθη σχημάτων διαρκούν πολύ.

Πλήθος χαρακτηριστικών πίνακα– Διάρκεια

Υπάρχει μια μεγάλη πλειοψηφία των πινάκων, σε όλα τα σύνολα δεδομένων, των οποίων το μέγεθος είναι μεταξύ 0 και 10 πεδία. Το ενδιαφέρον είναι πώς το μικρό μέγεθος του σχήματος τους δεν καθορίζει τη διάρκεια τους. Από την άλλη πλευρά όμως, παρατηρήθηκε ότι κάθε φορά που κάποιος πίνακας υπερβαίνει την κρίσιμη τιμή των ***10*** χαρακτηριστικών στο σχήμα του, οι πιθανότητες επιβίωσης είναι υψηλές. Παρατηρήθηκε ότι πολύ συχνά, η πλειοψηφία των wide πινάκων δημιουργήθηκαν νωρίς και δεν διαγράφονται αργότερα.

Στατιστικά στοιχεία για το πρότυπο Γ

Για να ελέχθη το πρότυπο Γ, πρέπει να ελεγχθεί η πιθανότητα ότι επιβιώνει ένας πίνακας με ευρύ σχήμα (ένα σχήμα είναι ευρύ, όταν είναι αυστηρά πάνω από 10 πεδία). Η κορυφαία διάρκεια (το ανώτερο τμήμα του σχήματος Γ) προσδιορίζεται ως το εύρος των τιμών που υπερβαίνουν το ***90%*** της μέγιστης διάρκειας (δηλαδή, το ανώτερο 10% των τιμών στην y -άξονα).

Θεωρείται ότι ένας πίνακα έχει γεννηθεί νωρίς, αν η γέννησή του είναι στο χαμηλότερο ***33%*** των εκδόσεων. Αντίστοιχα, γεννιούνται καθυστερημένα μετά το ***77%*** του αριθμού των εκδόσεων. Όλα τα σύνολα δεδομένων περιλαμβάνουν ευρείς πίνακες, με ποσοστά που συνήθως κυμαίνονται μεταξύ ***6%*** και ***17%***

Τα σύνολα δεδομένων επιβεβαιώνουν την υπόθεση ότι όταν ένας πίνακας είναι μεγάλος έχει υψηλές πιθανότητες επιβίωσης, με ένα ποσοστό υψηλότερο από ***85 %***. (fig. 6 του [VaZS16] – σελ. 30).

Η εξήγηση για το "*If you 're wide, you survive*" που αποτελεί μέρος του προτύπου Γ, οφείλεται στο ότι η επίδραση μιας διαγραφής ενός ευρύ πίνακα (ο οποίος έχει μεγαλύτερη πιθανότητα να ενεργεί ως πίνακας γεγονότων), είναι μεγάλη γιατί συχνά προσπελάζεται από ερωτήματα αιτήσεων που περιβάλλουν τη ΒΔ. Έτσι, η απομάκρυνσή του συνεπάγεται υψηλό κόστος συντήρησης, κάνοντας τους προγραμματιστές εφαρμογών και τους διαχειριστές μάλλον απρόθυμους να τον αφαιρέσουν.

Εντοπίστηκε ότι στα μισά σύνολα δεδομένων το ποσοστό των ευρέων πινάκων που δημιουργούνται νωρίς και δεν διαγράφονται στη συνέχεια, είναι πάνω από το ***70%*** και σε δύο από αυτά τα σύνολα δεδομένων, το ποσοστό αυτών των πινάκων είναι το ***ένα τρίτο***.

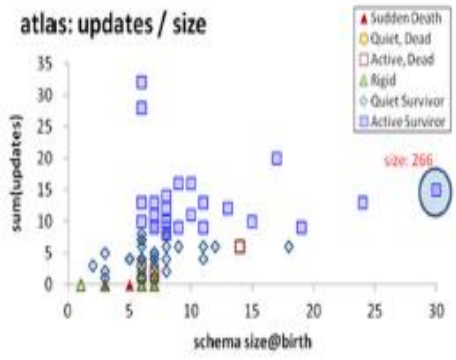
Για τους ευρείς πίνακες υπάρχει πολύ ισχυρή συσχέτιση μεταξύ αυτών που έχουν γεννηθεί νωρίς και έχουν μεγάλη διάρκεια, βρίσκονται δηλαδή στο πάνω μέρος του προτύπου Γ. (6η & 7η στήλη fig 6 του [VaZS16] – σελ. 11) Η συσχέτιση Pearson είναι συνολικά ***88%***, και ***100%*** για τα σύνολα δεδομένων με υψηλό ποσοστό πρόωρα γεννημένων ευρέων πινάκων. Συνολικά, τα δεδομένα είναι σαφώς διπολικά σε αυτό το πρότυπο: οι μισές περιπτώσεις απαντούν στην ερευνητική ερώτηση θετικά, με υποστήριξη υψηλότερη από 70%, ενώ οι υπόλοιπες περιπτώσεις το διαψεύδουν.

Σε όλα τα σύνολα δεδομένων, εάν ένας ευρύς πίνακας έχει μακρά διάρκεια, στο πάνω μέρος του προτύπου Γ, αυτό δηλώνει ότι ο πίνακας ήταν επίσης πρόωρα γεννημένος και επέζησε.

3.4.2 Μεγέθους του Σχήματος κατά τη Γέννηση - Αριθμός Ενημερώσεων ενός Πίνακα

Αρχικά οι συγγραφείς παρατηρούν τα scatterplots που αντιπροσωπεύουν οπτικά τη συσχέτιση μεταξύ του μεγέθους του σχήματος κατά τη γέννηση και του αριθμού των ενημερώσεων ενός πίνακα. Χρησιμοποιούν ένα scatterplot ανά ΒΔ, με κάθε σημείο του scatterplot να αναφέρεται σε ένα πίνακα (fig 4, 7 του [VaZS16] – σελ. 28- 31). Στο x -άξονα απεικονίζεται το μέγεθος του σχήματος του πίνακα και στον y -άξονα τα updates του κάθε πίνακα. Υπάρχουν δύο κύριες συστάδες σημείων: (α) μία μεγάλη, ένα πυκνό σύμπλεγμα κοντά την αρχή των αξόνων, που υποδηλώνει μικρό μέγεθος και μικρό ποσό της μεταβολής και (β) ένα αραιό σύνολο των ακραίων τιμών, σπασμένο σε δύο υποκατηγορίες: (β1) πίνακες μεσαίου μεγέθους με μεσαίες έως μεγάλες ποσότητες αλλαγών και (β2) "ευρείς" πίνακες με μεγάλα μεγέθη σχημάτων με μικρές έως μεσαίες ποσότητες αλλαγών.

Οι συγγραφείς ονόμασαν αυτή την κατανομή δηλ. μια κεντρική ομάδα κοντά στην αρχή των αξόνων μαζί με δύο ουρές έξω από αυτήν, ως πρότυπο ***κομήτη***.



Σχήμα 3-7 - The Comet pattern από το VaZS16 με άδεια των συγγραφέων – σελ 35

Ο πυρήνας και οι ουρές του κομήτη

Το πρώτο σύμπλεγμα "πυρήνα" τυπικά περιέχεται μέσα ένα "κουτί" μεγέθους ***10 X 10*** (δηλαδή όχι περισσότερα από 10 πεδία συνήθως έχουν ως αποτέλεσμα όχι περισσότερες από 10 αλλαγές). Αυτό αποδίδεται στον μικρό ρυθμό της αλλαγής που υφίστανται οι πίνακες, ως αποτέλεσμα των μικρών πιθανοτήτων για ενημερώσεις χαρακτηριστικών σε στενούς πίνακες. Την ίδια στιγμή, στα περισσότερα από τα σύνολα δεδομένων, οι πίνακες με το μεγαλύτερο ποσό αλλαγής είναι όχι απαραίτητα οι μεγαλύτεροι όσον αφορά τα πεδία, αλλά πίνακες των οποίων το σχήμα είναι κατά μέσο όρο μία τυπική απόκλιση πάνω από το μέσο όρο.

Τυπικά, οι πίνακες μεσαίου μεγέθους δείχνουν στρώματα όλων των ειδών στη συμπεριφορά αλλαγής, καθώς καλύπτουν ολόκληρο τον y -άξονα, ενώ οι (λίγοι) πίνακες με μεγάλο μέγεθος δείχνουν αξιοσημείωτη δραστηριότητα αλλαγής (δηλ. όχι μηδέν ή μικρή), η οποία βρίσκεται λίγο κάτω από τη μέση του y -άξονα σε πολλές περιπτώσεις.

Στατιστικά στοιχεία για το πρότυπο κομήτη

Για να αξιολογηθεί το πρότυπο κομήτη αξιολογήθηκαν τρεις μετρήσιμες ποσότητες. Συγκεκριμένα: (α) το ποσοστό του πληθυσμού που βρίσκεται εντός του πλαισίου των 10 × 10 (το πολύ 10 ιδιότητες στο σχήμα και 10 ενημερώσεις κατά τη διάρκεια της ζωής τους), (β) το μέγεθος του σχήματος στη γέννηση των πινάκων που τελικά δείχνουν μεγάλο αριθμό ενημερώσεων κατά τη διάρκεια της ζωής τους και (γ) τα updates σε πίνακες μεγάλων μεγεθών.

Συνήθως, περίπου το ***70%*** των πινάκων μιας βάσης δεδομένων βρίσκεται εντός του πλαισίου 10 × 10, δηλαδή στον πυρήνα του κομήτη (δηλαδή πίνακες με μικρό μέγεθος και ήσυχη συμπεριφορά ενημέρωσης).

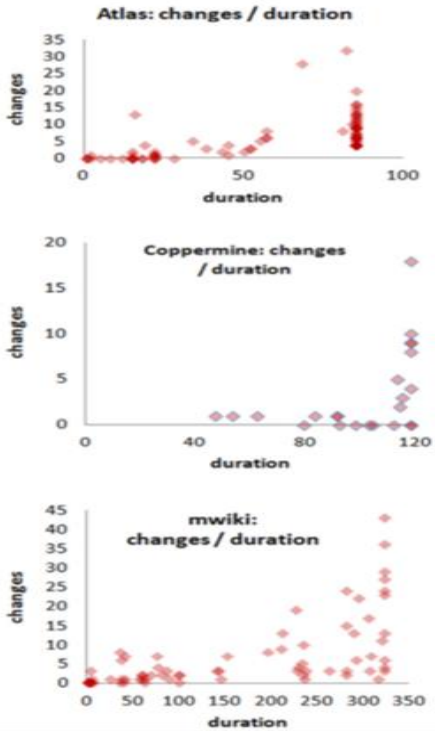
Προκειμένου να ελέγξουν που τείνει το μέγεθος των πινάκων με πολλές αλλαγές, οι συγγραφείς επέλεξαν το ***5%*** των πινάκων με το μέγιστο συνολικό αριθμό ενημερώσεων (updates). Οι συγγραφείς υπολόγισαν τη μέση τιμή του μεγέθους του σχήματος κατά τη γέννηση, και τη μέση τιμή για το 5% των πινάκων. Αξιοσημείωτο είναι το γεγονός ότι σε ***5*** από τις ***8*** περιπτώσεις, το μέσο σχήμα των κορυφαίων πινάκων σε σχέση με το άθροισμα των ενημερώσεων είναι εντός ***0,4*** και ***0,5*** της μέγιστης τιμής (κατά μέσο όρο, φτάνει το ***48%*** της μέγιστης τιμής - πρακτικά, η μέση τιμή του domain) και ποτέ πάνω από το ***0,65***. Συνολικά, βάσει αυτού του στοιχείου, μπορεί να εξαχθεί το συμπέρασμα *ότι πίνακες με μεγάλους αριθμούς ενημερώσεων τείνουν να έχουν τυπικά μεσαία μεγέθη σχήματος* (ισοδύναμα: μπορεί να βεβαιωθεί ότι η ουρά του κομήτη αφορά πίνακες με κορυφαία ποσά ενημερώσεων και μεσαίου μεγέθους σχήματα)

Προκειμένου να ελέγξουν που τείνουν οι αλλαγές των wide πινάκων κατά την ημερομηνία γέννησης τους, οι μελετητές πήραν το μέγιστο 5% όσον αφορά το μέγεθος του σχήματος κατά τη γέννησή τους, και παρατηρήθηκε η σχέση της συμπεριφοράς ενημέρωσής τους σε αντίθεση με την συμπεριφορά ενημέρωσης του υπολειπόμενου του συνόλου δεδομένων. Απομονώθηκε λοιπόν το 5% των πινάκων σε σχέση με το μέγεθος του σχήματος κατά τη γέννηση και κατά μέσο όρο το ποσό των ενημερώσεών τους. Σε ***5*** από τις ***8*** περιπτώσεις, ο μέσος όρος ενημερώσεων για τους κορυφαίους πίνακες είναι εντός του ***35-45%*** της μέγιστης τιμής για την ποσότητα των ενημερώσεων. Επιπλέον, σε 4 από τα 8 σύνολα δεδομένων, ο πίνακας με τον μεγαλύτερο αριθμό ενημερώσεων συμπεριλήφθηκε στο σύνολο των κορυφαίων πινάκων (*ισοδύναμα*: *η ουρά του κομήτη αφορά ευρείς πίνακες με μεσαία ποσά ενημερώσεων*).

3.4.3 Σχέση Διάρκειας - Αριθμός Ενημερώσεων ενός Πίνακα

Η μελέτης της σχέσης της διάρκειας με τα updates ενός πίνακα ξεκινά από την οπτική παρατήρηση των αντίστοιχων scatterplots, (η διάρκεια στο x -άξονα και τα updates στον y –άξονα – βλέπε σχήμα 3-8). Παρατηρήθηκε ένα φαινόμενο που

ονομάστηκε το ***πρότυπο αντίστροφου Γ***: Οι πίνακες με μικρή διάρκεια υποβάλλονται σε λίγες αλλαγές, οι πίνακες με μεσαία διάρκειας υφίστανται λίγες ή μεσαίου αριθμού αλλαγές και *long-lived* πίνακες επιδεικνύουν όλα τα είδη της συμπεριφοράς σε σχέση με τα *updates*.



Σχήμα 3-8-The Inverse Γ pattern in all datasets.- από το VaZS16 με άδεια των συγγραφέων–σελ 35

Πρότυπο Αντίστροφου Γ: Μια πρώτη γεύση της βαρύτητας της ακαμψίας

Η συντριπτική πλειοψηφία των πινάκων έχει "ήρεμη" ζωή, χωρίς υπερβολική δραστηριότητα αλλαγής. Υπάρχει συνεπώς μεγάλη πιθανότητα ένας πίνακας να είναι χαμηλά στον *y* –άξονα. Λεπτομερώς, όλα τα scatterplots παρουσιάζουν τρεις ομαδοποιήσεις:

- μια σύντομης διάρκειας ζωή έχει λίγες (συνήθως κοντά στο μηδέν) αλλαγές.
- η μέση διάρκεια έχει κάποιες μεγαλύτερες πιθανότητες ώστε να συμβούν 5-10 αλλαγές
- ψηλά στον άξονα y (πολλές αλλαγές) βρίσκονται πίνακες με μεγάλη διάρκεια ζωής.

Ακόμα, αντί για ένα τρίγωνο, το γράφημα δείχνει κυρίως ένα <u>αντίστροφο Γ</u>: υπάρχει μια εντυπωσιακή σπανιότητα των πινάκων με μέσες διάρκειες και μεσαία αλλαγή που θα ήταν μέρος της διαγώνιου (ή θα γέμιζε το εσωτερικό του τρίγωνο) Η πλειοψηφία των πινάκων έχει ήσυχη, ήρεμη ζωή και άρα ωθεί το τρίγωνο να γίνει ένα αντίστροφο Γ.

Οι συγγραφείς πιστεύουν ότι το πρότυπο αντίστροφου Γ πρότυπο είναι μια αρκετά ζωντανή επίδειξη της ***έλξης προς την ακαμψία*** στη ζωή των πινάκων. Παρατηρήθηκε μια τάση των πινάκων να "προσελκύονται" στις χαμηλότερες δυνατές τιμές, κοντά στον οριζόντιο άξονα. Φαίνεται σαν να υπάρχει μια δύναμη βαρύτητας που τους τραβάει προς το μηδέν με πολύ λίγους από αυτούς να ξεφεύγουν από αυτή τη δύναμη.

3.4.4 Ζωή και Θάνατος Πίνακα

Updates - επιβίωση

Για τους μη επιζώντες, αυτοί με αιφνίδιο θάνατο είναι οι πιο συχνοί, δεύτεροι έρχονται οι ήσυχοι μη -επιζήσαντες και τρίτοι ένα πολύ μικρό ποσοστό των πινάκων κοντά στο ***0 - 6 %*** που πεθαίνουν μετά από κάποια έντονη δραστηριότητα.

Για τους επιζώντες, σημαντικό είναι το ποσοστό των άκαμπτων πινάκων, το τοπίο κυριαρχείται από ήσυχους πίνακες, συνήθως ξεπερνώντας ***30%*** των πινάκων και φτάνοντας μέχρι και το ***60%*** των πινάκων για μερικά σύνολα δεδομένων.

Οι ενεργοί επιζώντες είναι μια μικρή μειοψηφία (περίπου ***5%***) του πληθυσμού.

Καθώς η βάση δεδομένων ωριμάζει, το ποσοστό ενεργών πινάκων πέφτει (η δραστηριότητα της εξέλιξης φαίνεται να ηρεμεί).

Γέννηση πίνακα- Διάρκεια -Ενημερώσεις

Η επεξεργασία των δεδομένων αποκάλυψε ότι *πολύ συχνά*, οι πινάκες που βρίσκονται στη κορυφή των αλλαγών *γεννιούνται νωρίς*, *ζουν πολύ*, *έχουν υψηλό ATU και συνεπώς*, *ένα μεγάλο ποσό συνολικών updates*.

<u>*Οι επιζώντες top-changers ζουν πολύ*</u>: Από την κατακόρυφη γραμμή του *αντίστροφου* Γ *γίνεται κατανοητό ότι αρκετοί από τους top-changers έχουν* πολύ μεγάλη διάρκεια ζωής. Οι *top-changers* είναι αυτοί με το υψηλότερο ATU. Είναι ό-

μως θεωρητικά πιθανό (και στην πραγματικότητα, υπάρχει μια τέτοια περίπτωση στο Mediawiki, που φαίνεται κοντά στην αρχή των αξόνων) ένας πίνακας με μικρή διάρκεια ζωής με σχετικά χαμηλές συνολικές ενημερώσεις να είναι ένας top-changer. Στην πράξη, όμως, η μακροζωία και το υψηλό ATU έχουν υψηλή συσχέτιση.

Στατιστικά στοιχεία για τους top-changers

Ένας πίνακας, θεωρείται ότι έχει γεννηθεί νωρίς εάν η έκδοση γέννησης είναι στο χαμηλότερο 33% των εκδόσεων. Αντίστοιχα, οι late-deaths λαμβάνουν χώρα μετά το 77% του αριθμού των εκδόσεων. Οι διάρκειες ακολουθούν το ίδιο πρότυπο: οι short-durations είναι στο χαμηλότερο 33% της διάρκειας και οι long-durations στο ανώτερο 23% (αυστηρά υψηλότερο από το 77% των πινάκων μέγιστης διάρκειας). Ο όρος "λίγες ενημερώσεις" αναφέρεται στους ήσυχους και άκαμπτους πίνακες.

Τα δεδομένα καταδεικνύουν μια έντονη σχέση μεταξύ της διάρκειας, της επιβίωσης και της γέννησης.

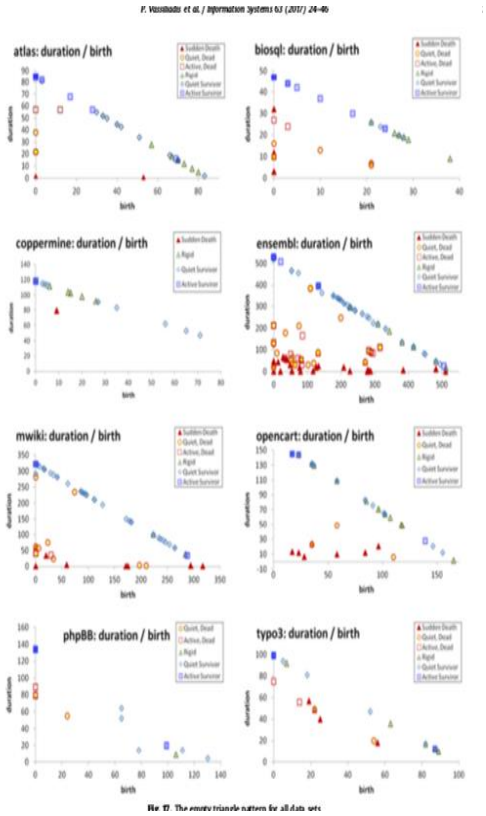
- Σε όλα τα σύνολα δεδομένων, οι ενεργοί πίνακες γεννιούνται νωρίς με ποσοστό που υπερβαίνει το 75%.
- Με τις εξαιρέσεις δύο συνόλων δεδομένων, οι ενεργοί πίνακες επιβιώνουν με ποσοστό μεγαλύτερο από 70%. (Στο Mediawiki το ποσοστό είναι 60% και στο Ensemble, όπου πάρα πολλοί πίνακες πεθαίνουν, το ποσοστό είναι 48%.)
- Η πιθανότητα ενός ενεργού πίνακα να έχει μεγάλη διάρκεια είναι υψηλότερη από το 50% σε 6 από τα 8 σύνολα δεδομένων.
- Ένας ενεργός πίνακας με μεγάλη διάρκεια έχει γεννηθεί νωρίς και επιβίωσε με πιθανότητα 100%.
- Ένας ενεργός – επιζών πίνακας που έχει μακρά διάρκεια γεννήθηκε νωρίς με πιθανότητα 100%.

Θάνατος Πίνακα - Διάρκεια – Ενημερώσεις

Οι πίνακες που έχουν αφαιρεθεί από το σχήμα είναι ήσυχοι, νωρίς γεννημένοι, με μικρή διάρκεια ζωής και αρκετά συχνά και τα τρία. Στο (fig. 17 του [VaZS16] – σελ. 39) βλέπουμε τη συσχέτιση της γέννησης, της διάρκειας, της επιβίωσης και ενημέρωσης σε ένα μόνο διάγραμμα, για τα οκτώ σύνολα δεδομένων.

Υπάρχει μεγάλη συγκέντρωση των διαγραμμένων πινάκων σε μια ομάδα νεαρών πινάκων, που έχουν αφαιρεθεί γρήγορα, με λίγες ή καθόλου ενημερώσεις. Λίγοι πίνακες που διαγράφηκαν δείχνουν είτε καθυστερημένη γέννηση, είτε μακρά διάρκεια, ή υψηλό ATU ή μεγάλο αριθμό ενημερώσεων. Στο τρίγωνο που σχηματίζεται στο σχήμα από τους δύο άξονες και τη γραμμή των επιζώντων, η διαγώνιος των επιζώντων είναι αναμενόμενη: αν ο πίνακας είναι επιζών, όσο νωρίτερα γεννήθηκε, τόσο περισσότερο ζει. Το νέο εύρημα στο τρίγωνο είναι: ***(α)*** η διαγώνιος γίνεται σχεδόν αποκλειστικά από επιζώντες και (β) το τρίγωνο είναι κυρίως κενό.

Είναι αποδεκτό για τους μη επιζώντες να έχουν μεγάλη διάρκεια και να έχουν γεννηθεί στη μέση του κύκλου ζωής της ΒΔ: αν δεν είχαμε την εμφάνιση των μη-επιζώντων στην αρχή των αξόνων, θα μπορούσαν να εμφανίζονται μη επιζώντες στην περιοχή της διαγώνιου και φυσικά, μια ομοιόμορφη εξάπλωση στο εσωτερικό του τρίγωνου. Στην πραγματικότητα, στην περίπτωση του Mediawiki, βλέ-



Σχήμα 3-9-The empty triangle pattern for all data sets-από το VaZS16 με άδεια των συγγραφέων–σελ 35

πουμε ένα ζευγάρι (αλλά μόνο ένα ζευγάρι) τέτοιων περιπτώσεων και για τις δύο προαναφερθείσες περιπτώσεις.

Συνολικά, διαπιστώθηκε ότι το κενό του τριγώνου είναι μια απροσδόκητη αποκάλυψη και επομένως, αναφέρεται αυτή η διαπίστωση ως ***Empty Triangle pattern.***

Είναι πολύ σπάνιο πίνακες να απομακρύνονται σε γήρας, αν και σε κάθε σύνολο δεδομένων υπάρχουν μερικές τέτοιες περιπτώσεις, συνήθως, η περιοχή με μεγάλη διάρκεια κατοικείται κατά συντριπτική πλειοψηφία από επιζώντες.

Στατιστικά στοιχεία για τη σχέση Ηλικίας - Πιθανότητα απομάκρυνσης
Οι πίνακες που καταργούνται είναι πίνακες με λίγες ενημερώσεις (πάνω από ***85%*** σε 6 από τις 8 περιπτώσεις) και έχουν γεννηθεί νωρίς.

Πίνακες με σύντομη διάρκεια: σε 3 σύνολα δεδομένων εντοπίστηκαν πολύ λίγοι νεκροί πίνακες, ενώ στα άλλα σύνολα δεδομένων, το ποσοστό των πινάκων που πεθαίνουν μετά από μια σύντομη ζωή υπερβαίνουν το ***75%***.

Στην μελέτη δεν εμφανίστηκε περίπτωση αφαίρεσης πίνακα με μακρά διάρκεια ζωής. Η εξήγηση γι' αυτό είναι ότι έχουν χτιστεί εφαρμογές γύρω από αυτό και το κόστος συντήρησης για τα υπάρχοντα ερωτήματα μειώνει την πιθανότητα διαγραφής αυτών των πινάκων.

3.5 Οδηγός Εξέλιξης του Σχήματος των Πινάκων: Η Αποφυγή μιας Άκαμπτης Παιδικής Ηλικίας Οδηγεί στην Πορεία προς μια Ήσυχη Ζωή

Στην εργασία [VaZa17] μελετήθηκαν οι παράγοντες που σχετίζονται με την επιβίωση ενός πίνακα στο πλαίσιο της εξέλιξης του σχήματος σε λογισμικό ανοιχτού κώδικα. Η έρευνα διεξήχθη στο σχήμα 8 προγραμμάτων λογισμικού ανοιχτού κώδικα που περιλαμβάνουν σχεσιακές βάσεις δεδομένων και από τις οποίες εξήχθησαν πληροφορίες που σχετίζονται με την επιβίωση ή το θάνατο των πινάκων τους. Ερευνούν κατά πόσο η δραστηριότητα και η διάρκεια σχετίζονται και με την επιβίωση.

Τα ευρήματα συνοψίζονται ως επί το πλείστον από ένα πρότυπο, το οποίο ονομάστηκε από τους συγγραφείς ***ηλεκτρόλυση***, λόγω της διαγραμματικής του

αναπαράστασης, δηλώνοντας ότι οι νεκροί (δηλαδή οι πίνακες που κάποια στιγμή αφαιρέθηκαν οριστικά από το σχήμα της ΒΔ) και οι επιζώντες (δηλαδή οι πίνακες που υπάρχουν στην ΒΔ μέχρι την χρονική στιγμή κατά την οποία μελετάται η ΒΔ) ζουν πολύ διαφορετικές ζωές.

Εξίσου σημαντική είναι η απόδειξη ότι η εξέλιξη του σχήματος "πάσχει" από τον ανταγωνισμό της βαρύτητας με την ακαμψία, δηλαδή την τάση να ελαχιστοποιηθεί η εξέλιξη όσο το δυνατόν περισσότερο, προκειμένου να ελαχιστοποιηθεί η επίπτωση στον περιβάλλοντα κώδικα.

Η αξιολόγηση ενός πίνακα γίνεται με βάση τις α) πληροφορίες σχετικά με την έκδοση/ημερομηνία γέννησης και θανάτου (κατά περίπτωση), και τη διάρκεια, β) πληροφορίες σχετικά με το αρχικό, το τελικό και το μέσο μέγεθος του σχήματος (σε αριθμό χαρακτηριστικών), μαζί με την αναλογία αλλαγής μεγέθους μεταξύ της πρώτης και της τελευταίας έκδοσης, γ) πληροφορίες σχετικά με την ενημέρωση, όπως ο αριθμός των αλλαγών του έγιναν, το ποσοστό αλλαγής κλπ.

Ταξινομήθηκαν οι πίνακες με βάση ***(α)*** την επιβίωσή τους (δηλαδή την παρουσία τους στην τελευταία έκδοση του ιστορικού του σχήματος ή όχι), που χαρακτηρίζουν τους πίνακες ως επιζώντες ή νεκρούς, ***(β)*** τη συμπεριφορά τους ως προς τη δραστηριότητα χαρακτηρίζοντας τους ως άκαμπτους (μηδέν ενημερώσεις), ενεργούς όταν ATU $> 0{,}1$ (***ATU*** = Μέσος αριθμός ενημερώσεων ανά μετάβαση - είναι ο λόγος του συνόλου των ενημερώσεων προς την διάρκεια) και ήσυχους και ***(γ)*** από τον συνδυασμό των παραπάνω μέσω του καρτεσιανού γινομένου τους τους, το οποίο ονομάζουμε LifeAndDeath class

Οι συγγραφείς μελέτησαν 8 σχεσιακά σχήματα ως προς την εξέλιξή τους σε λογικό επίπεδο. Αυτά τα σχήματα είναι (α) ενσωματωμένα σε ανοικτό - source λογισμικό, και (β) έχουν σημαντικό αριθμό δεσμευμένων εκδόσεων, μαζί με μια αρκετά μακρά ιστορία. Για κάθε σύνολο δεδομένων συγκεντρώθηκαν όσες εκδόσεις σχήματος (αρχεία DDL) ήταν δυνατόν να εξαχθούν από τα δημόσια αποθετήρια αποθήκευσης πηγαίου κώδικα (cvs, svn, git).

Προκειμένου να μεγιστοποιηθεί η εγκυρότητα του έργου, οι συγγραφείς επέλεξαν μόνο τις αλλαγές του τμήματος της βάσης δεδομένων του έργου όπως αυτές ενσωματώνονται στον κορμό του έργου. Ως εκ τούτου, συνέλεξαν μόνο τις εκδόσεις των ΒΔ, που αφορούν στον κορμό ή στον κύριο κλάδο. Τα αρχεία συλλέχθηκαν τον Ιούνιο 2013. Τα αρχεία επεξεργάστηκαν από το εργαλείο ***Hecate***, που

αναπτύχθηκε από τους συγγραφείς, και ανιχνεύει τις αλλαγές τόσο σε επίπεδο χαρακτηριστικών όσο και σχέσεων. Για κάθε μετάβαση μεταξύ συνεχόμενων εκδόσεων του σχήματος, ανιχνεύει:

(α) ποιοι πίνακες εισήχθησαν και ποιοι διαγράφησαν και

(β) ποιοι πίνακες ενημερώθηκαν σε επίπεδο χαρακτηριστικών και ειδικά, για κάθε επηρεαζόμενο πίνακα, τα χαρακτηριστικά που εισάγονται, διαγράφονται, τα χαρακτηριστικά στα οποία αλλάχθηκε ο τύπος δεδομένων ή αλλάχθηκε το πρωτεύον κλειδί.

3.5.1 Σχέση Μεγέθους Σχήματος και ΄Ετους Γέννησης με την Επιβίωση

Μέγεθος σχήματος στο τέλος - Επιβίωση

Είναι ενδιαφέρον, ότι σε όλα τα σύνολα δεδομένων, η συσχέτιση του μέσου μεγέθους του σχήματος με του μέγεθος του σχήματος στο τέλος είναι πολύ ισχυρή (κατά μέσο όρο, ο συσχετισμός Kendall είναι 0,95). Το μέγεθος του σχήματος κατά τη γέννηση έχει μια συσχέτιση Kendall ***0,89*** με το μέσο μέγεθος του σχήματος και ***0,86*** σε σχέση με το μέγεθος του σχήματος στο τέλος. Όλες αυτές οι πληροφορίες σημαίνουν ότι το μέγεθος του σχήματος στο τέλος δίνει μια καλή εικόνα για τα μεγέθη ενός πίνακα.

Τα στατιστικά στοιχεία, σε 6 από τα 8 σύνολα δεδομένων είναι:

Thin Tables. (1-5 πεδία) Σε 5 από τις 7 περιπτώσεις, το ποσοστό των νεκρών πινάκων για αυτό τον αριθμό πινάκων είναι μεγαλύτερο από 60%. Αυτό σημαίνει ότι οι νεκροί πίνακες έχουν μια ισχυρή τάση να είναι "λεπτοί". Οι επιζώντες συσπειρώθηκαν επίσης με μεγάλα ποσοστά σε αυτό το εύρος (σε μία μόνο περίπτωση ήταν χαμηλότερο από τους νεκρούς). Σε 5 από τις 7 περιπτώσεις, οι πίνακες που κυμαίνονται από 1-5 πεδία έχουν μεγαλύτερη πιθανότητα να αφαιρεθούν από τη μέση πιθανότητα απομάκρυνσης στο συνόλου δεδομένων.

Wide Tables (>10 πεδία) Σε απόλυτους αριθμούς, ο αριθμός των ευρέων πινάκων που τελικά πέθαναν δεν είναι μεγαλύτερος από 4 πίνακες σε οποιαδήποτε από τα σύνολα. Η πιθανότητα αφαίρεσης ενός ευρύ πίνακα είναι αρκετά χαμηλότερη από τη μέση πιθανότητα αφαίρεσης του συνόλου δεδομένων. Οι ευρείς πίνακες

αποτελούν πολύ σημαντικό τμήμα της ομάδας των επιζώντων. Τα ποσοστά των ευρέων επιζώντων πινάκων είναι πάνω από ***73%*** για όλα τα σύνολα δεδομένων.

Αλλαγή μεγέθους σχήματος - Επιβίωση

Η κατανομή των πινάκων σε κατηγορίες με βάση τη συμπεριφορά κατά την αλλαγή μεγέθους του σχήματος είναι:

(i) καμία κλίμακα ***(NS)*** ή επίπεδοι πίνακες, οι οποίοι έχουν μια αναλογία SchemaResize ακριβώς 1,0

(ii) ScaledUp (***S-U***) με ένα λόγο SchemaResize αυστηρά υψηλότερο από 1.0

(iii) ScaledDown (***S - D***) πίνακες, με ένα λόγο SchemaResize αυστηρά χαμηλότερο από 1.0

Τα ποσοστά που υπολογίσθηκαν στα σύνολα δεδομένων είναι:

- Ο μέσο όρο των NS πινάκων στα σύνολα δεδομένων είναι ***63%***. Σχεδόν 2 στους 3 πίνακες δεν έχουν κάποια τροποποίηση του μεγέθους του σχήματος τους.
- Ο μέσο όρο των S-U πινάκων είναι ***32%***
- Ο μέσο όρο των S-D πινάκων είναι ***5%***

Παρατηρήθηκε ότι:

- Στην πολύ μεγάλη πλειοψηφία τους, οι πίνακες που μεγαλώνουν (S-U), είναι επιζώντες, με μέσο όρο ***88%***.
- Στην αντίστροφη ερώτηση, δηλαδή, ποιο ποσοστό των επιζώντων αυξάνεται, η μέση τιμή είναι ***37%*** (το ποσοστό των S-U πινάκων επί των επιζώντων).
- Περίπου ***5%*** κατά μέσο όρο των επιζώντων πινάκων είναι οι S-D

Η αλλαγή μεγέθους του σχήματος είναι σπάνια, όμως οι πίνακες στους οποίους αυξάνεται το μέγεθος τους, είναι επιζώντες.

Έτος γέννησης - Επιβίωση

Η έρευνα στα σύνολα δεδομένων έδειξε ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των νεκρών και των επιζώντων πινάκων όσον αφορά το έτος γέννησής τους. Υπάρχουν μικρές εξαιρέσεις αλλά η συνολική διαμόρφωση δείχνει σαφώς ότι μοιράζονται την ίδια ποσοστιαία κατανομή ανά έτος γέννησης και οι νεκροί και οι επιζώντες πίνακες.

3.5.2 Σχέση Διάρκειας και Δραστηριότητας με την Επιβίωση

Από προηγούμενες μελέτες παρατηρήθηκε ότι:

- όταν η διάρκεια σχετίζεται με το ποσό των ενημερώσεων που έγιναν σε κάθε πίνακα, υπάρχει μια έντονη συγκέντρωση μεγάλων αλλαγών σε μεγάλες διάρκειες και μια έλλειψη μεσαίου μεγέθους ενημερώσεων σε μεσαίες διάρκειες. Αυτό συνοψίστηκε ως το πρότυπο αντίστροφου Γ.
- Μια δεύτερη σχετική παρατήρηση είναι στο empty triangle pattern που σχετίζεται η διάρκεια με τη γέννηση, παρατηρήθηκε η απουσία νεκρών πινάκων σε μεσαίες ή υψηλές διάρκειες (που σημαίνει ότι απομακρύνθηκαν σύντομα μετά από τη γέννησή τους)

Παρακάτω γίνεται από τους συγγραφείς μια προσπάθεια ώστε να εξηγηθεί με ακρίβεια πώς συμπεριφέρεται η διάρκεια σε σχέση με την επιβίωση.

Σχέση της Δραστηριότητας με την Επιβίωση

Υπολογίστηκε για τους πίνακες η τιμή *Activity Class* και με βάση αυτό ταξινομήθηκαν οι πίνακες σε 3 κατηγορίες:

(α) άκαμπτοι (rigid), χωρίς μεταβολή καθ ’όλη τη διάρκεια της ζωής τους

(β) ενεργοί (active), εάν έχουν ποσοστό ATU άνω του 0,1 και 5 αλλαγές στο σύνολο της ζωής τους

(γ) ήσυχοι (quiet)

Για να ανακαλυφθεί πώς σχετίζεται η δραστηριότητα ενός πίνακα με την επιβίωσή του, (πιθανότητα θανάτου/επιβίωσης) για καθεμία από τις 3 κατηγορίες δραστηριοτήτων (άκαμπτη, ήσυχη και ενεργή), υπολογίστηκε η αντίστοιχη πιθανότητα για τη συνολική πιθανότητα θανάτου/επιβίωσης επί του αντίστοιχου συνόλου δεδομένων.

Κατηγορίες Δραστηριότητας:

• *Άκαμπτοι* πίνακες: παρουσιάζουν (συχνά σημαντικά) υψηλότερες πιθανότητες να αφαιρεθούν σε σχέση με τους υπόλοιπους πίνακες. Ενώ υπάρχει μεγαλύτερη πιθανότητα από τον μέσο όρο για να πεθάνει ένας άκαμπτος πίνακας, υπάρχει μικρότερη πιθανότητα από τον μέσο όρο για να επιβιώσει ένας άκαμπτος πίνακας.

• Όταν πρόκειται για *ήσυχους* πίνακες (που αποτελούν την πλειοψηφία του πληθυσμού), σε 6 από τα 8 σύνολα δεδομένων υπάρχει μείωση της πιθανότητας

θανάτου (αντίστοιχα: αύξηση της πιθανότητας επιβίωσης) σε σύγκριση με τη μέση πιθανότητα θανάτου του συνόλου δεδομένων και η διαφορά αυτή υπερβαίνει το 10% σε 4 περιπτώσεις.

- Οι *ενεργοί* πίνακες έχουν μεικτή συμπεριφορά

Διαφορές στη Δραστηριότητα μεταξύ νεκρών και επιζώντων

Για να αξιολογηθεί αν η δραστηριότητα των νεκρών πινάκων είναι διαφορετική από την αντίστοιχη των επιζώντων, έγινε κατανομή των πινάκων ανά επιβίωση και κατηγορία δραστηριότητας.

Παρατηρήθηκαν 2 κύριες ομάδες σχημάτων:

- Η πρώτη ομάδα περιλαμβάνει σχήματα που είναι πολύ άκαμπτα, με πάρα πολλούς άκαμπτούς πίνακες τόσο για τους νεκρούς (όπου η ακαμψία είναι πάνω από ***50%*** για το σύνολο) όσο και για τους επιζώντες (με ποσοστό μεταξύ 30% και 47%).
- Η δεύτερη ομάδα περιλαμβάνει σύνολα δεδομένων με τους νεκρούς πίνακες να είναι κυρίως άκαμπτοι (οι μισοί από τους νεκρούς) και πάρα πολλοί επιζώντες πίνακες να είναι ήσυχοι.

Αιφνίδιοι θάνατοι:

- άκαμπτο πίνακες με ποσοστό από 46% - 100% (με 1 εξαίρεση)
- ήσυχοι πίνακες με 20% -40%
- ενεργές με 0% -22%.

Επιζώντες:

- ήσυχοι (σε 6 από τα 8 σύνολα δεδομένων) με ποσοστό μεταξύ 48% -90
- ενεργοί με 4% - 50%
- Άκαμπτοι με 4% -54%

Αντίθετα λοξές διάρκειες (Oppositely Skewed Durations)

Η διάρκεια των πινάκων κατανεμήθηκε σε διαστήματα (buckets) 10 εκδόσεων, δημιουργώντας ένα ιστόγραμμα διαρκειών. Αντιστοιχήθηκε ο αριθμός των νεκρών και επιζώντων πινάκων που ανήκουν σε κάθε bucket. Οι συγγραφείς παρατήρησαν ένα φαινόμενο, το οποίο το ονόμασαν ***oppositely skewed durations or opposite skews pattern.***

Οι νεκροί πίνακες έχουν έντονη παρουσία για σύντομες διάρκειες (left - heavy), συχνά με πολύ μεγάλα ποσοστά τα οποία απομακρύνονται πολύ σύντομα

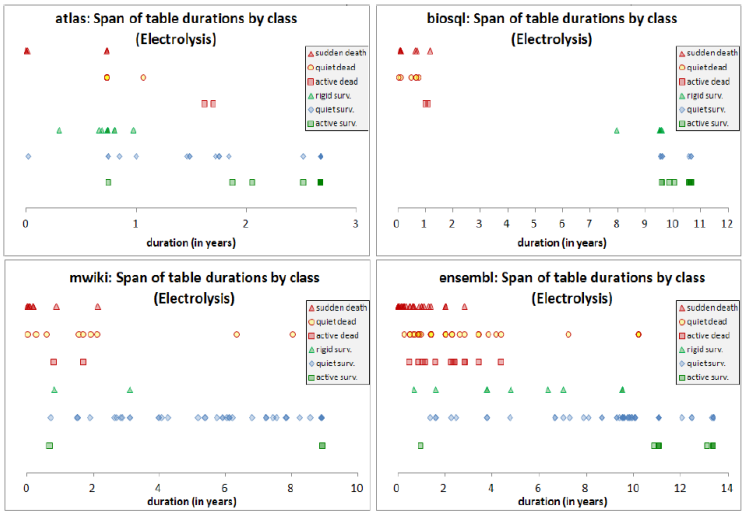
μετά τη γέννηση. Την αντίθετη συμπεριφορά παρουσιάζουν οι επιζώντες πίνακες (right-heavy).

Η τάση να έχουν σύντομες διάρκειες οι διαγραμμένοι πίνακες, αποδίδεται, στο κόστος που έχουν οι διαγραφές για τη συντήρηση του λογισμικού που περιβάλλει τη βάση δεδομένων. Όσο νωρίτερα αφαιρείται ένας πίνακας, τόσο μικρότερο είναι το κόστος της διατήρησης του περιβάλλοντος κώδικα. (***Βαρύτητα στην ακαμψία***)

Αυτό το γεγονός, σε συνδυασμό με το ότι οι αρχικές εκδόσεις της βάσης δεδομένων περιλαμβάνουν ήδη ένα μεγάλο ποσοστό του συνολικού αριθμού των πινάκων, έχει ως αποτέλεσμα (για 6 από τα 8 σύνολα δεδομένων), οι διάρκειες των επιζώντων πινάκων να υπερβαίνουν το 45% στο τελικό cluster.

Το πρότυπο ηλεκτρόλυσης

Το πρότυπο της ηλεκτρόλυσης συνδέει τη ***διάρκεια*** με τη ***δραστηριότητα***. Ομαδοποιεί τους πίνακες σύμφωνα με τη κλάση *LifeAndDeath* η οποία εκφράζει το προφίλ ενός πίνακα συνδυάζοντας τους δύο τομείς {dead, survivor} × {άκαμπτο, ήσυχο, ενεργό} σε καρτεσιανές συντεταγμένες. Κάθε σημείο του γραφήματος αναφέρεται σε έναν πίνακα με (α) τη διάρκεια του στο x-άξονα και (β) την κλάση LifeAndDeath (συμπεριλαμβανομένων της επιβίωσης και της δραστηριότητας) στον άξονα y.



Σχήμα 3-10: The Electrolysis pattern. Each point refers to a table with (a) its duration at the x-axis and (b) its LifeAndDeath class at the y-axis (also its symbol).
Points-από το VaZa17 με άδεια των συγγραφέων–σελ 18

Πρότυπο ηλεκτρόλυσης: Οι νεκροί πίνακες δείχνουν πολύ μικρότερη διάρκεια ζωής από τους επιζήσαντες και μπορούν να εντοπιστούν με σύντομες ή μεσαίες διάρκειες, και όχι σε μεγάλες διάρκειες. Με λίγες εξαιρέσεις, οι λιγότερο ενεργοί νεκροί πίνακες έχουν μεγαλύτερη πιθανότητα να φτάσουν σε μικρότερες διάρκειες. Οι επιζώντες εκθέτουν την αντίστροφη συμπεριφορά, δηλαδή βρίσκονται ως επί το πλείστον σε μεσαίες ή υψηλές διάρκειες. Όσο μεγαλύτερη δραστηριότητα παρουσιάζουν οι επιζήσαντες, τόσο εντονότερα οδηγούνται σε υψηλές διάρκειες.

Τα κυριότερα σημεία του σχεδίου:

- Η πλήρης απουσία νεκρών πινάκων σε υψηλές διάρκειες.
- Η συσσώρευση άκαμπτων νεκρών σε χαμηλές διάρκειες, η εξάπλωση των ήσυχων νεκρών πινάκων σε χαμηλές ή μεσαίες διάρκειες και τη περιστασιακή παρουσία των λίγων ενεργών νεκρών, που βρίσκονται επίσης σε χαμηλές και μεσαίες διάρκειες, αλλά με συσσωρευμένο τρόπο.
- Η ακραία συσσώρευση ενεργών επιζώντων σε υψηλές διάρκειες.
- Η ευρύτερη εξάπλωση των (αρκετά) ήσυχων επιζώντων σε μεγάλο εύρος διαρκειών.
- Η εξάπλωση των άκαμπτων επιζώντων σε όλα τα είδη των διαρκειών (συχνά, όχι τόσο ψηλά όσο οι ήσυχοι και οι ενεργοί επιζώντες).

Οι <u>άκαμπτοι νεκροί</u> πίνακες είναι η πιο πυκνοκατοικημένη ομάδα της κατηγορίας των νεκρών πινάκων, αλλά έχουν τη συντομότερη δυνατή έκταση.

Οι <u>άκαμπτοι επιζώντες</u>, που είναι η 2^{η} πιο πυκνοκατοικημένη κατηγορία του συνόλου του πληθυσμού, παρουσιάζουν όλα τα είδη συμπεριφοράς. Ωστόσο, στις περισσότερες περιπτώσεις, είναι δυσανάλογα συγκεντρωμένες και δεν εξαπλώνονται σε όλες τις κατηγορίες.

Οι <u>ενεργά επιζώντες</u> είναι επίσης δυσανάλογα τοποθετημένοι σε μεγάλες διάρκειες. Συνολικά, με εξαίρεση τους ήσυχους επιζώντες που πράγματι εκτείνονται σε μια μεγάλη διακύμανση των διαρκειών, στις υπόλοιπες κατηγορίες, η διασπορά του χρόνου είναι δυσανάλογη με το μέγεθος του πληθυσμού (αριθμός σημείων στο γράφημα) της αντίστοιχης κατηγορίας.

Διεξοδική μελέτη των διαρκειών

Εκφράστηκαν οι διάρκειες των πινάκων ως ποσοστά κατά τη διάρκεια ζωής του σχήματος. Στη συνέχεια, για καθένα υπολογίστηκε η τιμή LifeAndDeath και για

κάθε χρονική περίοδο που αντιστοιχεί σε 5% της διάρκειας της ΒΔ, υπολογίστηκε το ποσοστό των πινάκων των οποίων η διάρκεια εμπίπτει σε αυτό το εύρος. (σχήμα 14 του [VaZa17] –σελ. 20).

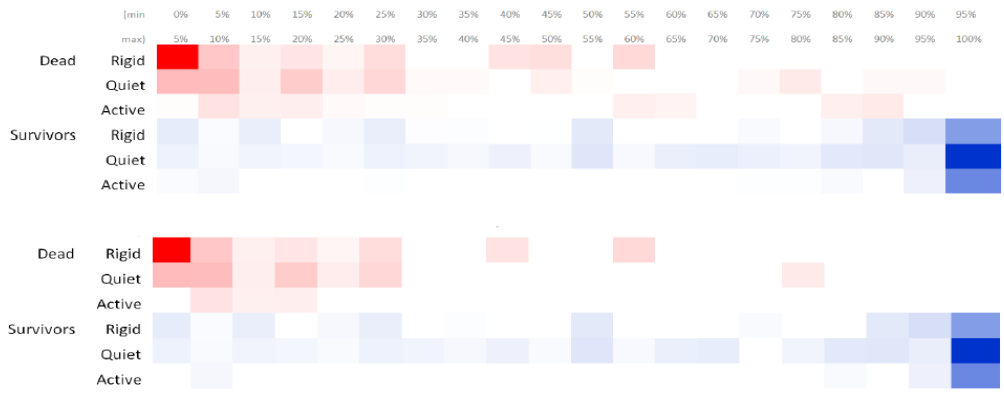
Από την παραπάνω διαδικασία εντοπίστηκαν τα δεδομένα σε 3 χρονικά διαστήματα:

(α) διάρκειες μικρότερες από το ***20%*** της βάσης δεδομένων (που προσελκύει μεγάλο αριθμό νεκρών πινάκων, κυρίως, τους άκαμπτους), δηλαδή χαμηλές διάρκειες.

(β) διάρκειες υψηλότερες από το 80% της διάρκεια ζωής της βάσης δεδομένων (όπου εντοπίζονται πάρα πολλοί επιζώντες, κυρίως οι ενεργοί), δηλαδή υψηλές διάρκειες.

(γ) υπόλοιπες διάρκειες που σχηματίζουν μια ενδιάμεση κατηγορία, τις μεσαίες διάρκειες.

Στη συνέχεια κατασκευάστηκε ένας χάρτης θερμότητας (Εικόνα 15 του [VaZa17] – σελ. 20) για τα παραπάνω δεδομένα. Η ένταση κάθε χρώματος υποδηλώνει πόσο μεγάλο είναι το αντίστοιχο ποσοστό.



Σχήμα 3-11 : Ηλεκτρόλυση-όλα-σε-ένα. Χάρτης θερμότητας για τη συμμετοχή κάθε τιμής LifeAndDeath σε μια ορισμένη περιοχή διάρκειας, κατά μέσο όρο για όλα τα 8 σύνολα δεδομένων (πάνω: αμιγή νούμερα, κάτω: χωρίς τις ακραίες τιμές) - από το VaZa17 με άδεια των συγγραφέων – σελ 20

Παρατηρήσεις και ευρήματα

Τα ποσοτικά μας ευρήματα του σχεδίου ηλεκτρόλυσης είναι:

Νεκροί Πίνακες. Σχεδόν οι μισοί νεκροί πίνακες είναι άκαμπτοι και παρουσιάζουν μικρές διάρκειες.

Η έλξη των νεκρών πινάκων, ιδιαίτερα άκαμπτων, σε (κυρίως) χαμηλές ή (δευτερευόντως) μεσαίες διάρκειες είναι σημαντική και μόνο λίγοι πίνακες στην κατηγορία των νεκρών πινάκων διαφεύγουν αυτού του κανόνα.

- Οι ***άκαμπτοι*** νεκροί πίνακες, είναι η πολυπληθέστερη κατηγορία νεκρών πινάκων, ομαδοποιημένοι στην κατηγορία των χαμηλών διαρκειών (χαμηλότερο από το 20% της διάρκειας ζωής της βάσης δεδομένων) με ποσοστά ***90%*** - ***100%*** σε 3 από τα 6 σύνολα δεδομένων. Ακολουθεί ο άτλας με ένα μεγάλο ποσοστό ***57%*** σε αυτό το εύρος. Υπάρχουν δύο εξαιρέσεις: opencart και typo3, έχοντας τους περισσότερους από τους νεκρούς πίνακες στο μεσαίο εύρος.
- Οι ***ήσυχοι*** νεκροί πίνακες, που είναι μια κατηγορία που αποτελείται από λίγους πίνακες, βρίσκονται κυρίως στην περιοχή χαμηλής διάρκειας για τα 5 σύνολα δεδομένων. (Η ΒΔ *Atlas* έχει 100% των πινάκων στο μεσαίο εύρος και οι νεκροί ήσυχοι πίνακες της ΒΔ *phpBB* είναι χωρισμένοι στη μεσαία και στη μεγάλη διάρκεια).
- Οι ***ενεργοί*** νεκροί πίνακες, είναι μια κατηγορία που εμφανίζεται στα 6 από τα 8 σύνολα δεδομένων έστω και με 1 μόνο πίνακα. Δύο από αυτά έχουν 100% συγκέντρωση και ένα ***67%*** του πληθυσμού του σε χαμηλές διάρκειες. Για τα υπόλοιπα, η ΒΔ Atlas έχει το 100% των ενεργών νεκρών πινάκων στη μεσαία περιοχή, η phpBB έχει το ***100%*** των ενεργών νεκρών πινάκων στην περιοχή μεγάλων διαρκειών (σημειώνεται ότι η phpBB έχει μια ξεχωριστή συμπεριφορά) και η ΒΔ typo3 χωρίζεται στη μέση μεταξύ χαμηλών και μεσαίων διαρκειών.

Επιζώντες Πίνακες. Οι επιζώντες έχουν την αντίθετη τάση της ομαδοποίησης σε σύγκριση με τους νεκρούς. Έτσι, υπάρχουν αρκετές περιπτώσεις όπου οι επιζώντες πίνακες φτάνουν σε πολύ υψηλές συγκεντρώσεις στις υψηλές διάρκειες και όσο πιο ενεργοί είναι οι πίνακες, τόσο υψηλότερη συσσώρευσή έχουν σε μεγάλες διάρκειες.

- Οι ***άκαμπτοι*** επιζώντες επιδεικνύουν μια μεγάλη ποικιλία συμπεριφορών. Είναι η δεύτερη πιο πολυσύχναστη κατηγορία πινάκων μετά από τους ήσυχους επιζώντες και επιδεικνύουν πάρα πολλά προφίλ ομαδοποίησης.

- Οι ***ήσυχοι*** επιζώντες, που είναι μεγάλη πλειοψηφία των επιζώντων, παρουσιάζουν ως επί το πλείστον μεγάλες διάρκειες και δευτερευόντως μεσαίες. Σε 6 από 8 σύνολα δεδομένων, το ποσοστό των ήσυχων επιζώντων πινάκων που υπερβαίνουν το ***80%*** της διάρκειας ζωής της ΒΔ υπερβαίνει το 50%. Στις δύο εξαιρέσεις, οι μεσαίες διάρκειες είναι οι μεγαλύτερες υποομάδα των ήσυχων επιζώντων πινάκων. Ιδιαίτερα, σε όλα τα σύνολα δεδομένων, υπάρχουν ήσυχοι επιζώντες που φτάνουν στη μέγιστη διάρκεια.
- Είναι εξαιρετικά εκπληκτικό το γεγονός ότι η μεγάλη πλειοψηφία των ***ενεργών*** επιζώντων υπερβαίνει το ***80%*** της διάρκειας ζωής της ΒΔ σε όλα τα σύνολα δεδομένων. Σε τρία συνόλων δεδομένων κυμαίνονται από ***67%*** -***75%*** ενώ σε 2 σύνολα δεδομένων αγγίζουν το απόλυτο). Οι ενεργοί επιζώντες πίνακες δεν είναι πάρα πολλοί, ωστόσο, η συσσώρευση τους στις υψηλές διάρκειες (που υποδηλώνει πρόωρη γέννηση) είναι εκπληκτική (νεογέννητοι και επιζώντες).

Απουσία εξέλιξης. Αν και η πλειοψηφία των επιζώντων πινάκων βρίσκονται στην ήσυχη τάξη, μπορούμε να πούμε με έμφαση ότι κυριαρχεί η απουσία της εξέλιξης. Οι επιζώντες ξεπερνούν κατά πολύ τους διαγραμμένους πίνακες. Ομοίως, οι άκαμπτοι πίνακες ξεπερνούν τους ενεργούς, τόσο στην επιβίωση όσο και στην κατηγορία των νεκρών. Το μέγεθος του σχήματος σπάνια αλλάζει μέγεθος και μόνο στους επιζώντες πίνακες. Οι ενεργοί πίνακες είναι λίγοι και κυρίως γεννήθηκαν σε πρώιμες φάσεις του κύκλου ζωής της βάσης δεδομένων.

ΚΕΦΑΛΑΙΟ 4

ΕΠΙΒΙΩΣΗ ΤΩΝ WIDE ΠΙΝΑΚΩΝ-ΜΟΤΙΒΟ ΓΑΜΜΑ

Η παρούσα εργασία στηρίζεται στο [Vass21] και τα δεδομένα που το συνοδεύουν. Στο [Vass21] συλλέχθηκαν 195 σύνολα δεδομένων, Για κάθε σύνολο δεδομένων (project) καταγράφηκαν πληροφορίες μεταξύ των οποίων περιλαμβάνονται τα ακόλουθα (α) η ημερομηνία γέννησης και θανάτου των πινάκων, από τις οποίες υπολογίζεται η διάρκεια ζωής του κάθε πίνακα, (β) πληροφορίες για το μέγεθος του σχήματός τους (τελευταία μέτρηση και μέσος όρος μεγέθους), καθώς και ο λόγος τελικού /αρχικό μεγέθους, γ) μετρήσεις σχετικά με την δραστηριότητα του πίνακα όπως το άθροισμα των αλλαγών, το πλήθος των εκδόσεων στις οποίες έγιναν αλλαγές, ο λόγος που αντιστοιχεί στο άθροισμα ενημερώσεων προς τη διάρκεια (το οποίο ονομάζεται ATU), και, δ) κατηγοριοποίηση των πινάκων με κριτήριο την επιβίωση σε Survival Class (επιζών ή νεκρός), Activity Class, και ο συνδυασμός τους ως Life And Death Class (LAD)

Οι πίνακες χαρακτηρίζονται με βάση το *Survival Class*, σε επιζώντες ***(survived)*** αν υφίστανται στην τελευταία γνωστή έκδοση ολόκληρου του σχήματος, αλλιώς χαρακτηρίζονται ως νεκροί ***(not survived or dead)***.

Με βάση το Activity Class ένας πίνακας είναι ***active*** όταν ο λόγος ATU (άθροισμα ενημερώσεων / διάρκεια) είναι υψηλότερος από 0,1 και έχει περισσότερες από 5 ενημερώσεις, ***rigid*** (άκαμπτος) εάν δεν υπάρχουν ενημερώσεις σε αυτόν και ***quiet*** (αθόρυβος) σε κάθε άλλη περίπτωση.

Η απεικόνιση των πινάκων ως προς το μέγεθος του σχήματος κατά τη γέννηση, ταξινομεί τους πίνακες που είναι μεγαλύτεροι από 10 χαρακτηριστικά σε φαρδιούς ***(wide)*** *και τους υπόλοιπους ως μη φαρδιούς* ***(not wide).***

Στη μελέτη που ακολουθεί ερευνάται *σε πόσα από τα 195 σύνολα δεδομένων μπορεί και έχει νόημα να μελετηθεί το μοτίβο Γάμμα*. Το μοτίβο Γάμμα στηρίζεται στη σχέση του μεγέθους ενός πίνακα κατά τη γέννηση του και της διάρκειας του πίνακα. Συμπεριλήφθηκαν, λοιπόν, στη μελέτη τα projects με βάση τα παρακάτω:

(α) πρέπει το project να περιέχει *wide* πίνακες. Σε περίπτωση που ένα project δεν περιλαμβάνει *wide* πίνακες, η δοκιμή του προτύπου Γ δεν έχει νόημα (η αξιολόγηση του προτύπου δεν ισχύει, καθώς και οι δύο τιμές, (1) μέγεθος και (2) διάρκεια wide πινάκων είναι μηδέν). Επομένως, για να έχει νόημα η δοκιμή, πρέπει να εξαιρεθούν τα projects που δεν περιέχουν wide πίνακες. Από τα 195 projects, τα 105 δεν περιείχαν κανένα μεγάλο πίνακα (54%) και εξαιρέθηκαν. Έτσι έμειναν για μελέτη 90 projects δηλαδή το 46%.

(β) πρέπει επίσης να υπάρχουν *dead* πίνακες στο project. Αυτό γιατί σε ένα σύνολο δεδομένων που δεν έχει dead πίνακες, η πιθανότητα επιβίωσης των wide πινάκων και η πιθανότητα επιβίωσης των not wide είναι 100%. Άρα δεν έχει νόημα η σύγκριση μεταξύ τους αφού και οι δύο πιθανότητες είναι 100%. Από τα 90 σύνολα δεδομένων που έμειναν από το (α), υπάρχουν 55 που έχουν wide πίνακες και δεν απεβίωσε κανένας, αλλά αυτά τα projects δεν έχουν πεθαμένους πίνακες σε καμιά κατηγορία.

Αυτό άφησε μόνο ***35*** projects για μελέτη.

Μια συνοπτική παρουσίαση των πληροφοριών που αφορούν τα 35 υπό μελέτη projects φαίνεται στο Πίνακα 4-1.

Πίνακας 4-1 Συνοπτική αποτύπωση των πληροφοριών των 35 συνόλων δεδομένων που συμμετέχουν στην μελέτη.

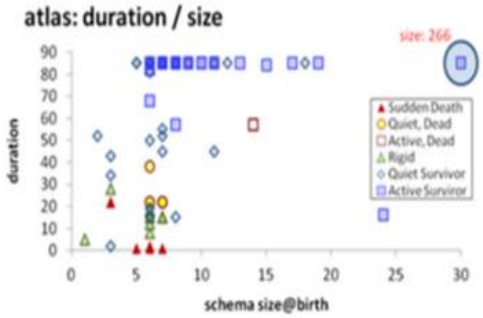
Project	Geometry?	FisherExecuted?	p-Value?	SurvWide	SurvNotWide	DeadWide	DeadNotWide	Wide	Dead	probSW	probSNW	WMLTS?	OverallClass
GoBelieveIO__im_service	FALSE	FALSE		0	2	6	8	6	14	0%	20%	FALSE	3_FocusedShot_n_LOW
keybase__node-client	FALSE	FALSE		0	2	1	2	1	3	0%	50%	FALSE	3_FocusedShot_n_LOW
nawork__nawork-uri	FALSE	FALSE		1	3	1	0	2	1	50%	100%	FALSE	2_MODERATE
nooku__joomla-todo	FALSE	FALSE		2	0	3	0	5	3	40%			4_ACTIVE
webnuts__post_json	FALSE	FALSE		1	3	1	0	2	1	50%	100%	FALSE	1_FocusedShot_n_FROZEN
blabla1337__skf-flask	TRUE	FALSE		1	18	0	6	1	6	100%	75%	TRUE	4_ACTIVE
builderscon__octav	TRUE	FALSE		2	20	0	5	2	5	100%	80%	TRUE	4_ACTIVE
energine-cmf__energine	TRUE	FALSE		3	57	0	5	3	5	100%	92%	TRUE	4_ACTIVE
EPICPaaS__appmsgsrv	TRUE	FALSE		2	21	0	1	2	1	100%	95%	TRUE	4_ACTIVE
imbo__imbo	TRUE	FALSE		1	3	0	2	1	2	100%	60%	TRUE	2_MODERATE
lamassu__lamassu-admin	TRUE	FALSE		1	4	0	2	1	2	100%	67%	TRUE	2_MODERATE
liujianping__scaffold	TRUE	FALSE		2	3	0	1	2	1	100%	75%	TRUE	1_FocusedShot_n_FROZEN
mozilla__mig	TRUE	FALSE		3	9	0	1	3	1	100%	90%	TRUE	2_MODERATE
processone__ejabberd	TRUE	FALSE		4	36	0	2	4	2	100%	95%	TRUE	4_ACTIVE
pw-press__web-project	TRUE	FALSE		1	10	0	4	1	4	100%	71%	TRUE	3_FocusedShot_n_LOW
spaceboats__busbus	TRUE	FALSE		4	6	0	5	4	5	100%	55%	TRUE	3_FocusedShot_n_LOW
symphonycms__symphony-2	TRUE	FALSE		1	19	0	1	1	1	100%	95%	TRUE	2_MODERATE
TalkingData__OWL-v3	TRUE	FALSE		7	26	0	8	7	8	100%	76%	TRUE	3_FocusedShot_n_LOW
teresko__palladium	TRUE	FALSE		1	0	0	1	1	1	100%	0%	TRUE	1_ALMOST_FROZEN
UlricQin__beego-blog	TRUE	FALSE		1	2	0	1	1	1	100%	67%	TRUE	1_ALMOST_FROZEN
yiier__forum	TRUE	FALSE		1	2	0	1	1	1	100%	67%	TRUE	1_ALMOST_FROZEN
anchorcms__anchor-cms	FALSE	TRUE	FALSE	1	9	1	9	2	10	50%	50%	FALSE	3_FocusedShot_n_LOW
cgrates__cgrates	FALSE	TRUE	FALSE	7	14	11	8	18	19	39%	64%	FALSE	4_ACTIVE
gem__oq-engine	FALSE	TRUE	FALSE	1	3	5	40	6	45	17%	7%	TRUE	1_FocusedShot_n_FROZEN
gugoan__economizzer	FALSE	TRUE	FALSE	2	6	2	9	4	11	50%	40%	TRUE	3_FocusedShot_n_LOW
hurad__hurad	FALSE	TRUE	FALSE	5	8	6	9	11	15	45%	47%	FALSE	3_FocusedShot_n_LOW
joomlatools__joomla-platform	FALSE	TRUE	FALSE	5	13	21	74	26	95	19%	15%	TRUE	4_ACTIVE
NPRA__EmissionCalculatorLib	FALSE	TRUE	FALSE	1	17	1	19	2	20	50%	47%	TRUE	1_FocusedShot_n_FROZEN
opencart__opencart	FALSE	TRUE	FALSE	14	121	13	135	27	148	52%	47%	TRUE	4_ACTIVE
foodcoopshop__foodcoopshop	TRUE	TRUE	FALSE	8	27	1	5	9	6	89%	84%	TRUE	4_ACTIVE
HaliteChallenge__Halite-II	TRUE	TRUE	FALSE	3	19	2	9	5	11	60%	68%	FALSE	4_ACTIVE
intelliants__subrion	TRUE	TRUE	FALSE	10	41	3	2	13	5	77%	95%	FALSE	4_ACTIVE
pinterest__teletraan	TRUE	TRUE	FALSE	5	17	2	16	7	18	71%	52%	TRUE	4_ACTIVE
quickapps__cms	TRUE	TRUE	FALSE	7	19	1	5	8	6	88%	79%	TRUE	4_ACTIVE
torrentpier__torrentpier	TRUE	TRUE	FALSE	11	39	2	20	13	22	85%	66%	TRUE	4_ACTIVE

Με βάση τις παραπάνω πληροφορίες, αποτυπώθηκαν για κάθε project οι παρακάτω μετρήσεις προκειμένου να ερευνηθεί αν το μοτίβο Γ ισχύει:

Geometry ? Είναι TRUE όταν η γραφική απεικόνιση σχηματίζει το μοτίβο Γ και FALSE στην αντίθετη περίπτωση. Όταν σχηματίζεται το μοτίβο Γ σημαίνει απουσία μεγάλων πινάκων με μικρή διάρκεια Η γεωμετρία ικανοποιείται όταν (α) ο αριθμός wide και survived πινάκων ξεπερνά τον αριθμό των wide και not survived πινάκων και (β) ο αριθμός των wide και not survived πινάκων είναι μικρότερος ή ίσος με 3 πίνακες. Με άλλα λόγια, πρέπει πράγματι

να υπάρχει ένας σχεδόν άδειος χώρος που να αντιστοιχεί στους wide και not survived πινάκες ώστε να κρατηθεί το μοτίβο, το οποίο δηλώνει μεγαλύτερη πιθανότητα επιβίωσης ενός wide πίνακα από την πιθανότητα να διαγραφεί. Το σχήμα Γ δηλώνει απουσία wide πινάκων με μικρή διάρκεια.

Στο Σχήμα 4-1 φαίνεται ως παράδειγμα η οπτική απεικόνιση του μοτίβου Γ στη βάση Atlas όπου στον άξονα x απεικονίζεται το μέγεθος (πλήθος attributes) του σχήματος κατά την γέννηση του πίνακα και στον άξονα y η διάρκεια ζωής. Φαίνεται στο σχήμα πως πίνακες με μικρά μεγέθη σχημάτων μπορεί να έχουν αυθαίρετες διάρκειες, ενώ οι πίνακες με μεγαλύτερα μεγέθη σχημάτων διαρκούν πολύ.



Σχήμα 4-1 : The Γ pattern -- από το [VaZS16] με άδεια των συγγραφέων – σελ 29

FisherExecuted Αν εκτελέστηκε το Fisher τεστ, το οποίο είναι ένα στατιστικό τεστ που εφαρμόστηκε στο σύνολο δεδομένων και δημιουργεί τιμές που αντιστοιχούν στον διαχωρισμό των πινάκων με βάση τα δύο χαρακτηριστικά (μέγεθος και επιβίωση). Ο διαχωρισμός αυτός γίνεται σε wide και not wide και σε survived και not survived για κάθε project. Οι τιμές αυτές τοποθετούνται σε έναν πίνακα όπου το ένα χαρακτηριστικό (μέγεθος) αποτελεί τις γραμμές του πίνακα και το άλλο χαρακτηριστικό (επιβίωση) τις στήλες του πίνακα. Το τεστ Fisher είναι κατάλληλο διότι μπορεί να δώσει αποτελέσματα ακόμη και σε μηδενικές ή πολύ μικρές τιμές. Σε όλες τις περιπτώσεις χρειάστηκε να δοκιμαστεί η ανεξαρτησία των δύο χαρακτηριστικών. Συγκεκριμένα, εκτός αν α-

	ναφέρεται διαφορετικά, η ερευνητική υπόθεση, σε όλες τις περιπτώσεις, διατυπώνεται ως δίδυμο: **H_0**: η μηδενική υπόθεση δηλώνει ότι το μέγεθος του πίνακα και η επιβίωση του, είναι μεγέθη ανεξάρτητα μεταξύ τους,. **H_1**: υπάρχει διαφορά στην επιβίωση του πίνακα ανάλογα με το μέγεθος του. Η ερευνητική υπόθεση H_0 εκφράζει την πιθανότητα τα παραγόμενα αποτελέσματα να οφείλονται στην τύχη.
p-Value	Έχει οριστεί ένα επίπεδο alpha ίσον με 5% ως όριο αποδοχής για την τιμή p-Value. Όταν η p-Value είναι κάτω από το επίπεδο alpha (5%) είναι στατιστικά αποδεκτό οποιοδήποτε αποτέλεσμα, ενώ όταν η p-Value είναι υψηλότερη από το 5% δεν μπορεί να απερρίφθη η μηδενική υπόθεση (δηλαδή ότι δεν υπάρχει διαφορά μεταξύ δύο μετρήσιμων χαρακτηριστικών ή ότι δύο δείγματα προέρχονται από τον ίδιο γενικό πληθυσμό). Αυτό οδηγεί στο συμπέρασμα ότι τουλάχιστον τα ακραία παραγόμενα αποτελέσματα που παρατηρούνται στο σύνολο δεδομένων οφείλονται πιθανώς στην τύχη. Η τιμή της p_value έχει νόημα εάν πέτυχε η εκτέλεση του fisher test.
SurvWide	Πόσοι wide πίνακες επιβίωσαν
SurvNotWide	Πόσοι not wide πίνακες επιβίωσαν
DeadWide	Πόσοι wide πίνακες διαγράφηκαν
DeadNotWide	Πόσοι not wide πίνακες διαγράφηκαν
Wide	Το πλήθος των φαρδιών πινάκων του έργου (και αυτοί που επιβίωσαν και αυτοί που διαγράφησαν)
Dead	Το πλήθος των διαγραμμένων πινάκων του έργου (και wide και not wide)
probSW	Η πιθανότητα να επιβιώσει ένας φαρδύς πίνακας
probSNW	Η πιθανότητα να επιβιώσει ένας στενός πίνακας
WMLTS?	(Wide More Likely To Survive?) TRUE όταν η πιθανότητα επιβίωσης των wide πινάκων είναι μεγαλύτερη από την πιθανότητα επιβίωσης των not wide πινάκων και FALSE στην αντίθετη περίπτωση.

OverallClass

Μια συνολική ταξινόμηση των projects με βάση την δραστηριότητα που αφορά αλλαγές πάνω στο σχήμα - [Vass21]. Στο [Vass21] παρουσιάζονται οι κατηγορίες (κλάσεις) στις οποίες ανήκουν τα data sets με βάση την δραστηριότητα που επιδεικνύουν ως προς τις αλλαγές.Οι μετρήσεις που καθορίζουν την κλάση είναι: ***(α) Active commit***: ουσιαστικές αλλαγές πάνω στο σχήμα και όχι αλλαγές που αφορούν σχολιασμό, δημιουργία δεικτών ή οτιδήποτε δεν διαφοροποιεί το σχήμα (A commit with sum(activity) > 0 attributes) ***(β) Reed***: μια "συγκεντρωμένα" μεγάλη, μαζική αλλαγή (A commit with sum(activity) >= 15 attributes) ***(γ)*** Total activity: πλήθος attributes που αλλάχθηκαν

Με βάση τα παραπάνω θα μελετηθούν οι παρακάτω κατηγορίες:

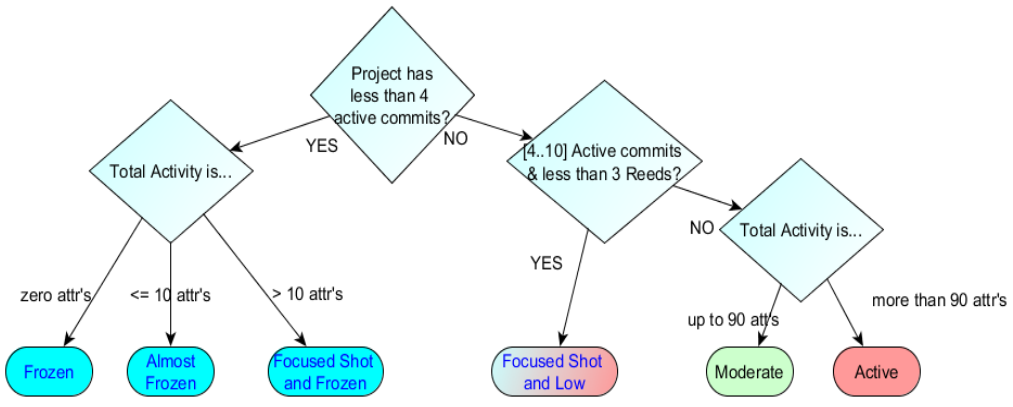
1_ALMOST_FROZEN : At most 3 active commits, totalActivity <= 10 updated attributes

1_FocusedShot_n_FROZEN : At most 3 active commits, totalActivity > 10 updated attributes

2_MODERATE : None of the rest, totalActivity <= 90 updated attributes

3_FocusedShot_n_LOW : Between 4 and 10 active commits, 1 or 2 reeds

5_ACTIVE : None of the rest, totalActivity > 90 updated attributes



Σχήμα 4-2 : Κατηγορίες των projects με βάση τα active commits και το πλήθος των attributes που αλλάχθηκαν-από το [Vass21] με άδεια των συγγραφέων

4.1 Πιθανότητα να ζήσουν οι Wide Πίνακες;

Στην συγκεκριμένη παράγραφο διερευνάται το ερώτημα : *Ποια είναι η πιθανότητα να ζήσουν οι wide πίνακες;*

Στον Πίνακα 4-2 φαίνεται ο συνολικός αριθμός των wide πινάκων (*sumWide*), ο αριθμός των wide πινάκων που επιβίωσαν (*survWide*) και το ποσοστό όσων πινάκων επιβίωσαν (*probSW*). Παρατηρούμε τα εξής :

- Στα 28 projects οι wide πίνακες έχουν πάνω από 50% πιθανότητα να επιβιώσουν και μόνο στα 7 το αντίθετο.
- Στο σύνολο των 35 projects που μελετήθηκαν, οι wide πίνακες επιβίωσαν σε ποσοστό 100% σε 16 από αυτά. (ποσοστό 47%). Αναλυτικά ανά κατηγορία τα ποσοστά επιβίωσης όλων των wide πινάκων είναι :
    - 1_ALMOST_FROZEN : 3 στα 3 (100%)
    - 1_FocusedShot_n_FROZEN : 1 στα 4 (25%)
    - 2_MODERATE : 4 στα 5 (80%)
    - 3_FocusedShot_n_LOW : 3 στα 8 (37,5%)
    - 5_ACTIVE : 5 στα 15 (33,3%)

Πίνακας 4-2 Πιθανότητα επιβίωσης των wide πινάκων ανά project

Project	OverallClass	sumWide	SurvWide	probSW
teresko__palladium	1_ALMOST_FROZEN	1	1	100%
UlricQin__beego-blog		1	1	100%
yiier__forum		1	1	100%
webnuts__post_json	1_FocusedShot_n_FROZEN	2	1	50%
liujianping__scaffold		2	2	100%
gem__oq-engine		6	1	17%
NPRA__EmissionCalculatorLib		2	1	50%
nawork__nawork-uri	2_MODERATE	2	1	50%
imbo__imbo		1	1	100%
lamassu__lamassu-admin		1	1	100%
mozilla__mig		3	3	100%
symphonycms__symphony-2		1	1	100%
GoBelieveIO__im_service	3_FocusedShot_n_LOW	6	0	0%
keybase__node-client		1	0	0%
pw-press__web-project		1	1	100%
spaceboats__busbus		4	4	100%
TalkingData__OWL-v3		7	7	100%
anchorcms__anchor-cms		2	1	50%
gugoan__economizzer		4	2	50%
hurad__hurad		11	5	45%
nooku__joomla-todo	4_ACTIVE	5	2	40%
blabla1337__skf-flask		1	1	100%
builderscon__octav		2	2	100%
energine-cmf__energine		3	3	100%
EPICPaaS__appmsgsrv		2	2	100%
processone__ejabberd		4	4	100%
cgrates__cgrates		18	7	39%
joomlatools__joomla-platform		26	5	19%
opencart__opencart		27	14	52%
foodcoopshop__foodcoopshop		9	8	89%
HaliteChallenge__Halite-II		5	3	60%
intelliants__subrion		13	10	77%
pinterest__teletraan		7	5	71%
quickapps__cms		8	7	88%
torrentpier__torrentpier		13	11	85%

Παρατηρήσεις :

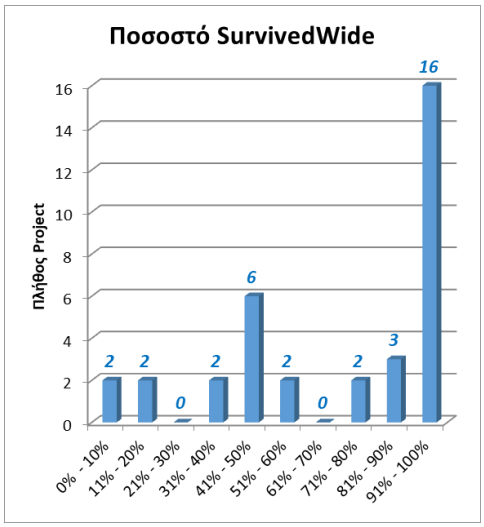
- Στα projects της κατηγορίας 1_ALMOST_FROZEN παρατηρούμε ότι οι πιθανότητα επιβίωσης είναι 100% κάτι που εξηγείται αφού συμβαίνουν πολύ λίγες αλλαγές και τα project πλησιάζουν στο να "παγώσουν".
- Σε αντίθεση στα projects της κατηγορίας 1_FocusedShot_n_FROZEN το ποσοστό επιβίωσης είναι πολύ μικρό και αυτό ίσως οφείλεται στην προσπάθεια να γίνουν αλλαγές ώστε να διατηρηθούν τα projects σε χρήση και να καλύψουν τις νέες απαιτήσεις των χρηστών.
- Για τις κατηγορίες 3_FocusedShot_n_LOW και 5_ACTIVE παρατηρούμε ότι γίνονται πολλές αλλαγές και κατά συνέπεια διαγράφονται και wide πίνακες αφού σε λίγα projects (μόνο 8) επιβιώνουν κατά 100% όλοι οι wide πίνακες.
- Στο σύνολο των 35 project υπάρχουν και 2 project στα οποία δεν επιβίωσε

κανένας wide πίνακας. Το ένα είχε 6 wide πίνακες που διαγράφηκαν όλοι ενώ τα άλλο είχε μόλις έναν wide πίνακα και ο οποίος διαγράφηκε. Και τα δύο αυτά project ανήκουν στην κατηγορία *3_FocusedShot_n_LOW*.

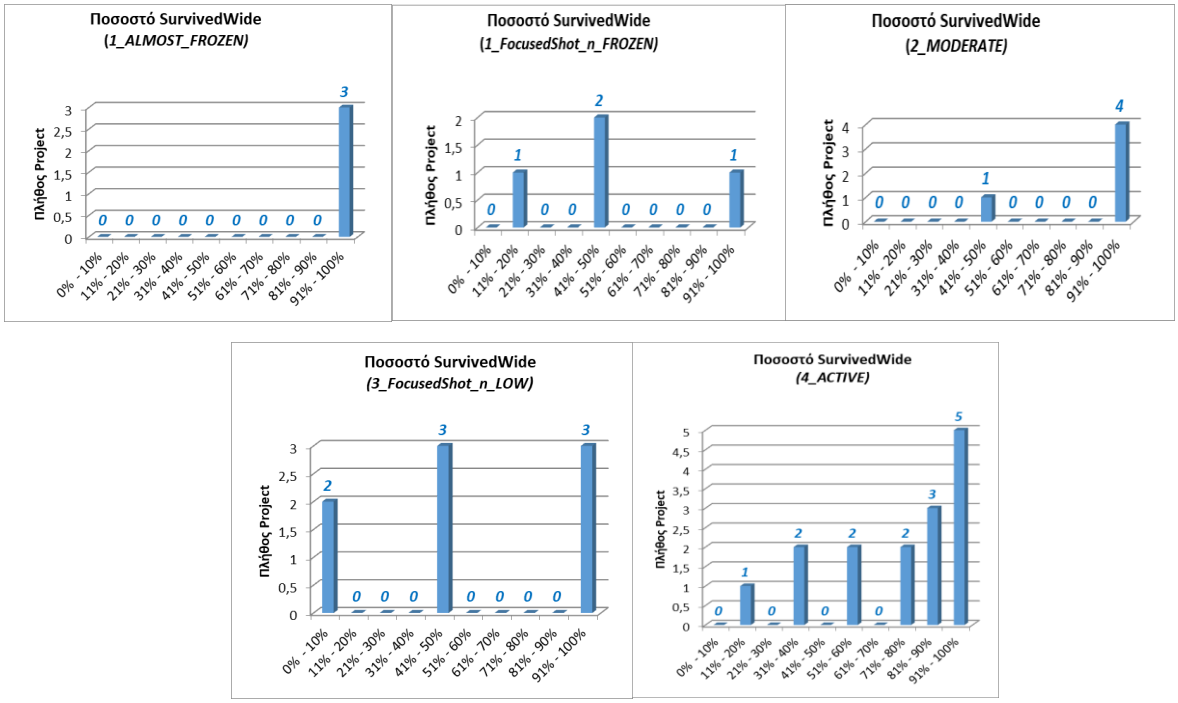
- Παρατηρούμε επίσης ότι τα μικρότερα ποσοστά επιβίωσης εμφανίζονται σε projects με μεγάλο πλήθος wide πινάκων. Χαρακτηριστικό παράδειγμα είναι τα δύο projects joomlatools_joomla-platform και cgrates_cgrates που ανήκουν στην κατηγορία 5_ACTIVE και το πρώτο έχει ποσοστό επιβίωσης των wide πινάκων 19% και το δεύτερο 39%. Στο πρώτο διαγράφηκαν 21 wide πίνακες και στο άλλο 11.

Προκειμένου να ερευνηθεί αν το πλήθος των projects με λίγους διαγραμμένους wide πίνακες υπερισχύει, δημιουργήθηκαν 10 διαστήματα στην περιοχή από το 0% έως το 100%. Για το κάθε διάστημα μετρήθηκε το πλήθος των συνόλων δεδομένων που το ποσοστό επιβίωσης των wide πινάκων ανήκει σε αυτό το διάστημα. Για παράδειγμα αν ένα σύνολο δεδομένων έχει ποσοστό επιβίωσης των wide πινάκων 77% τότε ανήκει στο διάστημα 71% με 80%.

Στα ιστογράμματα που φαίνονται στα σχήματα 4-3 και 4-4 (για το σύνολο των projects αλλά και για κάθε κατηγορία ξεχωριστά) εμφανίζεται το πλήθος των project για κάθε διάστημα που αντιστοιχεί στην πιθανότητα επιβίωσης των wide πινάκων. Στον άξονα x εμφανίζονται τα διαστήματα και στον άξονα y τα πλήθη των projects που καταμετρήθηκαν για το συγκεκριμένο διάστημα. Στο ιστόγραμμα που αντιστοιχεί στο σύνολο των projects φαίνεται πως σε 16 projects η πιθανότητα επιβίωσης είναι από 91% μέχρι 100% και μόνο σε δύο projects η πιθανότητα επιβίωσης είναι από 0% έως 10%. Σε όλες τις κατηγορίες (με εξαίρεση την κατηγορία 1_FocusedShot_n_FROZEN) η μεγαλύτερη τιμή εμφανίζεται στο διάστημα 91% - 100%.



Σχήμα 4-3 : Πλήθη των συνολικών projects για κάθε διάστημα (άξονας y) που αντιστοιχεί στη πιθανότητα επιβίωσης των wide πινάκων (άξονας x)



Σχήμα 4-4 : Πλήθη των projects για κάθε διάστημα (άξονας y) που αντιστοιχεί στη πιθανότητα επιβίωσης των wide πινάκων (άξονας x) ανά κατηγορία.

Προκειμένου να μελετηθεί η συμπεριφορά των wide πινάκων κατά την διάρκεια της ζωής μιας βάσης και κατά πόσο έχουν μια ισχυρή πιθανότητα να επιβιώσουν, ερευνάται και το υποερώτημα *αν οι wide πίνακες που τελικά επιβιώνουν είναι περισσότεροι από αυτούς που διαγράφονται.*

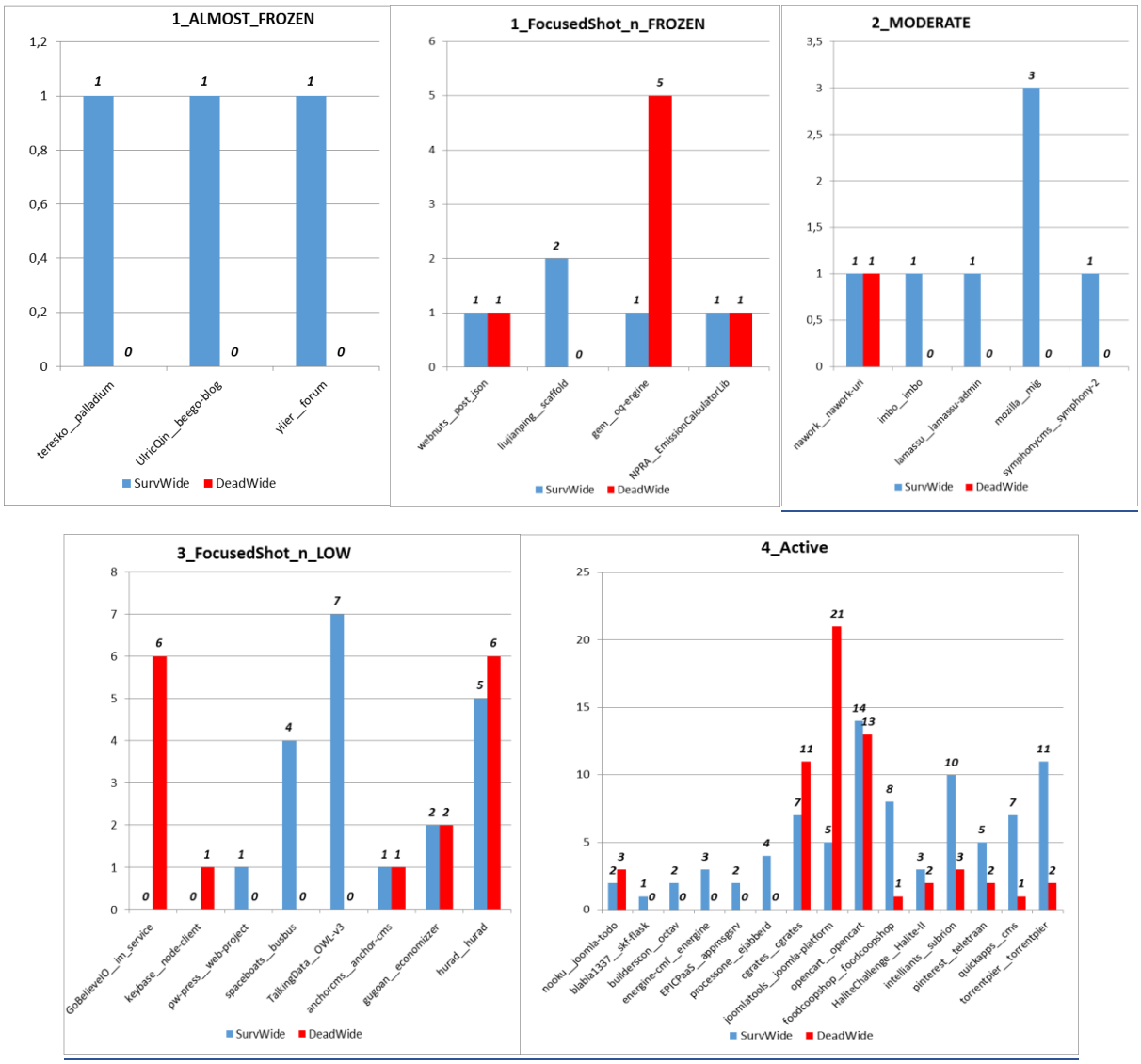
Από τα 35 συνολικά projects, που μελετήθηκαν, στα 23 projects το πλήθος των wide πινάκων που επιβίωσαν είναι μεγαλύτερο από το πλήθος των wide πινάκων που πέθαναν (ποσοστό ***65,7***), στα 5 projects τα 2 πλήθη είναι ίσα (ποσοστό ***14,3***) και τέλος στα 7 από τα 35 είναι μεγαλύτερη η πιθανότητα να διαγραφούν οι wide πίνακες (ποσοστό ***20%***). Τα ποσοστά επιβίωσης των wide πινάκων ανά κατηγορία είναι :

- 1_ALMOST_FROZEN : 3 στα 3 (100%)
- 1_FocusedShot_n_FROZEN : Σε 1 project από τα 4 το πλήθος των wide που επιβίωσαν είναι μεγαλύτερο από αυτούς που πέθαναν (25%), ενώ σε 2 projects το πλήθος αυτών που επιβίωσαν είναι ίσο με αυτών που διαγράφησαν.
- 2_MODERATE : Στα 4 από τα 5 (80%) είναι μεγαλύτερο και σε ένα ίσο.
- 3_FocusedShot_n_LOW : Στα 3 από τα 8 είναι μεγαλύτερο (37,5%) και σε 2 ίσο.
- 5_ACTIVE : Στα 12 από τα 15 (80%) είναι μεγαλύτερο.

Στα διαγράμματα του Σχήματος 4-5 φαίνεται το πλήθος των wide πινάκων που επιβίωσαν (SurvWide) και το πλήθος των wide πινάκων που διαγράφησαν (DeadWide), ξεχωριστά για τις 5 κατηγορίες.

Στις περισσότερες περιπτώσεις που οι dead – wide πίνακες είναι περισσότεροι από τους survived - wide πίνακες παρατηρείται μια μεγάλη διαφορά ανάμεσα στις αντίστοιχες στήλες του διαγράμματος. Μια ερμηνεία θα μπορούσε να είναι ότι σε αυτά τα projects έγινε μια μεγάλη αλλαγή στη βάση που ίσως να οφείλεται σε αλλαγή των αναγκών των χρηστών ή και σε λάθος αρχική σχεδίαση της βάσης.

Στα 1_ALMOST_FROZEN παρατηρείται, ότι δεν διαγράφηκε κανένας wide πίνακας. Πιθανόν γιατί η μικρή δραστηριότητα σε αυτά τα projects οδηγεί προς το να "παγώσουν" από το να γίνει προσπάθεια ώστε να διαγραφούν οι βασικοί πίνακες και να επανασχεδιαστούν.



Σχήμα 4-5 : Το πλήθος των wide πινάκων που επιβίωσαν σε σχέση με το πλήθος των wide πινάκων που διαγράφησαν για τις 5 κατηγορίες.

Συμπεραίνεται ότι υπάρχει ισχυρή πιθανότητα να επιβιώσει ένας wide πίνακας. Όσον αφορά την απάντηση στο αν ένας wide πίνακας έχει μεγαλύτερη πιθανότητα να επιβιώσει και όχι να διαγραφεί, μπορούμε να πούμε ότι αυτό ισχύει αφού η πιθανότητα επιβίωσης στο σύνολο των projects είναι 59%. (119 πίνακες επιβίωσαν και 83 διαγράφησαν). Αξιοσημείωτο είναι ότι οι πιθανότητες επιβίωσης των wide πινάκων εμφανίζουν μεγάλες αποκλίσεις στις τιμές τους ανά κατηγορία. Σε μια κατηγορία μάλιστα, την 1_FocusedShot_n_FROZEN, παρατηρείται ότι οι wide πίνακες που επιβίωσαν ήταν λιγότεροι σε σχέση με αυτούς που διαγράφησαν. Συνολικά όμως μπορούμε να πούμε ότι οι wide πίνακες έχουν μεγαλύτερη πιθανότητα να επιβιώσουν από το να διαγραφούν.

4.2 Πιθανότητα να επιβιώσουν οι wide σε σχέση με τους not wide;

Ένα άλλο ερώτημα που ερευνάται, *είναι αν υπάρχει μεγαλύτερη πιθανότητα να επιβιώσει ένας πίνακας αν είναι wide σε σχέση με έναν πίνακα ο οποίος είναι not wide*. Στο συγκεκριμένο ερώτημα μελετώνται 34 projects διότι υπάρχει ένα project χωρίς not wide πίνακες και συνεπώς δεν έχει νόημα η σύγκριση.

Όπως φαίνεται στο διάγραμμα του Σχήματος 4-6, στα ***14*** projects που ανήκουν στην κατηγορία 4_ACTIVE παρατηρούνται τα εξής :

- Στα 11 projects (79%) είναι μεγαλύτερη η πιθανότητα να επιβιώσουν οι wide πίνακες σε σχέση με τους not wide.
- Οι διαφοροποιήσεις μεταξύ των ποσοστών δεν υπερβαίνουν το 25%.
- Στα 8 από τα 14 projects (ποσοστό 54%), η διαφορά στα ποσοστά είναι μέχρι 8% που σημαίνει ότι σε αυτά τα projects η πιθανότητα να διαγραφεί ένας πίνακας δεν εξαρτιόταν και τόσο πολύ από το μέγεθος του.

Υπάρχει μια μικρή διαφορά στην πιθανότητα επιβίωσης των wide έναντι των not wide πινάκων αλλά αυτή η διαφορά φαίνεται να μην είναι σημαντική.

Στα ***3*** projects που ανήκουν στην κατηγορία 1_ALMOST_FROZEN παρατηρούνται τα εξής :

- Και στα 3 projects (100%) είναι μεγαλύτερη η πιθανότητα να επιβιώσουν οι wide πίνακες σε σχέση με τους not wide.
- Οι διαφοροποιήσεις μεταξύ των ποσοστών είναι σημαντικές και κυμαίνονται από 33% (σε 2 projects) μέχρι και 100% σε 1 project (δηλαδή επιβίωσαν όλοι οι wide και διαγράφησαν όλοι οι not wide).

Σε αυτή την κατηγορία παρατηρείται *αρκετά σημαντική διαφορά στην πιθανότητα επιβίωσης των wide έναντι των not wide πινάκων*.

Στα ***4*** projects που ανήκουν στην κατηγορία 1_FocusedShot_n_FROZEN παρατηρούνται τα εξής :

- Στα 3 projects (75%) είναι μεγαλύτερη η πιθανότητα να επιβιώσουν οι wide πίνακες σε σχέση με τους not wide.
- Οι διαφοροποιήσεις μεταξύ των ποσοστών κυμαίνονται από 3% μέχρι και 50%.

Σε αυτή την κατηγορία παρατηρείται *μια μικρή διαφορά στην πιθανότητα επιβίωσης των wide έναντι των not wide πινάκων χωρίς όμως να εξάγεται κάποιο σημαντικό συμπέρασμα.*

Στα ***5*** projects που ανήκουν στην κατηγορία 2_MODERATE παρατηρούνται τα εξής :

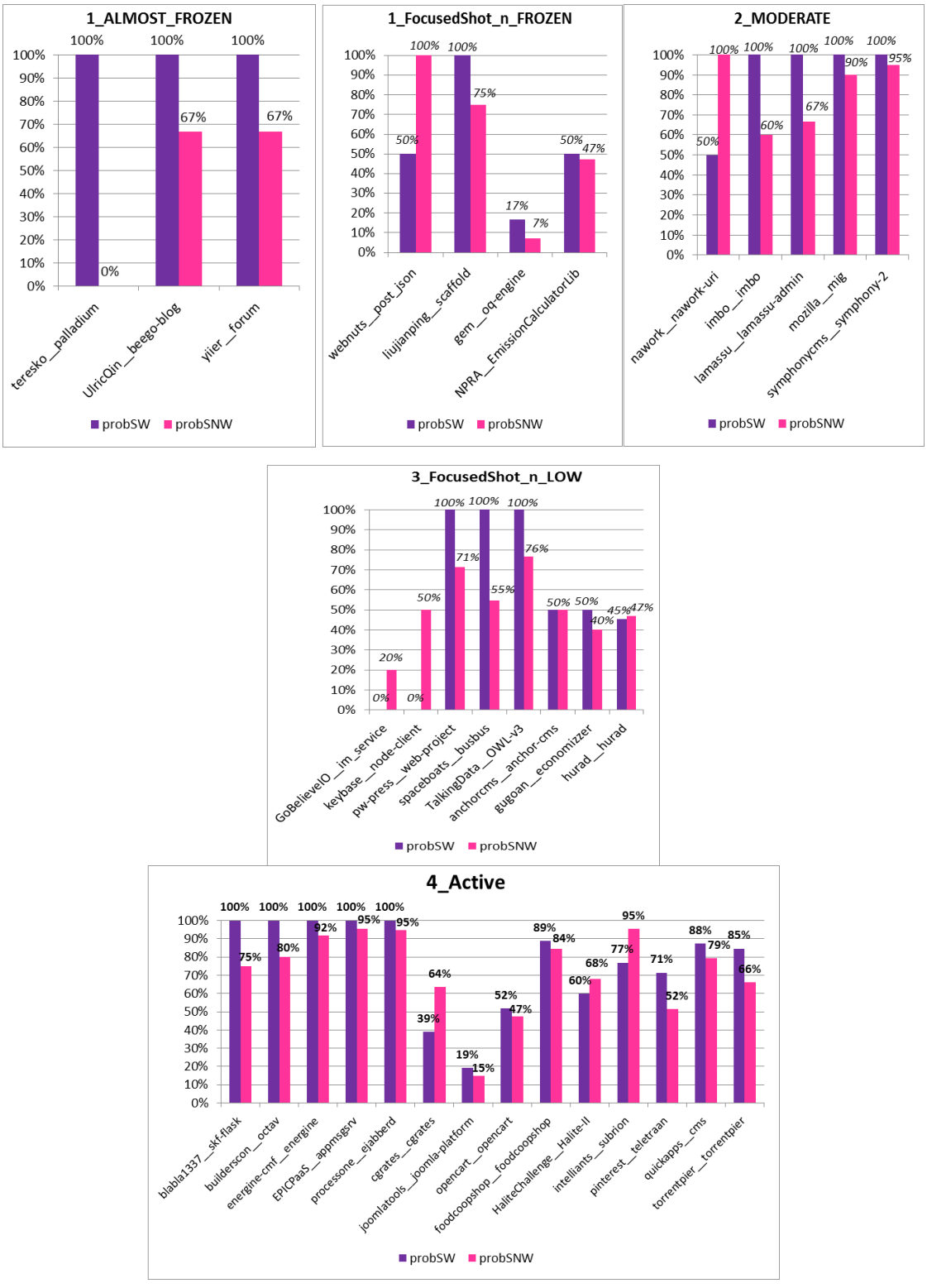
- Στα 4 από τα 5 projects (80%) είναι μεγαλύτερη η πιθανότητα να επιβιώσουν οι wide πίνακες σε σχέση με τους not wide.
- Οι διαφοροποιήσεις μεταξύ των ποσοστών κυμαίνονται από 5% μέχρι και 50%.

Σε αυτή την κατηγορία παρατηρείται μια *αρκετά σημαντική διαφορά στην πιθανότητα επιβίωσης των wide έναντι των not wide πινάκων*.

Στα ***8*** projects που ανήκουν στην κατηγορία 3_FocusedShot_n_LOW παρατηρούνται τα εξής :

- Στα 4 από τα 8 projects (50%) είναι μεγαλύτερη η πιθανότητα να επιβιώσουν οι wide πίνακες σε σχέση με τους not wide και σε ένα project η πιθανότητα είναι ίση.
- Οι διαφοροποιήσεις μεταξύ των ποσοστών είναι σημαντικές και κυμαίνονται από 0% (δηλαδή έχουν ίδιο ποσοστό επιβίωσης και οι wide και οι not wide) μέχρι και 50%

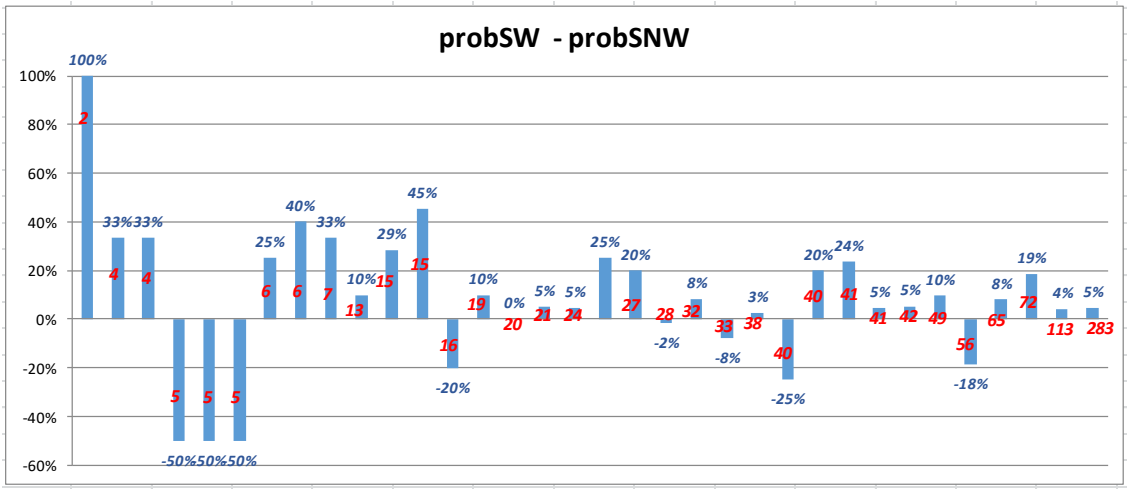
Σε αυτή την κατηγορία θα λέγαμε ότι *δεν υπερισχύει σημαντικά καμία πιθανότητα ως προς την επιβίωση των wide και των not wide πινάκων*. Αξιοσημείωτο είναι ότι υπάρχουν και δύο projects που επιβίωσαν μόνο οι not wide.



Σχήμα 4-6 Πιθανότητα να επιβιώσουν οι wide πινάκες (probSW) σε σχέση με την πιθανότητα να επιβιώσουν οι not wide πίνακες (probSNW) ανά κατηγορία

Σε όλες τις κατηγορίες υπερισχύει η πιθανότητα να επιβιώσουν οι wide έναντι των not wide χωρίς να παρατηρείται όμως κάποια σημαντική υπεροχή. Στο

Σχήμα 14-7 φαίνονται οι διαφορές στα ποσοστά των πιθανοτήτων επιβίωσης για κάθε data set. Οι γραμμές με αρνητικές τιμές αντιστοιχούν σε μεγαλύτερες πιθανότητες να επιβιώσουν οι not wide. Τα σύνολα δεδομένων είναι ταξινομημένα με βάση το πλήθος των πινάκων. Το πλήθος των πινάκων φαίνεται με κόκκινα γράμματα.



Σχήμα 4-7 Ιστόγραμμα με τις διαφορές που προκύπτουν αν από το ποσοστό της πιθανότητας επιβίωσης των wide πινάκων αφαιρέσουμε το αντίστοιχο των Not wide. (με κόκκινα γράμματα φαίνεται το πλήθος των πινάκων.

Οι μεγάλες διαφοροποιήσεις των ποσοστών εμφανίζονται σε projects με λίγους πίνακες όπως φαίνεται στις γραμμοσκιασμένες περιοχές του Πίνακα 4-3. Αν δεν λάβουμε υπ' όψιν αυτά τα projects με τους λίγους πίνακες τότε η διαφορά στα ποσοστά είναι μικρή όσον αφορά την πιθανότητα επιβίωσης των wide πινάκων σε σχέση με τους not wide.

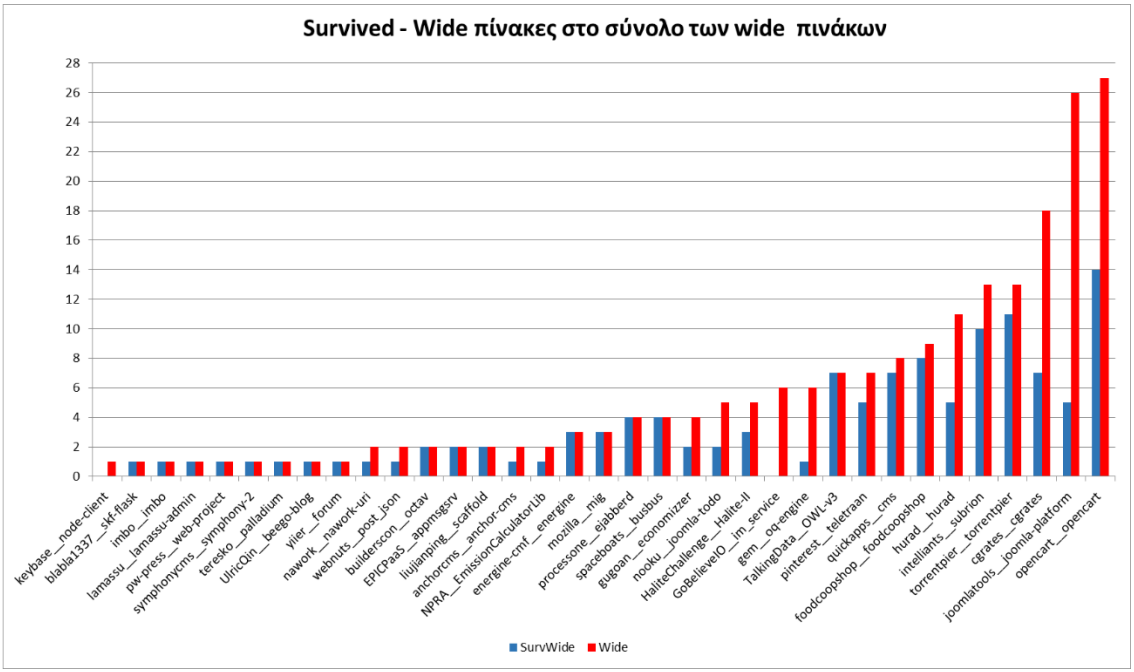
Με βάση την παραπάνω ανάλυση θα λέγαμε ότι δεν υπάρχει ισχυρή ένδειξη στα υπό μελέτη projects ότι οι wide πίνακες επιβιώνουν σε πολύ μεγαλύτερο ποσοστό από τους not wide πίνακες. Οι αλλαγές που συμβαίνουν στα projects οδηγούν σε διαγραφή πινάκων που έχουν μικρή συσχέτιση με το μέγεθος τους.

Πίνακας 4-3 Διαφορές στα ποσοστών επιβίωσης των wide και not wide πινάκων.

Project	TotalWide	TotalNotWide	probSW	probSNW	Διαφορά	OverallClass
teresko__palladium	1	1	100%	0%	**100%**	1_ALMOST_FROZEN
UlricQin__beego-blog	1	3	100%	67%	**33%**	
yiier__forum	1	3	100%	67%	**33%**	
webnuts__post_json	2	3	50%	100%	**-50%**	1_FocusedShot_n_FROZEN
liujianping__scaffold	2	4	100%	75%	**25%**	
gem__oq-engine	6	43	17%	7%	**10%**	
NPRA__EmissionCalculatorLib	2	36	50%	47%	**3%**	
nawork__nawork-uri	2	3	50%	100%	**-50%**	2_MODERATE
imbo__imbo	1	5	100%	60%	**40%**	
lamassu__lamassu-admin	1	6	100%	67%	**33%**	
mozilla__mig	3	10	100%	90%	**10%**	
symphonycms__symphony-2	1	20	100%	95%	**5%**	
GoBelieveIO__im_service	6	10	0%	20%	**-20%**	3_FocusedShot_n_LOW
keybase__node-client	1	4	0%	50%	**-50%**	
pw-press__web-project	1	14	100%	71%	**29%**	
spaceboats__busbus	4	11	100%	55%	**45%**	
TalkingData__OWL-v3	7	34	100%	76%	**24%**	
anchorcms__anchor-cms	2	18	50%	50%	**0%**	
gugoan__economizzer	4	15	50%	40%	**10%**	
hurad__hurad	11	17	45%	47%	**-2%**	
blabla1337__skf-flask	1	24	100%	75%	**25%**	4_ACTIVE
builderscon__octav	2	25	100%	80%	**20%**	
energine-cmf__energine	3	62	100%	92%	**8%**	
EPICPaaS__appmsgsrv	2	22	100%	95%	**5%**	
processone__ejabberd	4	38	100%	95%	**5%**	
cgrates__cgrates	18	22	39%	64%	**-25%**	
joomlatools__joomla-platform	26	87	19%	15%	**4%**	
opencart__opencart	27	256	52%	47%	**5%**	
foodcoopshop__foodcoopshop	9	32	89%	84%	**5%**	
HaliteChallenge__Halite-II	5	28	60%	68%	**-8%**	
intelliants__subrion	13	43	77%	95%	**-18%**	
pinterest__teletraan	7	33	71%	52%	**20%**	
quickapps__cms	8	24	88%	79%	**8%**	
torrentpier__torrentpier	13	59	85%	66%	**19%**	

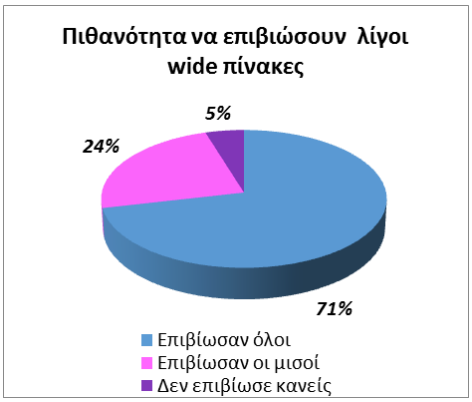
4.3 Επιβιώνουν οι wide πίνακες με μεγαλύτερη πιθανότητα όταν είναι λίγοι;

Ένα άλλο ερώτημα που προκύπτει είναι *αν οι wide πίνακες διαγράφονται πιο δύσκολα όταν είναι λίγοι*. Στο Σχήμα 4-8 απεικονίζεται η σχέση ***(α)*** των wide πινάκων που επιβίωσαν (SurvWide) με ***(β)*** το σύνολο των wide πινάκων (Wide) σε κάθε project. Τα projects ταξινομήθηκαν κατά αύξουσα σειρά του συνόλου των wide πινάκων και γίνεται εμφανές ότι τα δύο αυτά πλήθη συμβαδίζουν σε μεγάλο βαθμό στα projects που το σύνολο των wide πινάκων είναι μέχρι 4 πίνακες. Όσο το σύνολο των wide πινάκων απομακρύνεται από τον αριθμό 4 παρατηρείται μια αξιοσημείωτη διαφορά μεταξύ των δύο αυτών μεγεθών.



Σχήμα 4-8 Σχέση των Survived - Wide πινάκων με το σύνολο των Wide πινάκων κατά αύξουσα σειρά του πλήθους των wide πινάκων ανά project.

Στο πίνακα 4-4 εμφανίζονται τα projects που έχουν σύνολο wide πινάκων μικρότερο ή ίσο του 4. Στα 15 από τα 21 projects επιβίωσαν όλοι οι wide πίνακες, στα 5 από αυτά επιβίωσαν οι μισοί wide πίνακες και μόνο σε ένα διαγράφησαν όλοι οι wide πίνακες. Οι παραπάνω τιμές εμφανίζονται ποσοστιαία στο Σχήμα 4-9.



Σχήμα 4-9 Ποσοστά επιβίωσης των wide πινάκων όταν είναι <= 4

Πίνακας 4-4 Projects με συνολικό αριθμό wide πινάκων <= 4

Project	OverallClass	sumWide	SurvWide	probSW
teresko__palladium	1_ALMOST_FROZEN	1	1	100%
UlricQin__beego-blog		1	1	100%
yiier__forum		1	1	100%
webnuts__post_json	1_FocusedShot_n_FROZEN	2	1	50%
liujianping__scaffold		2	2	100%
NPRA__EmissionCalculatorLib		2	1	50%
nawork__nawork-uri	2_MODERATE	2	1	50%
imbo__imbo		1	1	100%
lamassu__lamassu-admin		1	1	100%
mozilla__mig		3	3	100%
symphonycms__symphony-2		1	1	100%
keybase__node-client	3_FocusedShot_n_LOW	1	0	0%
pw-press__web-project		1	1	100%
spaceboats__busbus		4	4	100%
anchorcms__anchor-cms		2	1	50%
gugoan__economizzer		4	2	50%
blabla1337__skf-flask	4_ACTIVE	1	1	100%
builderscon__octav		2	2	100%
energine-cmf__energine		3	3	100%
EPICPaaS__appmsgsrv		2	2	100%
processone__ejabberd		4	4	100%

Άρα όταν οι wide πίνακες είναι λίγοι *υπάρχει πολύ μεγάλη πιθανότητα να επιβιώσουν όλοι και αν όχι όλοι, τουλάχιστον οι μισοί*. Στο Πίνακα 4-5 απεικονίζονται τα ποσοστά επιβίωσης των wide πινάκων σε σχέση με το πλήθος των wide πινάκων στα υπό μελέτη projects.

Αξιοσημείωτο είναι ότι το μεγαλύτερο ποσοστό επιβίωσης (88,8%) εμφανίζεται στα projects με 1 wide πίνακα. Μόνο σε ένα project από τα 9 που είχαν ένα wide πίνακα διαγράφηκε αυτό ο ένας wide πίνακας.

Στη στήλη *Ποσοστό επιβίωσης αθροιστικά* παρατηρείται ότι το ποσοστό επιβίωσης των wide πινάκων αυξάνει μέχρι τα projects με συνολικό αριθμό wide πινάκων μέχρι 4 πίνακες. Σε αυτή τη γραμμή το ποσοστό επιβίωσης φτάνει στο 82,9%, το μεγαλύτερο αν εξαιρέσουμε το ποσοστό που αντιστοιχεί στα project με 1 wide πίνακα. Μετά από τα projects με 4 wide πίνακες το αθροιστικό ποσοστό εμφανίζει μικρότερες τιμές. Την μικρότερη τιμή την εμφανίζει στην τελευταία

γραμμή που αντιστοιχεί στα projects με πολλούς wide πίνακες. Όσο το πλήθος των wide πινάκων μεγαλώνει το ποσοστό επιβίωσης τους μειώνεται.

Συνεπώς γίνεται φανερό ότι στα υπό μελέτη projects η πιθανότητα να επιβιώσουν οι wide πίνακες όταν είναι λίγοι είναι αρκετά μεγάλη.

Πίνακας 4-5 Ποσοστά επιβίωσης των wide πινάκων σε σχέση με το πλήθος των wide πινάκων στα υπό μελέτη projects

	πλήθος projects	***ποσοστό επιβίωσης***	***πλήθος projects αθροιστικά***	***Ποσοστό επιβίωσης αθροιστικά***
Projects με 1 wide πίνακες	9	88,8%	9	88,8%
Projects με 2 wide πίνακες	7	71,4%	16	78,3%
Projects με 3 wide πίνακες	2	100%	18	82,7%
Projects με 4 wide πίνακες	***3***	***83,3%***	***21***	***82,9%***
Projects με 5 wide πίνακες	2	50%	23	76,5%
Projects με 6 wide πίνακες	2	8,3%	25	63,5%
Projects με 7 wide πίνακες	2	85,7%	27	67,5%
Projects με 8-11 wide πίνακες	3	71.4%	30	68,6%
Projects με 13 wide πίνακες	2	80.8%	32	71,0%
Projects με 18 wide πίνακες	1	38.9%	33	67,1%
Projects με 26-27 wide πίνακες	2	35.8%	35	58,9%

4.4 Μοτίβο Γάμμα

Όπως αναφέρθηκε παραπάνω, στην περίπτωση του μοτίβου Γ, εξετάζεται η σχέση μεταξύ (α) του μεγέθους του σχήματος κατά τη γέννηση και (β) της διάρκειας ζωής ενός πίνακα. Έγιναν δύο δοκιμές για να προσδιοριστεί εάν ένα σύνολο δεδομένων ακολουθεί το μοτίβο, μία γεωμετρική (Geometry ?) και μία στατιστική (FisherExecuted).

Για όλα τα projects, έχουμε ***(α)*** δοκιμή του γεωμετρικού μοτίβου που βασίστηκε στα γεωμετρικά χαρακτηριστικά των scatterplots που παράγονται από τους συνδυασμούς των δύο χαρακτηριστικών (μέγεθος – επιβίωση), και ***(β)*** στατιστική δοκιμή, για την ανεξαρτησία των δύο χαρακτηριστικών. Η δοκιμή αυτή θεωρείται επιτυχημένη όταν το fisher test εκτελείται και η τιμή της p_value είναι μικρότερη του 5%, (τιμή μικρότερη από 5% στην p_value δείχνει λιγότερο από 5% πιθανότητα τα συμπεράσματα να οφείλονται σε τυχαιότητα).

Στον πίνακα 4-6 φαίνεται το πλήθος των projects που ικανοποιούν τις δοκιμές στο συνολικό αριθμό των projects, καθώς και ανά κατηγορία.

Πίνακας 4-6 Πλήθη και αντίστοιχα ποσοστά που μετρήθηκαν : α) πιθανότητα να επιβιώσουν οι wide πίνακες > από τους not wide β) πέτυχε η δοκιμή του γεωμετρικού μοτίβου γ) εκτελέστηκε το fisher test

	WMLTS?= *TRUE*	***Πέτυχε η δοκιμή του γεωμετρικού μοτίβου***	***FISHER TEST = TRUE***	***FISHER TEST = TRUE & p_value = TRUE***
Σύνολο Projects (35)	***25 71%***	***22 63%***	***16 46%***	***Δεν ικανοποιείται σε κανένα project η στατιστική δοκιμή***
1_ALMOST_FROZEN (3)	***3 100%***	***3 100%***	***0 0%***	
1_FocusedShot_n_FROZEN (4)	***3 75%***	***1 25%***	***2 50%***	
2_MODERATE (5)	***4 80%***	***4 80%***	***0 0%***	
3_FocusedShot_n_LOW (8)	***4 50%***	***3 37,5%***	***3 37,5%***	
5_ACTIVE (15)	***11 79%*** *(υπολογίστηκε στα 14 projects)*	***11 73,3%***	***9 33,3%***	

Wide More Likely To Survive = TRUE

Σε 25 από τα 35 συνολικά projects (δηλαδή σε ποσοστό 71%) η πιθανότητα επιβίωσης των wide πινάκων είναι μεγαλύτερη από την πιθανότητα επιβίωσης των not wide πινάκων.

Στην κατηγορία 1_ALMOST_FROZEN και στα 3 project η πιθανότητα επιβίωσης των wide πινάκων είναι μεγαλύτερη από την πιθανότητα επιβίωσης not wide πινάκων, ποσοστό 100%.

Ακολουθεί η κατηγορία 2_MODERATE όπου είναι μεγαλύτερη η πιθανότητα επιβίωσης των wide πινάκων σε 4 από τα 5 projects, ποσοστό 80%.

Πολύ κοντά με ποσοστό σχεδόν 79% είναι η κατηγορία 5_ACTIVE. Το ποσοστό υπολογίστηκε στα 14 Projects της κατηγορίας αυτής, καθώς υπάρχει ένα project στο οποίο δεν μπορεί να υπολογιστεί η πιθανότητα επιβίωσης των not wide πινάκων διότι δεν επιβίωσε κανένας not wide.

Ακολουθεί η κατηγορία 1_FocusedShot_n_FROZEN με ποσοστό 75%.

Το ποσοστό πέφτει στο 50% για την κατηγορία 3_FocusedShot_n_LOW, όπου στα projects αυτής της κατηγορίας η πιθανότητα επιβίωσης των wide πινάκων είναι μεγαλύτερη μόνο στα μισά από αυτά.

Πέτυχε η δοκιμή του γεωμετρικού μοτίβου?

Σε 22 από τα 35 συνολικά projects (δηλαδή σε ποσοστό 63%) πέτυχε η δοκιμή του γεωμετρικού μοτίβου. Από τα 22 projects που πέτυχε το γεωμετρικό μοτίβο στα 20 (ποσοστό 91%) το WMLTS? είναι TRUE (η πιθανότητα επιβίωσης των wide πινάκων να είναι μεγαλύτερη από την πιθανότητα επιβίωσης των not wide πινάκων) Μόνο σε 2 projects (HaliteChallenge__Halite-II και intelliants_subrion) της κατηγορίας 5_ACTIVE έχουν το WMLTS? = FALSE.

Στην κατηγορία 1_ALMOST_FROZEN και στα 3 project η πέτυχε η δοκιμή του γεωμετρικού μοτίβου, ποσοστό 100%.

Ακολουθεί η κατηγορία 2_MODERATE όπου πέτυχε η δοκιμή του γεωμετρικού μοτίβου σε 4 από τα 5 projects, ποσοστό 80%. Στο project (nawork_nawork-uri) που δεν πέτυχε η δοκιμή είναι αυτό που η πιθανότητα επιβίωσης των wide πινάκων είναι μικρότερη από την πιθανότητα επιβίωσης not wide πινάκων έχει δηλαδή το WMLTS? = FALSE.

Πολύ κοντά με ποσοστό 73,3% είναι η κατηγορία 5_ACTIVE. Στα 4 projects που απέτυχε η δοκιμή του γεωμετρικού μοτίβου δεν υπάρχει ταύτιση με την τιμή του WMLTS? (στα 2 το WMLTS? Είναι TRUE, στο ένα είναι FALSE και στο άλλο δεν μπορεί να υπολογιστεί η τιμή του WMLTS? καθώς σε αυτό το project, το nooku_joomla-todo, δεν υπήρχαν not wide πίνακες)

Το ποσοστό πέφτει στο 37.5% για την κατηγορία 3_FocusedShot_n_LOW, όπου και στα τρία projects που πετυχαίνει η δοκιμή του γεωμετρικού μοτίβου έχουν WMLTS? = TRUE.

Το μικρότερο ποσοστό επιτυχίας του γεωμετρικού μοτίβου το έχει η κατηγορία 1_FocusedShot_n_FROZEN με ποσοστό 25% καθώς μόνο ένα project αυτής της κατηγορίας πετυχαίνει στην δοκιμή του γεωμετρικού μοτίβου. Και αυτό το project έχει WMLTS? = TRUE.

Πέτυχε η εκτέλεση του fisher test?

Σε 16 από τα 35 συνολικά projects (δηλαδή σε ποσοστό 46%) πέτυχε η εκτέλεση του fisher test. Και στα 16 projects που πέτυχε το WMLTS? είναι TRUE (η πιθα-

νότητα επιβίωσης των wide πινάκων να είναι μεγαλύτερη από την πιθανότητα επιβίωσης των not wide πινάκων).

Στην κατηγορία 1_ALMOST_FROZEN και στα 3 project πέτυχε η εκτέλεση του fisher test. Σε όλα τα projects αυτής της κατηγορίας πέτυχε η δοκιμή του γεωμετρικού μοτίβου και το WMLTS? είναι TRUE.

Ακολουθεί η κατηγορία 2_MODERATE όπου πέτυχε η η εκτέλεση του fisher test σε 4 από τα 5 projects, ποσοστό 80%. Σε αυτά τα 4 πέτυχε και η δοκιμή του γεωμετρικού μοτίβου (ποσοστό 80%) και το WMLTS? είναι TRUE (ποσοστό επίσης 80%).

Το ποσοστό πέφτει στο 37.5% για την κατηγορία 3_FocusedShot_n_LOW, όπου η η εκτέλεση του fisher test πετυχαίνει σε 3 projects. Μόνο σε αυτά τα 3 projects πετυχαίνει και η δοκιμή του γεωμετρικού μοτίβου και έχουν επίσης WMLTS? = TRUE. Σε αυτή την κατηγορία υπάρχει ένα project, το gugoan__economizzer, που έχει WMLTS? = TRUE αλλά αποτυχαίνει και στην δοκιμή του γεωμετρικού μοτίβου και στην εκτέλεση του fisher test.

Στη εκτέλεση του fisher test παρατηρείται ένα μικρό ποσοστό στην κατηγορία 5_ACTIVE που φτάνει μόλις το 33,3%. Στα 5 projects αυτής της κατηγορίας που πέτυχε η εκτέλεση του fisher test πετυχαίνει και η δοκιμή του γεωμετρικού μοτίβου και έχουν επίσης WMLTS? = TRUE.

Το μικρότερο ποσοστό επιτυχίας ζτην εκτέλεση του fisher test το έχει η κατηγορία 1_FocusedShot_n_FROZEN με ποσοστό 25% καθώς μόνο ένα project αυτής της κατηγορίας πετυχαίνει στην στατιστική δοκιμή.

Δοκιμή Γεωμετρίας - Εκτέλεση του fisher test

Στους πίνακες 4-7 και 4-8 φαίνεται το πλήθος των projects στην συσχέτιση της εκτέλεσης του fisher test και της γεωμετρίας στο σύνολο των projects και ανά κατηγορία αντίστοιχα.

Πίνακας 4-7 Συσχέτιση της εκτέλεσης του fisher test με την δοκιμή γεωμετρίας στο σύνολο των projects

		fisher test ?		
		True	***False***	
Geometry	***True***	**16**	**6**	**22**
	False	**0**	**13**	**13**
		16	**19**	**35**

Πίνακας 4-8 : Συσχέτιση της εκτέλεσης του fisher test με την δοκιμή γεωμετρίας ανά κατηγορία

1_ALMOST_FROZEN

		fisher test ?		
		True	False	
Geometry	True	3	0	3
	False	0	0	0
		3	0	3

1_FocusedShot_n_FROZEN

		fisher test ?		
		True	False	
Geometry	True	1	0	1
	False	0	3	3
		1	3	4

2_MODERATE

		fisher test ?		
		True	False	
Geometry	True	4	0	4
	False	0	1	1
		4	1	5

3_FocusedShot_n_LOW

		fisher test ?		
		True	False	
Geometry	True	3	0	3
	False	0	5	5
		3	5	8

5_ACTIVE

		fisher test ?		
		True	False	
Geometry	True		6	11
	False	0	4	4
		5	10	15

ΚΕΦΑΛΑΙΟ 5

ΜΟΤΙΒΟ ΑΝΤΙΣΤΡΟΦΟΥ ΓΑΜΜΑ

Στην περίπτωση του μοτίβου του αντίστροφου Γ, εξετάζεται η σχέση μεταξύ (α) της διάρκειας ζωής ενός πίνακα και (β) του πλήθους των ενημερώσεων, ενός πίνακα.

Η ουσία του μοτίβου είναι ότι στους πίνακες με μικρή διάρκεια γίνονται λίγες αλλαγές, στους πίνακες με μεσαία διάρκεια γίνονται λίγες ή μεσαίου αριθμού αλλαγές και οι πίνακες με μεγάλη διάρκεια επιδεικνύουν όλα τα είδη της συμπεριφοράς σε σχέση με τα updates.

Ως ***long-lived*** *πίνακες ή* ***HighDuration*** *πίνακες χαρακτηρίζονται οι πίνακες με διάρκεια ζωής ίση ή μεγαλύτερη* από το 90% της μέγιστης διάρκειας στο σύνολο δεδομένων.

Ένας πίνακας χαρακτηρίζεται ως πίνακας σημαντικής δραστηριότητας (***highUpd ή high changer***) όταν το άθροισμα των ενημερώσεων ξεπερνά ένα όριο. Αυτό το όριο υπολογίζεται με την ικανοποίηση 2 συνθηκών :

- Τουλάχιστον 5 ενημερώσεις για έναν πίνακα. Το όριο αυτό είναι το κατώφλι που χρησιμοποιείται για τον καθορισμό ενός πίνακα ως ενεργό και είναι ανεξάρτητο από τις ενημερώσεις που γίνονται στο σύνολο των δεδομένων.

- Το άθροισμα των ενημερώσεων που συνέβησαν σε ένα πίνακα είναι μεγαλύτερο του 15% του μέγιστου πλήθος των ενημερώσεων που συνέβησαν σε έναν πίνακα σε όλο το σύνολο δεδομένων. Με αυτό τον υπολογισμό υπάρχει μια συσχέτιση των αλλαγών σε ένα πίνακα με τις αλλαγές που γίνονται στο σύνολο δεδομένων. Αυτό είναι ένα όριο που μπαίνει επίτηδες πολύ χαμηλά και παρ' όλα αυτά λίγοι είναι οι πίνακες που χαρακτηρίζονται ως high changers. Ειδικά οι πίνακες με λίγα attributes είναι πολύ δύσκολο να χαρακτηριστούν ως high changers καθώς και το 15% μπορεί να είναι μεγάλος αριθμός αλλαγών γι' αυτούς.

Στη συγκεκριμένη εργασία εξετάσθηκε η συμπεριφορά 195 συνόλων δεδομένων. Τα 195 σύνολα δεδομένων ομαδοποιήθηκαν σε κατηγορίες ανάλογα με την δραστηριότητα που παρουσιάζουν στις αλλαγές (activity).

Οι 6 κατηγορίες στις οποίες ταξινομούνται τα data sets είναι :

- 0_FROZEN : totalActivity = 0
- 1_ALMOST_FROZEN : At most 3 active commits, totalActivity <= 10 updated attributes
- 1_FocusedShot_n_FROZEN : At most 3 active commits, totalActivity > 10 updated attributes
- 2_MODERATE : None of the rest, totalActivity <= 90 updated attributes
- 3_FocusedShot_n_LOW : Between 4 and 10 active commits, 1 or 2 reeds
- 5_ACTIVE : None of the rest, totalActivity > 90 updated attributes

Στον πίνακα 5.1 φαίνεται το πλήθος των data sets ανά κατηγορία.

Πίνακας 5-1 Το πλήθος των data sets ανά κατηγορία

	Πλήθος Data sets
0_FROZEN	**34**
1_ALMOST_FROZEN	**65**
1_FocusedShot_n_FROZEN	**25**
2_MODERATE	**29**
3_FocusedShot_n_LOW	**20**
5_ACTIVE	**22**
Σύνολο	**195**

Για κάθε data set έγινε η δοκιμή γεωμετρίας προκειμένου να εντοπιστεί αν πετυχαίνει το γεωμετρικό μοτίβο. Για να εντοπιστεί αν το γεωμετρικό μοτίβο πετυχαίνει, ορίζεται στο γράφημα μια *περιοχή* που αντιστοιχεί (α) στο 90% της μέγιστης διάρκειας (notHighDuration) και (β) στο 15% του μέγιστου αθροίσματος ενημερώσεων ενός πίνακα σε ολόκληρο το σύνολο δεδομένων (hghlUpdates) και αυτή η περιοχή καλύπτει το 0.9x0.85 = 76,5% του συνόλου της περιοχής του 2-διάστατου χώρου. Αν η περιοχή αυτή, έχει λιγότερο από το ***15%*** των πινάκων τότε η γεωμετρική δοκιμή πετυχαίνει. Οι πίνακες που ανήκουν στην περιοχή αυτή έχουν μικρή διάρκεια ζωής (notHighDuration) αλλά πολλές ενημερώσεις (highUpd). Συνεπώς η ύπαρξη πολλών πινάκων σε αυτή την περιοχή του σχήματος αντιτίθεται στην ουσία του μοτίβου.

5.1 Στατιστική δοκιμή

Το Fisher τεστ είναι ένα στατιστικό τεστ που εφαρμόστηκε στο κάθε σύνολο δεδομένων και δημιουργεί τιμές που αντιστοιχούν στον διαχωρισμό των πινάκων με βάση τα δύο χαρακτηριστικά (διάρκεια - αλλαγές). Ο διαχωρισμός αυτός γίνεται σε highDur και notHighDur και σε highUpd και notHighUpd για κάθε σύνολο δεδομένων. Οι τιμές αυτές τοποθετούνται σε έναν πίνακα όπου το ένα χαρακτηριστικό (διάρκεια) αποτελεί τις γραμμές του πίνακα και το άλλο χαρακτηριστικό (αλλαγές) τις στήλες του πίνακα. Ο πίνακας ***συσχέτισης*** (*contingency table*) που δημιουργείται για κάθε date set είναι της παρακάτω μορφής :

	highUpd	nothighUpd
highDur		
notHighDur		

Στο Fisher test ελέγχεται η ανεξαρτησία των δύο χαρακτηριστικών. Η ερευνητική υπόθεση, σε όλες τις περιπτώσεις, διατυπώνεται ως δίδυμο:

H_0: η μηδενική υπόθεση δηλώνει ότι η διάρκεια και οι αλλαγές είναι ανεξάρτητες μεταξύ τους, δηλαδή, δεν υπάρχει συσχέτιση των αλλαγών με την μικρή ή μεγάλη διάρκεια. Εκφράζεται από τη σχέση :

H_0: prob(highUpd | highDur) <= prob(highUpd | notHighDur)

H_A: υπάρχει διαφορά στις αλλαγές που γίνονται σε έναν πίνακα ανάλογα με την διάρκεια ζωής του. Εκφράζεται από τη σχέση :

$$H_A: prob(highUpd \mid highDur) > prob(highUpd \mid notHighDur)$$

Η ερευνητική υπόθεση H_0 εκφράζει την πιθανότητα τα παραγόμενα αποτελέσματα να οφείλονται στην τύχη.

Στον πίνακα 5-2 φαίνεται σε πόσα σύνολα δεδομένων της κάθε κατηγορίας εκτελέστηκε το fisher test, Επίσης φαίνεται σε πόσα από αυτά που εκτελέστηκε το fisher test πέτυχε η δοκιμή γεωμετρίας και επίσης σε πόσα από αυτά που δεν εκτελέστηκε το fisher test πέτυχε η δοκιμή γεωμετρίας.

Πίνακας 5-2 Πλήθος των data sets ανά κατηγορία που εκτελέστηκε το fisher test, και συσχέτιση με δοκιμή γεωμετρίας.

Fisher Test	**Πλήθος data set**	**Εκτελέστηκε το Fisher Test ?**	**Geometry = TRUE**	
			Σε αυτά που εκτελέστηκε το Fisher Test	**Σε αυτά που ΔΕΝ εκτελέστηκε το Fisher Test**
0_FROZEN	34	0 (*0%*)	0	34
1_ALMOST_FROZEN	65	0 (*0%*)	0	65
1_FocusedShot_n_FROZEN	25	2 (*8%*)	1	19
2_MODERATE	29	6 (*21%*)	4	19
3_FocusedShot_n_LOW	20	7 (*29%*)	5	12
5_ACTIVE	22	14 (*64%*)	11	6
ΣΥΝΟΛΟ των data sets	**195**	**29 (15*%*)**	**21**	**155**

Παρατηρείται ότι το fisher test εκτελέστηκε σε πολύ μικρό ποσοστό στο σύνολο των data sets (μόλις 15%). Στα σύνολα δεδομένων που ανήκουν στη κατηγορία 0_FROZEN και στη κατηγορία 1_ALMOST_FROZEN, το Fisher test δεν εκτελέστηκε ούτε σε ένα σύνολο δεδομένων. (αυτό εξηγείτε αφού οι συγκεκριμένες κατηγορίες έχουν ελάχιστες αλλαγές). Παρατηρείται ότι όσο μεγαλώνει το activity τόσο αυξάνει και το ποσοστό εκτέλεσης του Fisher test. Στα 5_ACTIVE data sets το ποσοστό αυτό φτάνει στο 64% που είναι και η μεγαλύτερη τιμή.

Προκύπτει λοιπόν το ερώτημα *αν θα πρέπει να ξεχωρίσουμε σε ποια από τα σύνολα δεδομένων στα οποία το Fisher test ήταν false δεν έπρεπε να εκτελεστεί καθόλου το τεστ σε αυτά, διότι δεν έχει νόημα η εκτέλεσή του; Δηλαδή αν πρέπει κάποια data sets να αφαιρεθούν από την μελέτη του μοτίβου;*

Στις κατηγορίες 0_FROZEN και 1_ALMOST_FROZEN δεν έχει νόημα να εκτελεστεί το fisher test διότι τα updates είναι τόσο λίγα που όλοι οι πίνακες είναι high changers με 0 αλλαγές, οπότε τα 2 από τα 4 κελιά του πίνακα είναι σχεδόν μηδέν.

Για να απαντηθεί το παραπάνω ερώτημα για τις υπόλοιπες κατηγορίες, έτρεξε ένα πρόγραμμα για κάθε data set αυτών των κατηγοριών (το οποίο ονομάζεται *Metrisis*) που υπολόγισε για κάθε data set τα παρακάτω :

 # tables : το πλήθος των πινάκων

maxDur : η μέγιστη διάρκεια στο σύνολο του data set.

MaxUpd : ο μέγιστος αριθμός των αλλαγών που εμφανίζεται σε κάποιον πίνακα στο σύνολο του data set.

High Changers : Το πλήθος των πινάκων που χαρακτηρίζονται ως πίνακες με τις μέγιστες αλλαγές (δηλαδή άνω του 15% των maxUpdates)

Πίνακας Συσχέτισης : το πλήθος των πινάκων για κάθε κελί του πίνακα συσχέτισης του fisher test.

Στους πίνακες 5-3, 5-4, 5-5, 5-6, παρουσιάζονται οι πληροφορίες που παρήχθησαν από το πρόγραμμα για κάθε κατηγορία. Όταν κάποια τιμή του πίνακα συσχέτισης είναι μηδέν φαίνεται με κόκκινο χρώμα. Στην τρίτη στήλη του πίνακα συσχέτισης, που αντιστοιχεί στους πίνακες με notHighDuration και hihgUpdates υπολογίστηκε και το ποσοστό αυτών των πινάκων επί του συνόλου των πινάκων. Στις περιπτώσεις που αυτό το ποσοστό είναι πάνω από 15% (τότε φαίνεται με μπλε χρώμα) αποτυγχάνει η δοκιμή γεωμετρίας. Σε αυτούς τους πίνακες φαίνεται ξεκάθαρα ότι η εκτέλεση του *fisher test αποτυγχάνει πάντα στις περιπτώσεις που στον πίνακα συσχέτισης υπάρχει μία ή περισσότερες μηδενικές τιμές*.

Ένα data set δεν έχει νόημα να μελετηθεί ως προς το μοτίβο inverse Gamma αν:

- αν δεν υπάρχει καθόλου activity (αυτά τα data sets φαίνονται στους πίνακες με κόκκινα γράμματα)
- αν το **MaxUpd** είναι τόσο μικρό που δεν μπορεί να γίνει διάκριση σε υψηλής – μεσαίας ή χαμηλής δραστηριότητας πίνακες. Επειδή ο χαρακτηρισμός ενός πίνακα σε high changer ή όχι γίνεται με βάση τις μέγιστες αλλαγές που εμφανίστηκαν σε κάποιον πίνακα, έχει νόημα αυτές οι αλλαγές να ξεπερνούν κάποιο όριο. Το όριο του MaxUpd για να έχει νόημα η μελέτη ενός συνόλου δεδομένων ορίστηκε στις 5 αλλαγές. Δη-

λαδή να υπάρχει στο data set τουλάχιστον ένας πίνακας με πάνω από 5 αλλαγές. (αυτά τα data sets φαίνονται στους πίνακες με μπλε γράμματα)

- αν όλοι οι πίνακες έχουν max duration, που σημαίνει ότι δεν υπάρχουν μεσαίας ή χαμηλής διάρκειας πίνακες. (αυτά τα data sets φαίνονται στους πίνακες με πράσινα γράμματα)

Στην κατηγορία 1_FocusedShot_n_FROZEN υπάρχουν ***3*** data sets που έχουν και τις 4 τιμές του πίνακα συσχέτισης 0, δηλαδή δεν έχουν καθόλου activity άρα δεν έγινε καμία αλλαγή στο συγκεκριμένο data set, ***11*** data sets με πολύ μικρή activity (το max Updates είναι 1 ή 2) και ***1*** data set που όλοι οι πίνακες είχαν την μέγιστη διάρκεια. Άρα σε ***15*** data sets δεν έχει νόημα να εκτελεστεί το fisher test. Για αυτή την κατηγορία το πλήθος των data sets στο συνδυασμό του fisher test και της δοκιμής γεωμετρίας είναι :

		Fisher Test ?		
		True	***False***	
Geome-try	***True***	1	19	**20**
	False	1	4	**5**
		2	**23**	**25**

Στην κατηγορία 2_MODERATE δεν υπάρχει κανένα data set χωρίς activity, ***5*** data sets με πολύ μικρή activity (το max Updates είναι 1 ή 2) και ***1*** data set που όλοι οι πίνακες είχαν την μέγιστη διάρκεια. Άρα σε ***6*** data sets δεν έχει νόημα να εκτελεστεί το fisher test. Για αυτή την κατηγορία το πλήθος των data sets στο συνδυασμό του fisher test και της δοκιμής γεωμετρίας είναι :

		Fisher Test ?		
		True	***False***	
Geome-try	***True***	4	19	**23**
	False	2	4	**6**
		6	**23**	**29**

Στην κατηγορία 3_FocusedShot_n_LOW δεν υπάρχει κανένα data set χωρίς activity και κανένα data set που όλοι οι πίνακες να έχουν την μέγιστη διάρκεια. Υπάρχουν ***2*** data sets με πολύ μικρή activity (το max Updates είναι 1 ή 2). Για αυτή την κατηγορία το πλήθος των data sets στο συνδυασμό του fisher test και της δοκιμής γεωμετρίας είναι :

		Fisher Test ?		
		True	False	
Geome-try	True	5	12	17
	False	2	1	3
		7	13	20

Στην κατηγορία 5_ACTIVE υπάρχει μόνο ***1*** data set με 0 αλλαγές. Για αυτή την κατηγορία το πλήθος των data sets στο συνδυασμό του fisher test και της δοκιμής γεωμετρίας είναι :

		Fisher Test ?		
		True	False	
Geome-try	True	11	6	17
	False	3	2	5
		14	8	22

Παρατηρείται πως αρκετά data sets με fisher test=false έχουν 0 στην $3^{η}$ στήλη του πίνακα συσχέτισης (notHighDur - highUpd) που σημαίνει ότι ακολουθούν σε μεγάλο βαθμό την ουσία του μοτίβου δηλ ότι σε πίνακες με μικρή διάρκεια γίνονται λίγες (εδώ μηδενικές) αλλαγές. Υπάρχουν ***5*** τέτοια data sets στη κατηγορία 1_FocusedShot_n_FROZEN, ***12*** στην κατηγορία 2_MODERATE, ***8*** στην κατηγορία 3_FocusedShot_n_LOW και ***1*** στην κατηγορία 5_ACTIVE. Γι αυτά αποτυγχάνει το fisher test (διότι έχει μια μηδενική τιμή στην $3^{η}$ στήλη του πίνακα συσχέτισης – και αναφέρθηκε ότι έστω και με μία μηδενική τιμή στον πίνακα συσχέτισης αποτυγχάνει) και πετυχαίνει το γεωμετρικό μοτίβο καθώς στην περιοχή που επιτρέπεται μόνο το 15% των πινάκων σε αυτά data sets δεν υπάρχει κανένας πίνακας,

Όπως φαίνεται στους πίνακες 5-3, 5-4, 5-5, 5-6, αν αφαιρεθούν τα data sets στα οποία δεν έχει νόημα η εκτέλεση του fisher test και κατά συνέπεια δεν μπορεί να μελετηθεί το μοτίβο του Αντίστροφου Γάμμα (δηλαδή αυτοί που είναι με *μπλε*- MaxUpd > 5, *κόκκινο*-χωρίς καθόλου activity και *πράσινο*- όλοι οι πίνακες έχουν max duration) μένουν για κάθε κατηγορία :

1_FocusedShot_n_FROZEN : 6 data sets. Η γεωμετρία πέτυχε στα 4 (67%).

2_MODERATE → 15 data sets. Η γεωμετρία πέτυχε στα 12 (80%).

3_FocusedShot_n_LOW → 13 data sets. Η γεωμετρία πέτυχε στα 10 (77%).

5_ACTIVE → 18 data sets. Η γεωμετρία πέτυχε στα 13 (72%).

Συνολικά λοιπόν από τα ***52*** σύνολα δεδομένων η γεωμετρία πέτυχε στα ***39*** από αυτά (75%).

Για να θεωρηθεί όμως ότι πέτυχε η στατιστική δοκιμή θα πρέπει να έχει υπολογισθεί TRUE και η τιμή της p_value εκτός από την επιτυχή εκτέλεση του fisher test. Τα σύνολα δεδομένων, από τα 195, που ικανοποιούν και τις δύο αυτές συνθήκες είναι :

- 1 σύνολο δεδομένων της κατηγορίας 3_FocusedShot_n_LOW (TalkingData__OWL-v3)
- 4 σύνολα δεδομένων της κατηγορίας 5_ACTIVE (cgrates__cgrates, joomla-tools__joomla-platform, MDSLab__s4t-iotronic-standalone, open-cart__opencart)

Άρα η στατιστική δοκιμή πετυχαίνει μόνο σε 5 σύνολα δεδομένων (ποσοστό 2,5%). Η δοκιμή γεωμετρίας πετυχαίνει σε 4 από αυτά,

Πίνακας 5-3 Αποτελέσματα του προγράμματος *Metrisis* για τα σύνολα δεδομένων της κατηγορίας 1_FocusedShot_n_FROZEN

1_FocusedShot_n_FROZEN	Geometry	Executed Fisher Test	# tables	maxDur	maxUpd	High Changers	*Πίνακας Συσχέτισης*			
							highDur		notHighDur	
							highUpd	notHighUpd	highUpd	notHighUpd
dlds__yii2-mlm	F	FALSE	5	4	12	4	3	1	1 20%	0
jasongrimes__silex-simpleuser	F		2	6	5	2	1	0	1 50%	0
tstack__lnav	F		2	5	1	1	0	1	1 50%	0
webadmin87__rzwebsys7	F		5	3	25	5	4	0	1 20%	0
3ev__tev_label	T		2	3	2	1	1	0	0 0%	1
accgit__acl	T		7	17	10	6	5	0	1 14%	1
BotBotMe__botbot-bot	T		3	2	12	3	3	0	0 0%	0
devture__silex-user-bundle	T		2	1	0	0	0	0	0 0%	0
duythien__blog	T		15	4	3	1	1	12	0 0%	2
EricDepagne__Astrodb	T		6	2	2	2	2	0	0 0%	4
fastpress__fastpress	T		5	2	3	2	2	0	0 0%	3
gem__oq-engine	T		49	9	13	2	2	0	0 0%	47
jadekler__git-go-d3-concertsap	T		8	8	2	2	2	3	0 0%	3
JRonak__OnlineJudge	T		6	3	1	1	1	3	0 0%	2
liujianping__scaffold	T		6	3	0	0	0	0	0 0%	0
magnus-lycka__gocddash	T		11	3	7	1	1	8	0 0%	2
mukatee__pypro	T		9	4	2	8	8	1	0 0%	0
NPRA__EmissionCalculatorLib	T		38	4	5	1	0	18	1 3%	19
royzhao__prot-coderun	T		5	2	1	1	1	0	0 0%	4
snakerflow__snakerflow	T		9	4	6	9	9	0	0 0%	0
Terry-Mao__gopush-cluster	T		5	2	0	0	0	0	0 0%	0
webnuts__post_json	T		5	3	2	2	2	1	0 0%	2
williamespindola__field	T		7	6	2	2	1	0	1 14%	5
archan937__cached_record	F	TRUE	5	5	1	2	1	1	1 20%	2
dburry__indexed_search	T		9	4	1	2	1	2	1 11%	5

Πίνακας 5-4 Αποτελέσματα του προγράμματος *Metrisis* για τα σύνολα δεδομένων της κατηγορίας 2_MODERATE

2_MODERATE	Geometry	Executed Fisher Test	# tables	maxDur	maxUpd	High Changers	*Πίνακας Συσχέτισης*			
							highDur		notHighDur	
							highUpd	notHighUpd	highUpd	notHighUpd
benoitletondor__TwitterBot	F	FALSE	4	10	8	2	0	1	2 50%	0
imbo__imbo	F		6	10	5	5	2	1	3 50%	0
lamassu__lamassu-admin	F		7	18	6	6	1	0	5 71%	1
thewhitetulip__Tasks	F		6	10	9	3	1	0	2 33%	3
aimeos__aimeos-typo3	T		3	6	6	1	1	0	0 0%	2
aiyi__go-user	T		6	10	11	6	6	0	0 0%	0
Attendly__maillist	T		7	9	2	6	6	0	0 0%	1
cartalyst__sentry	T		5	13	10	3	3	1	0 0%	1
gousiosg__github-mirror	T		23	14	8	1	1	20	0 0%	2
IamBc__abc	T		6	14	3	4	4	1	0 0%	1
imsamurai__cakephp-task-plugin	T		3	10	9	1	1	0	0 0%	1
jaybennett89__thorium-go	T		9	7	3	1	1	4	0 0%	4
lisong__code-push-server	T		13	7	4	5	5	6	0 0%	2
mozilla__mig	T		13	31	10	3	3	9	0 0%	1
mozilla__tls-observatory	T		4	43	26	2	2	2	0 0%	0
nats-io__nats-streaming-server	T		7	9	5	4	4	2	0 0%	1
nawork__nawork-uri	T		5	26	38	1	1	1	0 0%	3
scorelab__Bassa	T		3	6	9	1	1	1	0 0%	1
soapboxsys__ombudslib	T		5	9	4	3	3	1	0 0%	1
symphonycms__symphony-2	T		21	24	9	6	6	14	0 0%	1
teaminmedias-pluswerk__ke_sear	T		7	15	10	3	3	4	0 0%	0
wskm__deruv	T		22	10	4	5	5	15	0 0%	2
ZachBergh__spark-mysql-protocc	T		3	5	2	2	2	0	0 0%	1
MorpheusXAUT__eveauth	F	T	13	16	5	6	4	4	2 15%	3
ranaroussi__qtpylib	F		6	5	2	3	2	2	1 17%	1
byteball__byteballcore	T		68	12	2	18	17	48	1 1%	2
comforme__comforme	T		8	21	2	5	4	2	1 13%	1
mapbox__osm-comments-parser	T		7	13	12	3	2	3	1 14%	1
neocogent__sqlchain	T		9	8	7	4	3	1	1 11%	4

Πίνακας 5-5 Αποτελέσματα του προγράμματος Metrisis για τα σύνολα δεδομένων της κατηγορίας 3_FocusedShot_n_LOW

3_FocusedShot_n_LOW	Geometry	Executed Fisher Test	# tables	maxDur	maxUpd	High Changers	*Πίνακας Συσχέτισης*			
							highDur		notHighDur	
							highUpd	notHighUpd	highUpd	notHighUpd
milogert__ocdns	F	FALSE	7	9	10	5	2	0	3 43%	2
alextselegidis__easyappointmen	T		11	19	7	7	7	2	0 0%	2
brettkromkamp__topic_db	T		15	10	12	6	6	1	0 0%	8
curt-labs__GoSurvey	T		11	7	3	8	8	1	0 0%	2
GoBelieveIO__im_service	T		16	9	7	2	1	0	1 6%	14
h2oai__steam	T		24	10	12	16	16	8	0 0%	0
hurad__hurad	T		28	14	10	9	9	1	0 0%	18
joyplus__o2oadmin	T		13	10	1	1	0	2	0 0%	5
keybase__node-client	T		5	6	4	1	1	1	0 0%	3
kronusme__dota2-api	T		13	15	25	5	5	1	0 0%	7
lamassu__lamassu-scripts	T		7	16	11	1	1	4	0 0%	2
spaceboats__busbus	T		15	8	4	9	9	0	0 0%	6
TwitchScience__rs_ingester	T		6	7	8	1	1	1	0 0%	4
jasdel__harvester	F	TRUE	8	12	10	5	2	1	3 38%	2
n2n__rocket	F		15	7	7	5	2	3	3 20%	7
anchorcms__anchor-cms	T		20	9	3	5	3	7	2 10%	8
CityGrid__twonicorn	T		23	9	3	12	10	7	2 9%	4
gugoan__economizzer	T		19	6	2	4	3	8	1 5%	7
pw-press__web-project	T		15	11	10	5	4	4	1 7%	6
TalkingData__OWL-v3	T		41	14	14	16	15	4	1 2%	21

Πίνακας 5-6 Αποτελέσματα του προγράμματος Metrisis για τα σύνολα δεδομένων της κατηγορίας 5_ACTIVE

5_ACTIVE	Geometry	Executed Fisher Test	# tables	maxDur	maxUpd	High Changers	*Πίνακας Συσχέτισης*			
							highDur		notHighDur	
							highUpd	notHighUpd	highUpd	notHighUpd
blabla1337__skf-flask	F	FALSE	25	45	17	6	1	0	5 20%	19
HaliteChallenge__Halite-II	F		33	48	21	17	8	0	9 27%	16
AA-ALERT__frbcatdb	T		16	16	62	3	3	13	0 0%	0
arnoldasgudas__Hangfire.MySqlStorage	T		24	10	5	12	11	0	0 0%	12
nooku__joomla-todo	T		5	6	0	0	0	0	0 0%	0
pods-framework__pods	T		27	34	30	6	5	0	1 4%	21
processone__ejabberd	T		42	14	3	3	3	32	0 0%	7
torrentpier__torrentpier	T		72	126	27	7	7	40	0 0%	25
cgrates__cgrates	F	TRUE	40	190	94	14	5	2	9 23%	24
EPICPaaS__appmsgsrv	F		24	24	7	5	1	9	4 17%	10
tronsha__cerberus	F		22	24	8	10	6	6	4 18%	6
builderscon__octav	T		27	74	12	9	5	3	4 15%	15
energine-cmf__energine	T		65	29	12	10	9	34	1 2%	21
enova__landable	T		59	24	17	5	4	52	1 2%	2
foodcoopshop__foodcoopshop	T		41	67	33	10	6	21	4 10%	10
intelliants__subrion	T		56	266	15	25	19	19	6 11%	12
joomlatools__joomla-platform	T		113	17	4	6	4	5	2 2%	102
MDSLab__s4t-iotronic-standalone	T		43	26	32	6	5	8	1 2%	29
opencart__opencart	T		283	494	69	21	11	76	10 4%	186
pinterest__teletraan	T		40	52	56	6	1	16	5 13%	18
quickapps__cms	T		32	27	17	3	1	18	2 6%	11
studygolang__studygolang	T		59	46	14	14	11	24	3 5%	21

5.2 Δοκιμή γεωμετρίας

Όπως αναφέρθηκε στην αρχή του κεφαλαίου η δοκιμή γεωμετρίας πετυχαίνει ό-ταν στο γεωμετρικό μοτίβο, στην *περιοχή* που αντιστοιχεί (α) στους πίνακες που η διάρκεια τους δεν ξεπερνά το 90% της μέγιστης διάρκειας (notHighDuration) και (β) στους πίνακες που το άθροισμα των ενημερώσεών τους είναι αυστηρά με-γαλύτερο από το 15% του μέγιστου αθροίσματος ενημερώσεων ενός πίνακα σε ολόκληρο το σύνολο δεδομένων (hghlUpdates), υπάρχει λιγότερο από ***15%*** των πινάκων. Οι πίνακες που ανήκουν στην περιοχή αυτή έχουν μικρή διάρκεια ζωής (notHighDuration) αλλά πολλές ενημερώσεις (highUpd).

Στο πίνακα 5-7 φαίνεται για κάθε κατηγορία το πλήθος των data sets που ι-κανοποιούν το γεωμετρικό μοτίβο.

Πίνακας 5-7 πλήθος των data sets που ικανοποιούν το γεωμετρικό μοτίβο ανά κα-τηγορία και συνολικά

Δοκιμή Γεωμετρίας	***Πλήθος Data sets***	***Πέτυχε η δοκιμή Γεωμετρίας***
1_FocusedShot_n_FROZEN	25	20 (***80%***)
2_MODERATE	29	23 (***79%***)
3_FocusedShot_n_LOW	20	17 (***85%***)
5_ACTIVE	22	17 (***77%***)
ΣΥΝΟΛΟ των Data sets	**96**	**80 (*%*)**

Η δοκιμή γεωμετρίας αποτυχαίνει σε :

- 5 σύνολα δεδομένων για την κατηγορία 1_FocusedShot_n_FROZEN
- 6 σύνολα δεδομένων για την κατηγορία 2_MODERATE
- 3 σύνολα δεδομένων για την κατηγορία 3_FocusedShot_n_LOW
- 5 σύνολα δεδομένων για την κατηγορία 5_ACTIVE

Στη συνέχεια παρουσιάζονται τα χαρακτηριστικά των συνόλων δεδομένων που απέτυχαν σε κάθε κατηγορία.

Πίνακας 5-8 Data sets της κατηγορίας 1_FocusedShot_n_FROZEN που απέτυχε η δοκιμή γεωμετρίας για το inverseGamma μοτίβο.

GEOMETRY = FALSE 1_FocusedShot_n_FROZEN	Fisher Executed	p-Value	#Tables @Start	#Tables @End	Upd Period Months
archan937__cached_record	TRUE	FALSE	2	5	25
dlds__yii2-mlm	FALSE		5	5	38
jasongrimes__silex-simpleuser	FALSE		1	2	22
tstack__lnav	FALSE		1	2	116
webadmin87__rzwebsys7	FALSE		4	5	36

Παρατηρείται ότι μόνο σε ένα data set από αυτά που απέτυχε η δοκιμή γεωμετρία εκτελέστηκε το Fisher Test (ποσοστό 20%). Μια άλλη παρατήρηση είναι ότι και τα 5 data sets έχουν μικρό σχήμα, με μέγιστη τιμή τους 5 πίνακες, Και τα 5 data sets που αποτυγχάνουν έχουν μόνο ***έναν*** πίνακα στην κατηγορία notHighDur και highUpd (πίνακας 3) αλλά όταν οι πίνακες είναι λίγοι έστω και ένας πίνακας σε αυτή την κατηγορία μπορεί να κάνει false την γεωμετρική δοκιμή.

Πίνακας 5-9 Data sets της κατηγορίας 2_MODERATE που απέτυχε η δοκιμή γεωμετρίας για το inverseGamma μοτίβο.

GEOMETRY = FALSE 2_MODERATE	Fisher Executed	p-Value	#Tables @Start	#Tables @End	Upd Period Months
benoitletondor__TwitterBot	FALSE		1	4	48
imbo__imbo	FALSE		3	4	93
lamassu__lamassu-admin	FALSE		1	5	37
MorpheusXAUT__eveauth	TRUE	FALSE	9	12	4
ranaroussi__qtpylib	TRUE	FALSE	4	6	32
thewhitetulip__Tasks	FALSE		1	6	40

Στον πίνακα 5-9 παρατηρείται ότι από τα 6 σύνολα δεδομένων της κατηγορίας 2_MODERATE που απέτυχε η δοκιμή γεωμετρία, μόνο σε 2 από αυτά εκτελέστηκε το Fisher Test (ποσοστό 33%). Μια άλλη παρατήρηση είναι ότι και σε αυτή την κατηγορία τα 5 data sets που απέτυχαν έχουν μικρό σχήμα και μόνο ένα έχει πάνω από 10 πίνακες.

Πίνακας 5-10 Data sets της κατηγορίας 3_FocusedShot_n_LOW που απέτυχε η δοκιμή γεωμετρίας για το inverseGamma μοτίβο.

GEOMETRY = FALSE 3_FocusedShot_n_LOW	Fisher Executed	p-Value	#Tables @Start	#Tables @End	Upd Period Months
jasdel__harvester	TRUE	FALSE	4	6	0
milogert__ocdns	FALSE		3	5	2
n2n__rocket	TRUE	FALSE	10	10	48

Παρατηρείται ότι από τα 3 σύνολα δεδομένων της κατηγορίας 3_FocusedShot_n_LOW που απέτυχε η δοκιμή γεωμετρία, σε 2 από αυτά που εκτελέστηκε το Fisher Test (ποσοστό 67%).

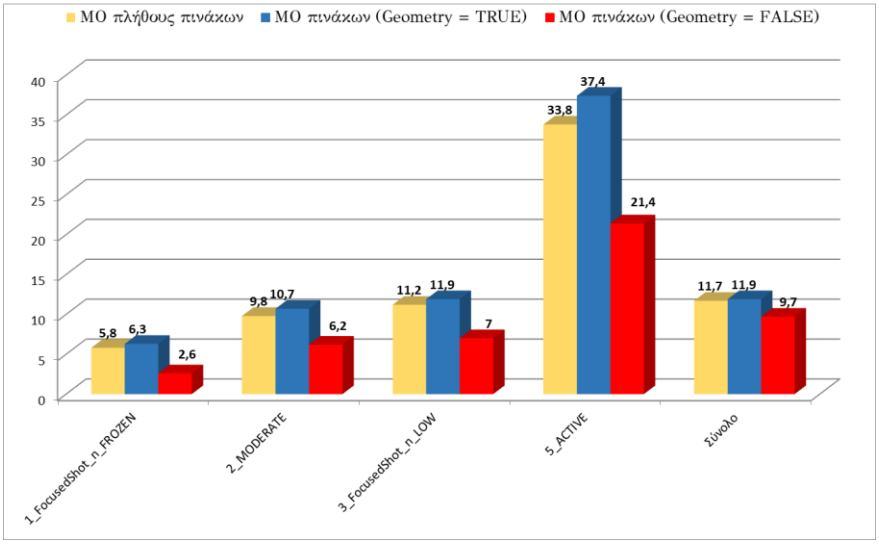
Πίνακας 5-11 Data sets της κατηγορίας 5_ACTIVE που απέτυχε η δοκιμή γεωμετρίας για το inverseGamma μοτίβο.

GEOMETRY = FALSE 5_ACTIVE	Fisher Executed	p-Value	#Tables @Start	#Tables @End	Upd Period Months
blabla1337__skf-flask	FALSE		4	19	51
cgrates__cgrates	TRUE	TRUE	9	21	88
EPICPaaS__appmsgsrv	TRUE	FALSE	10	23	6
HaliteChallenge__Halite-II	FALSE		9	22	39
tronsha__cerberus	TRUE	FALSE	11	22	77

Παρατηρείται ότι από τα 5 σύνολα δεδομένων της κατηγορίας 5_ACTIVE που απέτυχε η δοκιμή γεωμετρία, σε 3 από αυτά εκτελέστηκε το Fisher Test (ποσοστό 60%) και μόνο σε ένα data set η p-value ήταν κάτω του στατιστικά αποδεκτού ορίου του 5% (η p-value υπολογίστηκε true μόνο σε ένα σύνολο δεδομένων από αυτά που απέτυχε η δοκιμή γεωμετρίας, το cgrates__cgrates).

Από τα 19 σύνολα δεδομένων που απέτυχε η δοκιμή γεωμετρία το fisher test εκτελέστηκε σε 8 από αυτά και μόνο σε ένα η p-value ήταν κάτω του αποδεκτού ορίου του 5%. Όσον αφορά το μέγεθος του σχήματος υπάρχει μια ένδειξη ότι τα σύνολα δεδομένων που αποτυγχάνουν στη δοκιμή γεωμετρίας έχουν μικρό πλήθος πινάκων σε σχέση με τον μέσο όρο του πλήθος των πινάκων της αντίστοιχης κατηγορίας τους, χωρίς όμως να σημαίνει ότι σε όλα τα σύνολα δεδομένων με μικρό σχήμα αποτυγχάνει η δοκιμή γεωμετρίας.

Στο σχήμα 5-1 φαίνεται για τα σύνολα δεδομένων κάθε κατηγορίας και για το σύνολο των data sets, (α) ο μέσος όρος του πλήθους των πινάκων, (β) ο μέσος όρος του πλήθους των πινάκων για τα σύνολα των δεδομένων που ικανοποιούν τη δοκιμή της γεωμετρίας και (γ) ο μέσος όρος του πλήθους των πινάκων για τα σύνολα των δεδομένων που δεν ικανοποιούν τη δοκιμή της γεωμετρίας. Παρατηρείται ότι, ο μικρότερος μέσος όρος είναι αυτός που αντιστοιχεί στο πλήθος των πινάκων για τα σύνολα δεδομένων που δεν ικανοποιούν τη δοκιμή γεωμετρίας. Φαίνεται λοιπόν να υπάρχει μια ένδειξη ότι όταν ένα σύνολο δεδομένων έχει πολλούς πίνακες τείνει να ικανοποιεί την γεωμετρία.



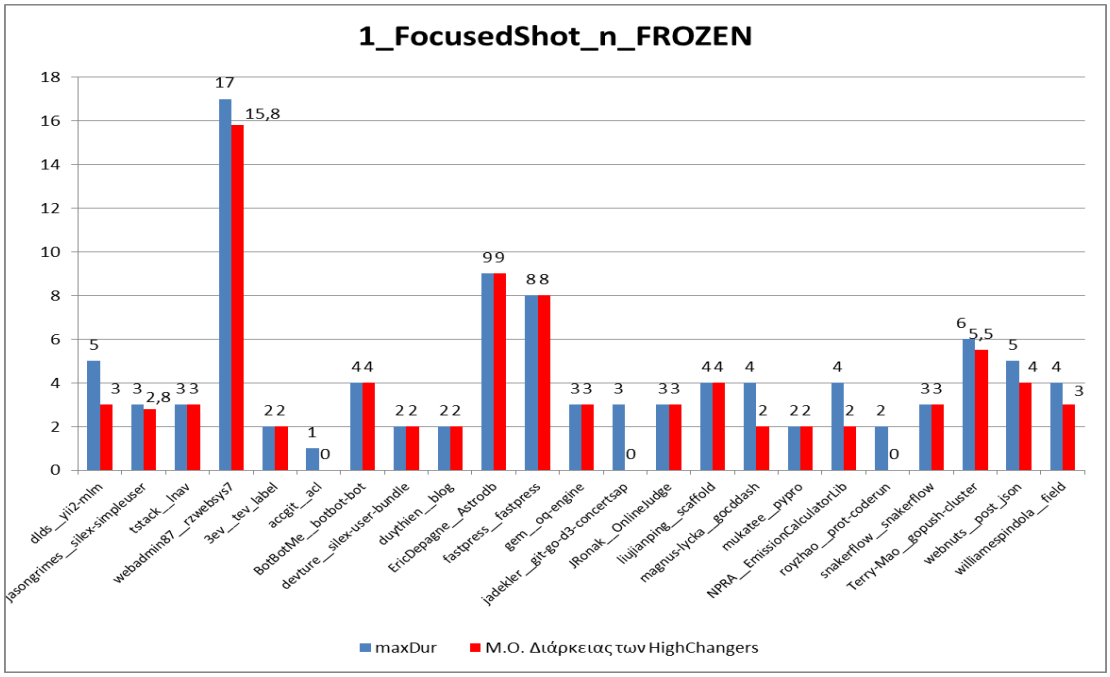
Σχήμα 5-1 – Συσχέτιση του (α) μέσου όρου του πλήθους των πινάκων, (β) μέσου όρου του πλήθους των πινάκων για τα σύνολα των δεδομένων που ικανοποιούν τη δοκιμή της γεωμετρίας (γ) μέσου όρου του πλήθους των πινάκων για τα σύνολα των δεδομένων που δεν ικανοποιούν τη δοκιμή της γεωμετρίας.

5.3 Ποια είναι η διάρκεια ζωής των high changers πινάκων;

Όπως περιγράφεται και στην αρχή του κεφαλαίου ως high changer χαρακτηρίζεται ένας πίνακας αν οι ενημερώσεις του είναι περισσότερες του 15% της μέγιστης τιμής των ενημερώσεων που έγιναν σε έναν πίνακα στο σύνολο δεδομένων.

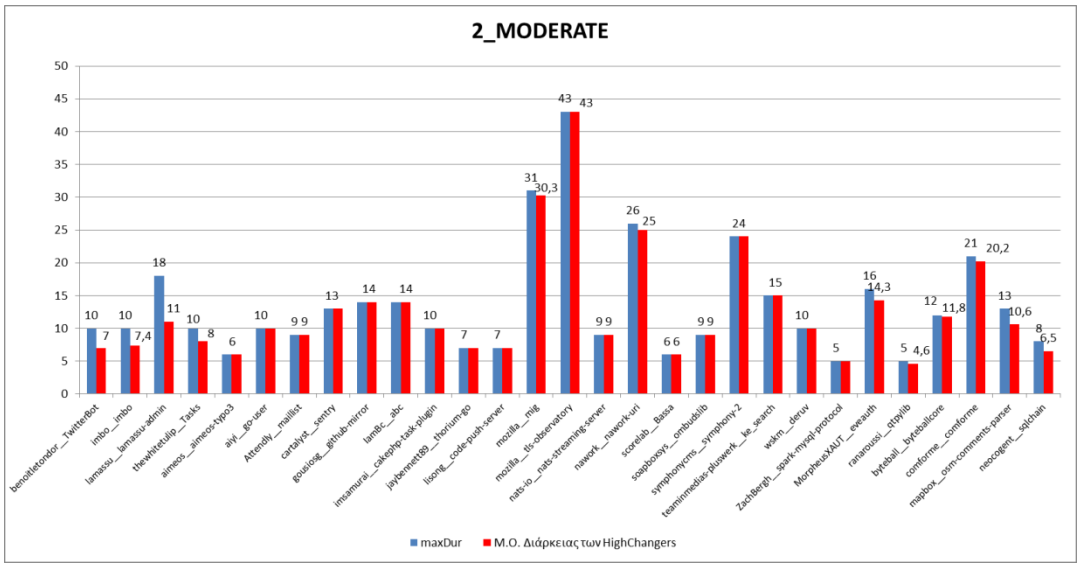
Ενδιαφέρον παρουσιάζει η μελέτη της διάρκειας των high changers (δηλαδή ο μέσος όρος για κάθε σύνολο δεδομένων) σε σχέση με την max duration του κάθε συνόλου δεδομένων.

Όπως φαίνεται στο σχήμα 5-2 στην κατηγορία 1_FocusedShot_n_FROZEN υπάρχουν 13 σύνολα δεδομένων από τα 25 (ποσοστό 52%) που όλοι οι high changers πίνακες έχουν την max duration και 4 σύνολα δεδομένων που η απόκλιση τους είναι μικρότερη το 0.5.



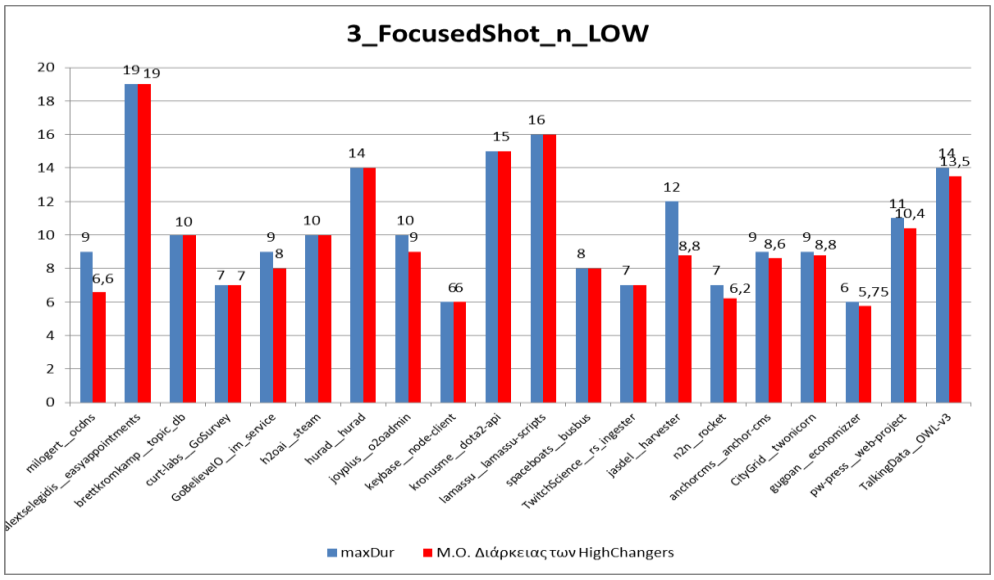
Σχήμα 5-2 – Συσχέτιση της διάρκειας των high changers με την max duration του κάθε συνόλου δεδομένων της κατηγορίας 1_FocusedShot_n_FROZEN.

Όπως φαίνεται στο σχήμα 5-3 στην κατηγορία 2_MODERATE υπάρχουν 16 σύνολα δεδομένων από τα 29 (ποσοστό 55,2%) που σε αυτά, όλοι οι high changers πίνακες έχουν την max duration και 2 σύνολα δεδομένων που η απόκλιση τους είναι μικρότερη το 0.5.



Σχήμα 5-3 – Συσχέτιση της διάρκειας των high changers με την max duration του κάθε συνόλου δεδομένων της κατηγορίας 2_MODERATE

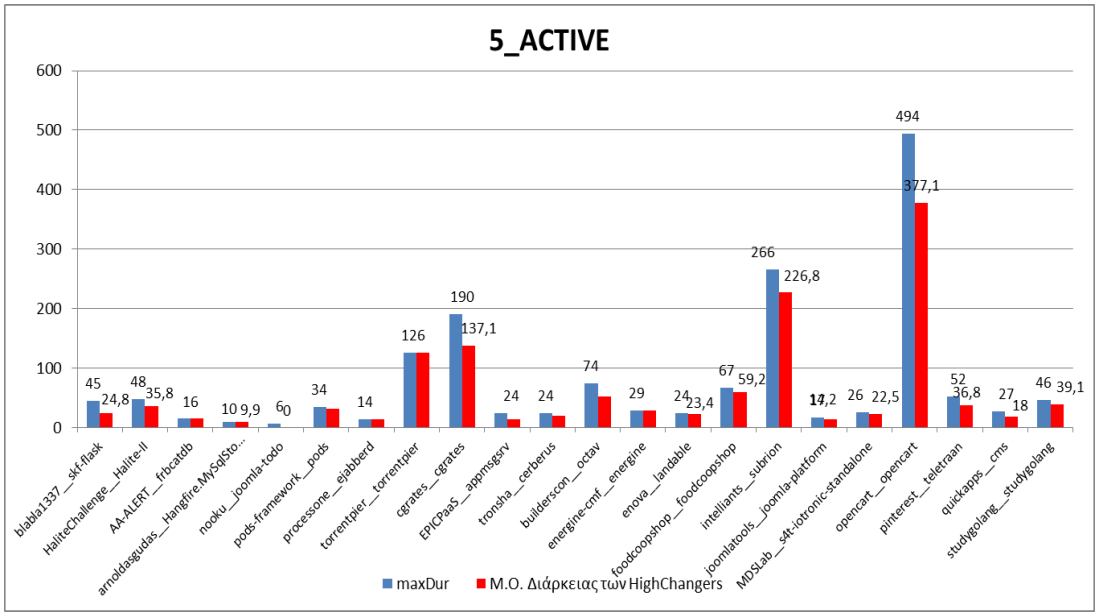
Όπως φαίνεται στο σχήμα 5-4 στην κατηγορία 3_FocusedShot_n_LOW υπάρχουν 10 σύνολα δεδομένων από τα 20 (ποσοστό 50%) που σε αυτά, όλοι οι high changers πίνακες έχουν την max duration και 4 σύνολα δεδομένων που η απόκλιση τους είναι μικρότερη το 0.5.



Σχήμα 5-4 - Συσχέτιση της διάρκειας των high changers με την max duration του κάθε συνόλου δεδομένων της κατηγορίας 3_FocusedShot_n_LOW

Όπως φαίνεται στο σχήμα 5-5 στην κατηγορία 5_Active υπάρχουν 3 σύνολα δεδομένων από τα 22 (ποσοστό 13.6%) που σε αυτά, όλοι οι high changers πίνα-

κες έχουν την max duration. Σε αυτή την κατηγορία υπάρχουν όμως και κάποιες σχετικά μεγάλες αποκλίσεις στην διάρκεια ζωής των high changers πινάκων και της μέγιστης διάρκειας του συνόλου δεδομένων.



Σχήμα 5-5 - Συσχέτιση της διάρκειας των high changers με την max duration του κάθε συνόλου δεδομένων της κατηγορίας 5_Active

Συνολικά μπορούμε να πούμε ότι οι πίνακες που έχουν τις περισσότερες αλλαγές σε ένα data set κατά ένα μεγάλο ποσοστό ζουν σχεδόν στη μέγιστη διάρκεια. Επίσης όπως παρατηρήθηκε τα σύνολα δεδομένων ικανοποιούν τη γεωμετρία σε μεγάλο βαθμό. Από αυτό εξάγεται το συμπέρασμα ότι γενικά υπάρχουν λίγοι high changers πίνακες με μικρή διάρκεια ζωής.

ΚΕΦΑΛΑΙΟ 6

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

6.1 Συμπεράσματα

Όσον αφορά το ερώτημα *τι έρευνα έχει γίνει πάνω στον τομέα της εξέλιξης του σχήματος και τι ερωτήματα έχουν απαντηθεί* στις υπό μελέτη έρευνες δίνονται οι εξής απαντήσεις *:*

(α) οι πολλές αλλαγές εντοπίζονται στην αρχή του κύκλου ζωής του σχήματος.

(β) η εξέλιξη του σχήματος παρουσιάζει μια έλξη προς την ακαμψία, δηλαδή την τάση να ελαχιστοποιήσει την εξέλιξη όσο το δυνατόν περισσότερο και

(γ) αν οι αλλαγές στο σχήμα της ΒΔ δεν γίνουν ταυτόχρονα με αλλαγές στον κώδικα της εφαρμογής μπορεί να οδηγήσουν σε απώλεια δεδομένων, application crash ή μείωση της απόδοσης.

Στο δεύτερο μέρος της εργασίας μελετήθηκαν ερωτήσεις για το μοτίβο Γάμμα και τους wide πίνακες σε 35 σύνολα δεδομένων.

- Ποια είναι η πιθανότητα να ζήσουν οι wide πίνακες; Υπάρχει ισχυρή πιθανότητα να επιβιώσει ένας wide πίνακας Σε 16 data sets επιβίωσαν όλοι οι wide πίνακες Σε 28 data sets επιβίωσαν πάνω από τους μισούς wide πίνακες

- Δεν υπάρχει ισχυρή ένδειξη στα υπό μελέτη projects ότι οι wide πίνακες επιβιώνουν σε πολύ μεγαλύτερο ποσοστό από τους not wide πίνακες. Οι αλλαγές που συμβαίνουν στα projects οδηγούν σε διαγραφή πινάκων που έχουν μικρή συσχέτιση με το μέγεθος τους
- *Στα υπό μελέτη projects η πιθανότητα να επιβιώσουν οι wide πίνακες όταν είναι λίγοι είναι αρκετά μεγάλη*
- ***WMLTS? (Wide More Likely To Survive)*** Σε 25 από τα 35 συνολικά projects (71%) η πιθανότητα επιβίωσης των wide πινάκων είναι μεγαλύτερη από την πιθανότητα επιβίωσης των not wide πινάκων.
- Από τα 22 projects που πέτυχε το ***γεωμετρικό μοτίβο*** στα 20 (ποσοστό 91%) το WMLTS? είναι TRUE
- Σε 16 από τα 35 συνολικά projects (δηλαδή σε ποσοστό 46%) πέτυχε η εκτέλεση του fisher test σε κανένα όμως από αυτά η p_value δεν ήταν κάτω του αποδεκτού ορίου του 5%.
- Όσον αφορά το Μοτίβο Gamma που μελετήθηκε σε 35 projects και συμπεραίνεται πως ισχύει σε μεγάλο βαθμό η γεωμετρική δοκιμή αλλά δεν επαληθεύεται η στατιστική δοκιμή.

Όσον αφορά το Μοτίβο του αντίστροφου Γ, ερευνήθηκε σε πόσα από τα 195 σύνολα δεδομένων υπάρχουν τα κατάλληλα χαρακτηριστικά ώστε να μπορεί να μελετηθεί το αντίστοιχο μοτίβο. Εντοπίστηκαν ***52*** σύνολα δεδομένων που διαθέτουν τα κατάλληλα χαρακτηριστικά. Από αυτά η γεωμετρία πέτυχε στα ***39*** (75%) και η στατιστική δοκιμή (εκτελέστηκε το fisher test και p_value μικρότερη του 5%) μόνο σε 4 σύνολα δεδομένων (***2,5%***)

Ανά κατηγορία το fisher test εκτελέστηκε σε

1_FocusedShot_n_FROZEN : 6 data sets. Η γεωμετρία πέτυχε στα 4 (67%).

2_MODERATE → 15 data sets. Η γεωμετρία πέτυχε στα 12 (80%).

3_FocusedShot_n_LOW → 13 data sets. Η γεωμετρία πέτυχε στα 10 (77%).

5_ACTIVE → 18 data sets. Η γεωμετρία πέτυχε στα 13 (72%).

Συνολικά μπορούμε να πούμε ότι οι πίνακες που έχουν τις περισσότερες αλλαγές (high changers) σε ένα data set κατά ένα μεγάλο ποσοστό ζουν σχεδόν στη μέγιστη διάρκεια. Στην μόνη κατηγορία που αυτό το ποσοστό μειώνεται πολύ είναι στην κατηγορία 5_Active.

6.2 Μελλοντικές Εργασίες

Σε μελλοντικές εργασίες θα μπορούσαν να αναλυθεί η ισχύς και των άλλων προτύπων, που παρουσιάζονται στις μελέτες του κεφαλαίου 3 ανά κατηγορία. Αυτά τα πρότυπα είναι :

- Το πρότυπο Κομήτη (Comet) που μελετά την αλληλεπίδραση του μεγέθους ενός πίνακα κατά τη γέννησή του με το συνολικό ποσό των ενημερώσεων.
- Το πρότυπο άδειου Τριγώνου (Empty Triangle) που εξετάζει την αλληλεπίδραση της γέννησης ενός πίνακα με τη συνολική διάρκεια.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[CGMM13] Anthony Cleve, Maxime Gobert, Loup Meurice, Jerome Maes, Jens Weber. Understanding database schema evolution: A case study. Science of Computer Programming 97 (2015), pp. 113–121.

[LiNe09] Dien-Yen Lin, Iulian Neamtiu. Collateral Evolution of Applications and Databases. In Proceedings of the IWPSE-Evol'09, August 24–25,2009, Amsterdam.

[QiLS13] D.Qiu, B.Li, Z.Su. An empirical analysis of the co-evolution of schema and code in database applications. In Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2013, pp.125–135.

[SkVZ14] I. Skoulis, P. Vassiliadis, A. Zarras. Open-source databases: Within, out-side, or beyond Lehman's laws of software evolution? In Proceedings of 26th International Conference on Advanced Information Systems Engineering (CAiSE 2014), pp 379-393

[SkVZ15] Ioannis Skoulis, Panos Vassiliadis, Apostolos V. Zarras. Growing up with stability: How open-source relational databases evolve. Information Systems 53 (2015), pp. 363–385.

[Vass21] P. Vassiliadis. Profiles of Schema Evolution in Free Open Source Software Projects. Accepted at 37th IEEE International Conf. on Data Engineering (ICDE 2021), 19-22 April 2021, Chania, Greece.

[VaZa17] Panos Vassiliadis, Apostolos V. Zarras. Schema Evolution Survival Guide for Tables: Avoid Rigid Childhood and You 're En Route to a Quiet Life. Journal of Data Semantics (JODS), 6(4) (2017), pp 221-241.

[VaZS16] Panos Vassiliadis, Apostolos V. Zarras, Ioannis Skoulis. Gravitating to Rigidity: Patterns of Schema Evolution–and its Absence–in the Lives of Tables. Information Systems 63(2017), pp. 24–46.

[WuNe11] S. Wu, I. Neamtiu. Schema evolution analysis for embedded databases. In Proceedings of the 2011 IEEE 27th International Conference on Data Engineering Workshops, ICDEW '11, (2011), pp 151-156.

Συντομο Βιογραφικο

Η Θεολογία Καλάκου είναι μεταπτυχιακή φοιτήτρια στο Τμήμα Επιστήμης και Μηχανικής Υπολογιστών του Πανεπιστημίου Ιωαννίνων. Έλαβε το πτυχίο από το Τμήμα Εφαρμοσμένης Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών το 1993. Εργάστηκε ως Προγραμματίστρια – Αναλύτρια Τραπεζικών και Βιομηχανικών εφαρμογών για 8 χρόνια και από το 2000 εργάζεται ως καθηγήτρια μέσης εκπαίδευσης.