ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΘΕΜΑ: Συστήματα Μητρώου Ειδικών στο πλαίσιο της Διαχείρισης Γνώσης (Yellow pages and Knowledge Management Systems)

Εκπόνηση διπλωματικής εργασίας: Βασιλική Κολιοπούλου (ΑΜ 141041)

Επιβλέπων Καθηγητής: Χρήστος Σκουρλάς

Αθήνα, Ιούλιος 2021

UNIVERSITY OF WEST ATTICA
FACULTY OF ENGINEERING
DEPARTMENT OF INFORMATICS AND COMPUTER SCIENCE

Diploma Thesis

Yellow pages and Knowledge Management Systems

Student: Vasiliki Koliopoulou
Registration Number: 141041

Supervisor: Prof. Christos Skourlas

Athens, July 2021

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Αθήνα, Ιούλιος 2021

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

Επιβλέπων καθηγητής:
Χρήστος Σκουρλάς

Christos Skourlas
Digitally signed by Christos Skourlas
Date: 2021.07.16 11:40:01 +03'00'

Μέλος επιτροπής:
Κλειώ Σγουροπούλου

KLEIO SGOUROPOULOU
Digitally signed by KLEIO SGOUROPOULOU
Date: 2021.07.16 12:34:24 +03'00'

Μέλος επιτροπής:
Βασίλειος Μάμαλης

Vasileios Mamalis
Digitally signed by Vasileios Mamalis
Date: 2021.07.17 01:20:43 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η κάτωθι υπογεγραμμένη Βασιλική Κολιοπούλου του Αθανασίου, με αριθμό μητρώου 141041, φοιτήτρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από εμένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

«Με επιφύλαξη παντός δικαιώματος δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία».

Η Δηλούσα

Κολιοπούλου Βασιλική

Υπογραφή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ Κολιοπούλου Βασιλική, Ιούλιος, 2021

Περίληψη

Στην παρούσα διπλωματική εργασία στο πλαίσιο του διεπιστημονικού τομέα της διαχείρισης γνώσης (Knowledge Management) γίνεται μελέτη των συστημάτων μητρώου ειδικών (Yellow Pages Systems). Τα συστήματα αυτά διευκολύνουν την αναζήτηση ειδικών (experts), δηλαδή ανθρώπων εντός ή εκτός οργανισμού με την κατάλληλη εξειδίκευση και τα κατάλληλα προσόντα. Στόχος είναι η επικοινωνία μαζί τους και η συνεργασία. Στο πλαίσιο της εργασίας αναπτύσσεται πιλοτική εφαρμογή ενός τέτοιου συστήματος και η εφαρμογή δοκιμάζεται με δεδομένα ελέγχου (datasets).
Γίνεται ανάπτυξη ενός συστήματος για την αποθήκευση και επεξεργασία των δεδομένων. Κατά την επεξεργασία των δεδομένων χρησιμοποιούνται blocking και filtering τεχνικές για να ομαδοποιηθούν τα δεδομένα. Για την υλοποίηση μελετήθηκαν ως κύρια εργαλεία το εργαλείο MonetDB και η γλώσσα προγραμματισμού python. Το εργαλείο MonetDB είναι σύστημα διαχείρισης βάσης δεδομένων ανοιχτού κώδικα με προσανατολισμό στις στήλες και είναι πολύ αποτελεσματικό στη διαχείριση μεγάλου όγκου δεδομένων.
Για τη μελέτη μας και την εξαγωγή των αποτελεσμάτων επιλέχθηκε μια μελέτη περίπτωσης σε έναν κλάδο ειδικών που είναι οι τεχνικοί επισκευής φωτογραφικών μηχανών. Για να επιτευχθεί ο σκοπός που είναι να δούμε πως μπορεί να δουλέψει ένα σύστημα μητρώου ειδικών έχοντας ένα μεγάλο όγκο δεδομένων, χρησιμοποιούνται τρία datasets. Το κεντρικό dataset είναι αυτό με τα βιογραφικά τεχνικών επισκευής φωτογραφικών μηχανών όπου υπάρχουν πληροφορίες για κάθε ειδικό και το επαγγελματικό του προφίλ. Για να εξεταστούν οι δυνατότητες που παρέχει το εργαλείο MonetDB χρησιμοποιούνται δύο μεγαλύτερα και γνωστά datasets τα οποία περιέχουν πληροφορίες σχετικές με φωτογραφικές μηχανές που υπάρχουν σε διάφορες ιστοσελίδες για πώληση.

Λέξεις-κλειδιά: Διαχείριση Γνώσης, Συστήματα Μητρώου Ειδικών, MonetDB, python

Abstract

This thesis is dedicated to the research of Yellow Pages Systems based on the interdisciplinary field of knowledge management (Knowledge Management). These systems enable the search for experts, who are people inside or outside an organization with the appropriate specialization and the appropriate qualifications. The goal is to communicate with them and cooperate. As part of the work, a pilot application of such a system is developed and the application is tested with control datasets.

Specifically, a system is developed concerning data storage and processing. Blocking and filtering techniques are used for their processing in order to group data. The MonetDB tool and the python programming language are the main tools that are used to implement the system. The MonetDB tool is a column-oriented open source database management system and is very effective in managing large scale data.

A case study in a branch of experts who are the camera repair technicians was selected to be used in this study. To achieve the goal of observing how a Yellow Page System can work with a large scale of data, three datasets are used. The central dataset is the one with the curriculums of camera repair technicians where there is information for each specialist and his professional profile. To test the capabilities of the MonetDB tool, two larger and well-known datasets are used which contain information about cameras that are available on various websites for sale.

Keywords: Knowledge Management, Community Yellow Pages, MonetDB, python

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέπων καθηγητή μου, κύριο Χρήστο Σκουρλά, για την εμπιστοσύνη που μου έδειξε εξ' αρχής, αναθέτοντάς μου το συγκεκριμένο θέμα, τις υποδείξεις του, τις συμβουλές του, την επιμονή του, το αμείωτο ενδιαφέρον του και τη συνεχή του υποστήριξη από την αρχή μέχρι το τέλος.

Τέλος, θα ήθελα εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου, Θανάση, Χριστίνα και Θοδωρή, και τις φίλες μου, Μαρίνα και Σύλβα, για όλη τη στήριξη, τη συμπαράσταση και την κατανόησή τους, καθ' όλη τη διάρκεια των σπουδών μου.

Πίνακας Περιεχομένων

Κεφάλαιο 1. Εισαγωγή

Στην παρούσα διπλωματική εργασία θα μας απασχολήσει η διαχείριση γνώσης (Knowledge Management) και συγκεκριμένα η μελέτη των Συστημάτων Μητρώου Ειδικών (Yellow Pages Systems).Η αποτελεσματική διαχείριση της γνώσης είναι ζωτικής σημασίας για τις επιχειρήσεις και τους οργανισμούς στη σημερινή οικονομία. Η δημιουργία και η διάδοση της γνώσης γίνονται όλο και πιο σημαντικοί παράγοντες ανταγωνισμού μεταξύ των οργανισμών. Πλέον οι οργανισμοί συχνά αναζητούν άτομα που διαθέτουν εξειδικευμένη γνώση σε τομείς που ενδιαφέρουν τον οργανισμό. Ένα μέσο για τη διευκόλυνση της αναζήτησης ανθρώπων εντός ή εκτός οργανισμού με την κατάλληλη εξειδίκευση και τα κατάλληλα προσόντα είναι η δημιουργία και χρήση των Συστημάτων Μητρώου Ειδικών. Αυτά διευκολύνουν την αναζήτηση ειδικών (experts) και στόχος τους είναι η επικοινωνία μαζί τους και η συνεργασία.

Η παρούσα διπλωματική εργασία χωρίζεται σε δύο μέρη.

Στο πρώτο μέρος, το οποίο αποτελεί το γενικό μέρος της, αναφέρεται στην έννοια της διαχείρισης γνώσης και αναλύονται τα Συστημάτων Μητρώου Ειδικών.

Στο δεύτερο μέρος στόχος είναι η ανάπτυξη ενός συστήματος για την αποθήκευση και επεξεργασία των δεδομένων. Για την υλοποίηση μελετήθηκαν και χρησιμοποιούνται ως κύρια εργαλεία το εργαλείο MonetDB και η γλώσσα προγραμματισμού python. Το εργαλείο MonetDB είναι σύστημα διαχείρισης βάσης δεδομένων ανοιχτού κώδικα με προσανατολισμό στις στήλες και είναι πολύ αποτελεσματικό στη διαχείριση μεγάλου όγκου δεδομένων.

Για τη μελέτη μας και την εξαγωγή των αποτελεσμάτων και των συμπερασμάτων επιλέχθηκε μια μελέτη περίπτωσης σε έναν κλάδο ειδικών που είναι οι τεχνικοί επισκευής φωτογραφικών μηχανών.

Σκοπός μας είναι να εξετάσουμε πως μπορεί να δημιουργηθεί και να χρησιμοποιηθεί ένα Σύστημα Μητρώου Ειδικών που βασίζεται σε μεγάλο όγκο δεδομένων.

Χρησιμοποιούνται τρία datasets. Το κεντρικό dataset περιλαμβάνει τα βιογραφικά τεχνικών επισκευής φωτογραφικών μηχανών, πληροφορίες για κάθε ειδικό και το επαγγελματικό του προφίλ. Το dataset αυτό κατασκευάστηκε για τις ανάγκες της Μελέτης περίπτωσης. Για να εξεταστούν οι δυνατότητες που παρέχει το εργαλείο MonetDB χρησιμοποιούνται και δύο μεγαλύτερα και γνωστά datasets τα οποία περιέχουν πληροφορίες σχετικές με φωτογραφικές μηχανές που υπάρχουν σε διάφορες ιστοσελίδες για πώληση.

Κεφάλαιο 2. Διαχείριση Γνώσης

Η γνώση και η διαχείριση γνώσης είναι θέματα που λαμβάνουν όλο και αυξανόμενη προσοχή από διάφορους κλάδους. Η διαχείριση γνώσης συγκεκριμένα είναι ένα φλέγον ζήτημα σήμερα στον κόσμο της βιομηχανίας, στον κόσμο της έρευνας και στον κόσμο των πληροφοριών.

Η αποτελεσματική διαχείριση της γνώσης είναι ζωτικής σημασίας στη σημερινή οικονομία και συνδέεται με δραστηριότητες καταγραφής της γνώσης, ανακάλυψης της γνώσης, διαμοιρασμού της γνώσης κ.λπ. Η δημιουργία και η διάδοση της γνώσης γίνονται όλο και πιο σημαντικοί παράγοντες ανταγωνισμού. Η γνώση θεωρείται κεφάλαιο του οργανισμού (asset) και η σχετική γνώση ενσωματώνεται σε προϊόντα, χρησιμοποιείται στη διαχείριση έργων και στη λήψη αποφάσεων κ.λπ.

Ενδιαφέρει ιδιαίτερα η άρρητη γνώση των υπαλλήλων που είναι διασκορπισμένοι σε διάφορες φυσικές τοποθεσίες και συνδέονται με υπολογιστές, smartphone και άλλες συσκευές μέσω του παγκόσμιου Διαδικτύου.

Ενώ η γνώση θεωρείται όλο και πιο συχνά ως διανοητικό κεφάλαιο του οργανισμού, υπάρχουν μερικά χαρακτηριστικά της γνώσης που είναι ριζικά διαφορετικά από άλλους πολύτιμους πόρους. Σε αυτά τα γνωσιακά χαρακτηριστικά περιλαμβάνουν τα ακόλουθα (Dalkir 2011):

• Η χρήση της γνώσης δεν συνεπάγεται την κατανάλωσή της.

• Η μεταφορά γνώσης δεν οδηγεί σε απώλεια της γνώσης.

• Η γνώση είναι άφθονη, αλλά η ικανότητα αποτελεσματικής χρήσης της είναι συνήθως περιορισμένη.

• Πολλές από τις πολύτιμες γνώσεις ενός οργανισμού δεν παραμένουν εντός του οργανισμού για μελλοντική χρήση και αξιοποίηση

Η έλευση του Διαδικτύου και ο Παγκόσμιος Ιστός δημιούργησε «απεριόριστους» πόρους γνώσης που είναι συχνά διαθέσιμοι σε όλους. Κάποιοι ειδικοί αναφέρονται στην αυγή της Εποχής της Γνώσης σε αντικατάσταση της βιομηχανικής εποχής. Σύμφωνα με τη βιβλιογραφία, π.χ., (Drucker 1994), (Barth, 2000) παλαιότερα σχεδόν οι μισοί από όλους τους εργαζόμενους στις βιομηχανικές χώρες συνεισέφεραν στην παραγωγή βιομηχανικών προϊόντων. Μέχρι το 2000, μόνο το 20 τοις εκατό των εργαζομένων συνεισέφερε στη βιομηχανική εργασία. Ο Davenport (2005, σελ. 5) γράφει ότι για τη διαχείριση της γνώση ασχολούνται 25-50% του εργατικού δυναμικού των ΗΠΑ.

Το έργο γνώσης απαιτεί περισσότερη συνεργασία. Μια εταιρεία είναι ανταγωνιστική και τελικά η βιωσιμότητά της εξαρτάται από το αν βασίζει τη δραστηριότητά της σε ότι γνωρίζει συλλογικά, από το πόσο αποτελεσματικά χρησιμοποιεί αυτό που γνωρίζει και από το πόσο γρήγορα αποκτά και χρησιμοποιεί τη νέα γνώση (Davenport και Prusak 1998).

Ένας οργανισμός που διδάσκεται (learning organization) στην εποχή της γνώσης είναι αυτός που μαθαίνει, θυμάται και ενεργεί με βάση τις καλύτερες διαθέσιμες

πληροφορίες, γνώσεις και τεχνογνωσία.

Όλες αυτές οι εξελίξεις έχουν δημιουργήσει μια έντονη ανάγκη για μια στοχευμένη και συστηματική προσέγγιση για τη δημιουργία και τη διανομή της βάσης γνώσης της εταιρείας. Η βάση γνώσης, είναι γνωστή και ως Οργανωσιακή Γνώση (ή μνήμη). Δεν αποσκοπεί στην αντικατάσταση της ατομικής γνώσης αλλά στη συμπλήρωσή της κάνοντας την ισχυρότερη, πιο συνεκτική και ευρύτερα εφαρμόσιμη.

Η Διαχείριση Γνώσης καθορίζει μια στοχευμένη και συστηματική προσέγγιση στη δημιουργία και την πλήρη αξιοποίηση της βάσης γνώσης του οργανισμού, σε συνδυασμό με τις ανθρώπινες δυνατότητες: ατομικές δεξιότητες, ικανότητες, σκέψεις, καινοτομία και ιδέες. Σκοπός είναι η δημιουργία μιας πιο αποτελεσματικής οργάνωσης της εταιρείας.

Όλο και περισσότερο, οι εταιρείες αλλάζουν με βάση τη γνώση (αυτά που γνωρίζουν). Η νέα επιχείρηση πρέπει να γνωρίζει πώς να κάνει νέα πράγματα καλά και γρήγορα. (Davenport and Prusak 1998, 13).

2.1 Η Γνώση στη βιβλιογραφία

Με σκοπό την καλύτερη κατανόηση του όρου διαχείριση γνώσης είναι χρήσιμο να γίνει μια αναφορά στο πως ορίζεται η γνώση στη βιβλιογραφία. Οι Davenport και Prusak ορίζουν την γνώση σαν ένα μείγμα των εμπειριών, των αξιών και των πληροφοριών αλλά και της διορατικότητας και της διαίσθησης των ανθρώπων του οργανισμού. Η γνώση παρέχει ένα περιβάλλον και ένα πλαίσιο για νέες εμπειρίες και πληροφορίες. Η γνώση δημιουργείται και εφαρμόζεται από τους ανθρώπους του οργανισμού. Η καταγεγραμμένη γνώση υπάρχει στα έγγραφα, στις αποθήκες δεδομένων αλλά και στις διαδικασίες, τις πρακτικές, και τους κανόνες που υιοθετεί ο οργανισμός (Davenport & Prusak, 1998), (Dalkir, 2011, σελ. 14).

Η γνώση μπορεί να διαχωριστεί σε διάφορες κατηγορίες ανάλογα με το είδος της. Ένας διαδεδομένος διαχωρισμός της είναι αυτός σε ρητή (explicit) και άρρητη (tacit).

Άρρητη γνώση είναι η γνώση που έχουν οι εργαζόμενοι του οργανισμού. Η γνώση αυτή είναι δύσκολο να τυποποιηθεί και να καταγραφεί με κείμενο ή με σχέδια κ.λπ. Δημιουργείται και διατηρείται στο μυαλό των ανθρώπων του οργανισμού και συχνά είναι αποτέλεσμα της εμπειρίας τους που προέρχεται από μια διαδικασία δοκιμής και λάθους όταν εργάζονται για να επιλύσουν προβλήματα που ανακύπτουν στην πράξη. Η άρρητη γνώση πρέπει να καταγράφεται (ως καλές πρακτικές και κανόνες) και με τον τρόπο αυτό να ενσωματώνεται στις επιχειρηματικές διαδικασίες.

Ρητή γνώση είναι η γνώση που έχει καταγραφεί με κάποιο τρόπο όπως κείμενο, ήχος, εικόνα κ.λπ. Βρίσκεται σε βάσεις εγγράφων, σε αποθήκες δεδομένων κ.λπ.

Στην πράξη η γνώση είναι κάτι που μπορεί να μεταδίδεται/μεταφέρεται εύκολα από κάποια άτομα ενώ είναι πολύ δύσκολο να εξωτερικευθεί από κάποια άλλα άτομα. Πολλά εξειδικευμένα άτομα συχνά δυσκολεύονται να εκφράσουν την τεχνογνωσία τους ενώ κάποιοι νέοι υπάλληλοι είναι ικανοί να ορίσουν πιο εύκολα αυτό που

προσπαθούν να κάνουν επειδή ακολουθούν συνήθως ένα εγχειρίδιο ή ένα τρόπο επεξεργασίας.

Η άρρητη γνώση συχνά είναι πολύτιμη και μπορεί να οδηγήσει σε επιτυχή δραστηριότητα όταν καταγραφεί. Στη διαχείριση γνώσης η γνώση αντιμετωπίζεται στις δύο μορφές της: άρρητη γνώση και ρητή γνώση.

2.2 Ορισμός Διαχείρισης Γνώσης

“Knowledge management was initially defined as the process of applying a systematic approach to the capture, structuring, management, and dissemination of knowledge throughout an organization to work faster, reuse best practices, and reduce costly rework from project to project”

«Η διαχείριση γνώσης ορίστηκε αρχικά ως η διαδικασία εφαρμογής μιας συστηματικής προσέγγισης για τη σύλληψη, τη δόμηση, τη διαχείριση και τη διάδοση των γνώσεων σε έναν οργανισμό με στόχο να εργαστεί γρηγορότερα, να επαναχρησιμοποιήσει τις βέλτιστες πρακτικές και να μειώσει τις δαπανηρές επανεπεξεργασίες από έργο σε έργο» (Dalkir, 2011, p. 3)

Ο ορισμός αυτός βασίζεται σε πολλές ερευνητικές εργασίες (Nonaka 2007), (Pasternack and Viscio 1998), (Pfeffer and Sutton, 1999), (Ruggles and Holtshouse, 1999).

Η διαχείριση γνώσης χαρακτηρίζεται συχνά από μια λανθασμένη προσέγγιση σύμφωνα με την Dalkir (2011). Είναι συνηθισμένο να αντιμετωπίζεται, ως εξής: «αποθηκεύστε το, μπορεί να αποδειχθεί χρήσιμο κάποια στιγμή μελλοντικά».

Ως αποτέλεσμα αυτής της προσέγγισης, πολλά ή όλα τα έγγραφα είναι αποθηκευμένα, και υποτίθεται ότι οι εξελιγμένες μηχανές αναζήτησης μπορούν να ανακτήσουν μεγάλο μέρος αυτού του περιεχομένου. Η άποψη αυτή έχει πολλούς περιορισμούς. Αντίθετα οι λύσεις οι βασιζόμενες στη διαχείριση γνώσης έχουν αποδειχθεί πιο επιτυχημένες στη σύλληψη, αποθήκευση και διάδοση της γνώσης υπό τη μορφή ρητής γνώσης, π.χ., διδάγματα (learning stories) και βέλτιστες πρακτικές (best practices) (Μαρινάγη και Σκουρλάς, 2021).

Ακολουθεί ένας άλλος καλός ορισμός της διαχείρισης γνώσης:

“Knowledge management is the deliberate and systematic coordination of an organization’s people, technology, processes, and organizational structure in order to add value through reuse and innovation. This is achieved through the promotion of creating, sharing, and applying knowledge as well as through the feeding of valuable lessons learned and best practices into corporate memory in order to foster continued organizational learning” (Dalkir, 2011, p.4).

«Η διαχείριση της γνώσης είναι ο σκόπιμος και συστηματικός συντονισμός των ατόμων, της τεχνολογίας, των διαδικασιών και της οργανωτικής δομής ενός οργανισμού προκειμένου να προσθέσουν αξία μέσω της επαναχρησιμοποίησης και της καινοτομίας. Αυτό επιτυγχάνεται μέσω της προώθησης της δημιουργικότητας,

της κοινοποίησης και της εφαρμογής γνώση καθώς και μέσω της τροφοδοσίας της εταιρικής μνήμης με στοιχεία πολύτιμων μαθημάτων και βέλτιστων πρακτικών για την προώθηση της συνεχιζόμενης οργανωσιακής μάθησης».

Σύμφωνα με τον Nonaka (2007) «Διαχείριση Γνώσης είναι η δημιουργία, ανάπτυξη, συλλογή και διάχυση της γνώσης καθώς και η μετατροπή της ατομικής σε συλλογική γνώση».

Σημαντική είναι η έννοια της διαχείρισης του πνευματικού κεφαλαίου (intellectual assets) του οργανισμού. Η έννοια επικεντρώνεται σε γνώση που έχει επιχειρηματική αξία για τον οργανισμό

Ο Stewart (1997) ορίζει το πνευματικό κεφάλαιο ως οργανωμένη γνώση που μπορεί να χρησιμοποιηθεί για την παραγωγή πλούτου.

Το πνευματικό κεφάλαιο αποτελείται από τεχνογνωσία, εμπειρία και εμπειρογνωμοσύνη που υπάρχει στο μυαλό των υπαλλήλων και βέβαια από καταγεγραμμένη γνώση, διπλώματα ευρεσιτεχνίας, πνευματική ιδιοκτησία) κ.λπ. (Klein 1998), (Stewart 1997).

Η διαχείριση γνώσης σύμφωνα με την Dalkir βασίζεται σε μεγάλο αριθμό διαφορετικών τομέων όπως (βλέπε και Πίνακα).

- Organizational science
- Cognitive science
- Linguistics and computational linguistics
- Information technologies such as knowledge-based systems, document and information management, electronic performance support systems, and database technologies
- Information and library science
- Technical writing and journalism
- Anthropology and sociology
- Education and training
- Storytelling and communication studies
- Collaborative technologies such as Computer-Supported Collaborative Work (CSCW) and groupware as well as intranets, extranets, portals, and other web technologies

Δηλαδή, η Διαχείριση Γνώσης είναι ένα διεπιστημονικό γνωστικό πεδίο με στοιχεία από διάφορους τομείς, όπως:

• Η οργανωσιακή επιστήμη

• Η γνωσιακή επιστήμη

• Η γλωσσολογία και η υπολογιστική γλωσσολογία

• Οι τεχνολογίες πληροφοριών, όπως συστήματα βασισμένα στη γνώση, τα

συστήματα διαχείρισης εγγράφων και πληροφοριών, τα ηλεκτρονικά συστήματα υποστήριξης απόδοσης και οι τεχνολογίες βάσεων δεδομένων

• Η επιστήμη πληροφόρησης και βιβλιοθηκονομίας

• Η τεχνική γραφή και η δημοσιογραφία

• Η ανθρωπολογία και η κοινωνιολογία

• Η εκπαίδευση και κατάρτιση

• Η μελέτη αφηγήσεων και η επικοινωνία

• Οι συνεργατικές τεχνολογίες όπως η Συνεργατική Υποστήριξη Υπολογιστών (CSCW), groupware, intranets, extranets, portal και οι άλλες τεχνολογίες ιστού.

Σύμφωνα με τη Dalkir η διεπιστημονική φύση της διαχείρισης γνώσης έχει δύο πλευρές:

Η διαχείριση γνώσης είναι πλεονέκτημα επειδή όλοι μπορούν να βρουν χρήσιμα στοιχεία πάνω στα οποία θα βασίσουν μια πρακτική της διαχείρισης γνώσης.

Αλλά η ποικιλομορφία της διαχείρισης γνώσης οδηγεί τους σκεπτικιστές να υποστηρίζουν ότι η διαχείριση γνώσης δεν μπορεί να θεωρηθεί ως ξεχωριστός τομέας.

Τελικά, η διαχείριση γνώσης ασχολείται και με τη γνώση και με την πληροφορία. Η γνώση συχνά βασίζεται σε βιωματικές ή ατομικές αξίες, αντιλήψεις και εμπειρίες.

2.3 Σημασία της διαχείρισης γνώσης

Υπάρχουν τέσσερις επιχειρηματικοί παράγοντες πίσω από το αυξημένο ενδιαφέρον για την εφαρμογή της διαχείρισης γνώσης σύμφωνα με την Dalkir.

1. Η παγκοσμιοποίηση των επιχειρήσεων. Οι επιχειρηματικοί οργανισμοί σήμερα είναι παγκόσμιοι, πολύγλωσσοι και πολυπολιτισμικοί.

2. Leaner organizations. Οι οργανισμοί βασίζονται στην ιδέα ότι πρέπει να κάνουν περισσότερα και να τα κάνουν γρηγορότερα και αν είναι εφικτό με λιγότερους πόρους. Δηλαδή, πρέπει να εργάζονται πιο έξυπνα αξιοποιώντας τη γνώση.

3. Η εταιρική «αμνησία». Οι εργαζόμενοι εναλλάσσονται πιο γρήγορα στους οργανισμούς και η γνώση τους συχνά «χάνεται» για τον οργανισμό.

4. Η τεχνολογική πρόοδος. Οι εξελίξεις της πληροφορικής έκαναν τη συνδεσιμότητα πανταχού παρούσα και άλλαξαν ριζικά τις προσδοκίες.

Πρέπει επίσης να έχουμε στο μυαλό μας ότι το σημερινό εργασιακό περιβάλλον είναι πιο περίπλοκο λόγω και της αύξησης του αριθμού των ζητημάτων και των αντικειμένων που πρέπει να παρακολουθούνται καθημερινά.

Η διαχείριση της γνώσης παρέχει οφέλη σε μεμονωμένους υπαλλήλους, σε κοινότητες πρακτικής και στον ίδιο τον οργανισμό σύμφωνα με την Dalkir.

Κάποια από τα οφέλη για τους μεμονωμένους υπαλλήλους είναι ότι τους βοηθά να εξοικονομήσουν χρόνο όταν κάνουν την δουλειά τους, ενώ παράλληλα δημιουργείται μια αίσθηση ότι υπάρχει ένας δεσμός μεταξύ των υπαλλήλων του οργανισμού και παρέχονται ευκαιρίες στους υπαλλήλους για να συνεισφέρουν.

Στην κοινότητα πρακτικής η διαχείριση γνώσης βοηθά να αναπτυχθούν επαγγελματικές δεξιότητες, διευκολύνει την αποτελεσματική δικτύωση και συνεργασία και αναπτύσσει μία κοινή γλώσσα μεταξύ της κοινότητας.

Στους οργανισμούς η διαχείριση γνώσης βοηθά στην προώθηση της στρατηγικής, δείχνει τις βέλτιστες πρακτικές, διασταυρώνει τις ιδέες και αυξάνει τις ευκαιρίες για καινοτομία, επιτρέπει στους οργανισμούς να παραμείνουν μπροστά από τον ανταγωνισμό και βελτιώνει τις γνώσεις που ενσωματώνονται σε προϊόντα και υπηρεσίες.

3. Συστήματα Μητρώου Ειδικών

Σήμερα η νέα γνώση λαμβάνεται πολύ πιο γρήγορα από ότι μπορεί να καταγραφεί και να αποθηκευτεί. Σε οποιονδήποτε οργανισμό εμπλέκεται στη μεταφορά και ανταλλαγή γνώσεων, ένα βασικό σημείο που πρέπει να ληφθεί υπόψη είναι η ικανότητα εντοπισμού εμπειρογνωμοσύνης.

Η ατομική γνώση είναι σημαντική αλλά και το να γνωρίζεις ποιος μέσα στην εταιρεία γνωρίζει κάτι συγκεκριμένο είναι ακόμα πιο σημαντικό. Έτσι οι οργανισμοί θέλουν να έχουν εύκολη πρόσβαση στο να βρουν το σωστό άτομο που κατέχει αυτή την γνώση για να απευθυνθούν. Για αυτό τον λόγο πολλοί οργανισμοί δημιουργούν Συστήματα Μητρώου Ειδικών (Yellow Pages Systems) ώστε να είναι δυνατό για τους εργαζόμενους να εντοπίσουν άλλους εργαζόμενους με συγκεκριμένη εμπειρία και δεξιότητες.

Τα Συστήματα Μητρώου Ειδικών είναι ένα εσωτερικό εργαλείο που δημιουργήθηκε για να βοηθήσει τους ανθρώπους να βρουν υπαλλήλους που διαθέτουν εμπειρία ή και γνώση σε μια συγκεκριμένη περιοχή. Αναφέρονται επίσης ως «Κατάλογοι εμπειρογνωμόνων» ή «Κατάλογοι εξειδίκευσης». Αποκαλύπτουν την εμπειρία που υπάρχει σε έναν οργανισμό.

Αυτοί οι “κατάλογοι” δημιουργούνται για να βοηθήσουν στην απάντηση της ερώτησης “ποιος ξέρει τι” ή σε κάποιες περιπτώσεις “ποιος ξέρει ποιος” (Dalkir, 2011).

Αυτά τα αρχεία μπορούν να κατασκευαστούν σαν εταιρικός κατάλογος, να διατίθενται στο διαδίκτυο, και να περιλαμβάνουν όλους τους υπαλλήλους. Σε πιο φιλόδοξη προσέγγιση, δεν περιορίζονται σε στοιχεία επικοινωνίας, τίτλο εργασίας και θέσεις στον οργανισμό, αλλά περιλαμβάνουν στοιχεία εξειδίκευσης, τρόπο επικοινωνίας κ.λπ.

3.1 Στοιχεία συστημάτων μητρώου ειδικών

Η ανάπτυξη συστημάτων μητρώου ειδικών δεν είναι πάντα εύκολη και πολλές φορές τα συστήματα καταλήγουν να μην είναι χρήσιμα ή να μην ενημερώνονται ή και να είναι απαρχαιωμένα. Υπάρχουν κάποια σημαντικά σημεία στη δημιουργία τέτοιων συστημάτων έτσι ώστε να αξιοποιούνται και να εκπληρώνουν το σκοπό τους (Dalkir),(Collison 2005):

- Να ευθυγραμμίζονται με ένα σαφές σχετικό όραμα για συνεργασία, διάχυση της γνώσης κ.λπ.
- Οι εμπλεκόμενοι να δημιουργούν και να τροποποιούν τις καταχωρήσεις έτσι ώστε να είναι ενημερωμένο.
- Να εξισορροπείται το «ανεπίσημο» και το επίσημο περιεχόμενο. Για παράδειγμα, να δίνεται η δυνατότητα στους ανθρώπους να μοιράζονται και

τα ενδιαφέροντά τους πέρα από τις γνώσεις τους.

- Να αναρτώνται και φωτογραφίες όταν είναι δυνατόν
- Το σύστημα να είναι ευέλικτο και χωρίς αποκλεισμούς για όλους. Να είναι εύκολο στη χρήση για τους εμπλεκόμενους και να επιτρέπει όχι μόνο να εισάγουν τις πληροφορίες τους αλλά και να εξάγουν πληροφορίες από αυτό.
- Να διατηρούνται ενημερωμένα τα στοιχεία των ατόμων
- Να ενθαρρύνεται η χρήση της με διάφορους τρόπους ώστε να βλέπουν οι συμμετέχοντες την αξία του συστήματος.

Κοινά πεδία/στοιχεία στα Συστήματα Μητρώου Ειδικών είναι:

όνομα, θέση εργασίας, τμήμα ή ομάδα εργασίας, περιγραφή της θέσης, σχετικά επαγγελματικά προσόντα, ένα ενημερωμένο βιογραφικό, τομείς γνώσης και εξειδίκευσης, ενδιαφέροντα, συμμετοχή σε άλλα δίκτυα γνώσης, προσωπικές πληροφορίες, φωτογραφία και πληροφορίες επικοινωνίας.

Ο Lamont (2003) υπογραμμίζει τη συμβολή των Συστημάτων Μητρώου Ειδικών στις οργανωτικές πρωτοβουλίες μάθησης, στη διευκόλυνση προγραμμάτων καθοδήγησης, στον προσδιορισμό αναγκών για γνώση και στην παροχή, υποστήριξη, αξιολόγηση εκπαιδευτικών δραστηριοτήτων.

Παράδειγμα

Στη συνέχεια θα εξετάσουμε μια τέτοια εφαρμογή για μια μεγάλη εταιρεία πώλησης, συντήρησης και επισκευής φωτογραφικών μηχανών. Ακολουθεί παράδειγμα Yellow Pages.

Τμήματα	Βιβλιοθήκες Γνώσης	Περιοχές Συζητήσεων	Υποστήριξη
Πωλήσεις/προϊόντα (φωτογραφικές κάμερες)	Best practices	Φόρουμ συζητήσεων	Γλωσσάρι όρων
Έργα (projects)	Lessons Learned	Διαχείριση έργου	Συχνές ερωτήσεις
Προμηθευτές	Stories	Risk management	
Κοινότητα Πρακτικής, π.χ., Δίκτυο εξειδικευμένων τεχνικών	Σεμινάρια		

Δίκτυο ειδικών

Θέσεις	Γεωγραφική περιοχή	Επιχειρηματικές δραστηριότητες	**Ειδικότητα**
Αναπληρωτής Πρόεδρος	Βόρεια Ευρώπη	Πωλήσεις	Content Management
Γενικός Διευθυντής	Κεντρική Ευρώπη	Χρηματοοικονομικά	Διαχείριση Γνώσης
Υπεύθυνος Τμημάτων Διανομής	Νότια Ευρώπη	Διανομή	Αναβάθμιση λογισμικού
Υπεύθυνος Τεχνικών Επισκευής	Νότια Ευρώπη	Συντήρηση & Επισκευές	Επιδιόρθωση υλικού

Ειδικότητα		
Αναβάθμιση λογισμικού		
Ντόγια Σύλβα	Κεντρικά Γραφεία	6999999999
Μαρίνα Δεληγιάννη	Κεντρικά Γραφεία	6999999998
Επιδιόρθωση Υλικού		
Δήμητρα Ιωάννου	Ηρ. Πολυτεχνείου 7	6999999997
Γιώργος Σουρπής	Ηρ. Πολυτεχνείου 7	6999999996

Κεφάλαιο 4. Υλοποίηση

Η υλοποίηση της παρούσας διπλωματικής πραγματοποιήθηκε υπό τις παρακάτω προϋποθέσεις.

Το λειτουργικό σύστημα στο οποίο αναπτύχθηκε ο κώδικας αποτέλεσε το Ubuntu 20.04. Η έκδοση της monetdb v11.39.17 και τέλος η έκδοση της python ήταν η 3.8. Οι συγκεκριμένες εκδόσεις των παραπάνω εργαλείων αποτελούν τον απαραίτητο συνδυασμό για την ομαλή ανάπτυξη και εκτέλεση της εργασίας.

4.1 Δημιουργία σχήματος βάσης

Εφαλτήριο βήμα πριν από οποιαδήποτε άλλη ενέργεια αποτέλεσε η δημιουργία του σχήματος της βάσης το οποίο πραγματοποιήθηκε εκτελώντας τις παρακάτω εντολές μέσω του mclient.

```
shell> monetdbd create /path/to/mydbfarm
shell> monetdbd start /path/to/mydbfarm
shell> monetdb create project
shell> monetdb release project
shell> mclient -u monetdb -d project
password:<monetdb>
```

Για τερματισμό της σύνδεση με τον διακομιστή

```
sql>\q
```

4.2 Εργαλεία που χρησιμοποιήθηκαν

Για τη διαχείριση των δεδομένων, για την αποθήκευση, επεξεργασία και ανάκτηση των δεδομένων χρησιμοποιήθηκε η monetDB. Η monetDB χρησιμοποιείται στη διαχείριση βάσης δεδομένων ανοιχτού κώδικα με προσανατολισμό στις στήλες. Αναπτύχθηκε στο Centrum Wiskunde & Informatica (CWI) (Κέντρο Ερευνών στα πεδία των μαθηματικών και της θεωρητικής επιστήμης των υπολογιστών) στην Ολλανδία όπως επίσης και η γλώσσα προγραμματισμού υψηλού επιπέδου γενικού σκοπού Python. Είναι αποτελεσματική σε πολύπλοκα queries όπως για παράδειγμα στο συνδυασμό πινάκων με εκατοντάδες στήλες και εκατομμύρια σειρές. Έχει χρησιμοποιηθεί σε διαδικτυακή αναλυτική επεξεργασία, εξόρυξη δεδομένων, συστήματα γεωγραφικών πληροφοριών, πλαίσιο περιγραφής πόρων, ανάκτηση κειμένου κ.λπ. (βικιπαίδεια)

Σαν γλώσσα προγραμματισμού χρησιμοποιήθηκε η Python. Η Python είναι διερμηνευμένη, γενικού σκοπού και υψηλού επιπέδου, γλώσσα προγραμματισμού. Ανήκει στις γλώσσες προστακτικού προγραμματισμού και υποστηρίζει τόσο το διαδικαστικό όσο και το αντικειμενοστραφές προγραμματιστικό υπόδειγμα. Είναι δυναμική γλώσσα προγραμματισμού. Δημιουργήθηκε από τον Ολλανδό Guido van Rossum στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) το 1989 και κυκλοφόρησε για πρώτη φορά το 1991.(βικιπαίδεια)

Τέλος, ως εναλλακτικό εργαλείο αντί του mclient, χρησιμοποιήθηκε το dbeaver, το οποίο βοήθησε στο πιο φιλικό UI για το χρήστη για τη σύνδεση στη βάση, καθώς και στην πιο ομαλή ανάπτυξη του κώδικα.

Κεφάλαιο 5. Δεδομένα

Η μελέτη πραγματοποιήθηκε χρησιμοποιώντας τρεις συλλογές δεδομένων που είναι οι ακόλουθες:

- `cv.gz`
- `camera_specs.tar.gz`
- `sigmod_medium_labelled_dataset.csv`

Η συλλογή δεδομένων του φακέλου cv δημιουργήθηκε για τις ανάγκες της συγκεκριμένης μελέτης και οι άλλες δύο συλλογές δεδομένων υπάρχουν διαθέσιμες στο διαδίκτυο.

Ο πρώτος φάκελος περιέχει βιογραφικά (cv) ειδικών, τα οποία είναι αποθηκευμένα σε .json αρχεία. Τα .json αρχεία περιέχουν πληροφορίες σχετικά με τα βιογραφικά ειδικών τεχνικών στον τομέα της επισκευής φωτογραφικών μηχανών. Αναλυτικότερα, η δομή των .json αρχείων φαίνεται παρακάτω. Τα αρχεία περιλαμβάνουν τα στοιχεία των φυσικών προσώπων των τεχνικών όπως τις προσωπικές πληροφορίες, την ειδικότητά τους και την εργασιακή τους εμπειρία.

Παρατίθενται δύο δείγματα τέτοιων αρχείων:

```
{
      "fullname": "Sotiris Psaras",
      "mail": "sotirisp@gmail.com",
      "birthday": "08/05/1970",
      "location": "Trikala, Greece",
      "Specialized": "casio, polaroid, panasonic",
      "Experience": [
      {
      "company": "Easyrepair Athens",
      "loc" :  "Athens, Greece",
      "date": "1993 to 1998",
      "description" : "repairing of cameras and photo lens of
sony.General Check of the smooth operation of the camera"
      },
      {
      "company": "fotoklik ",
      "loc" :  "Athens, Greece",
      "date": "1999 to 2003",
      "description" : "repairing of cameras and General Check
of the smooth operation of the camera"
      },
      {
      "company" : "Andreou cameras experts",
      "loc" : "Trikala, Greece",
      "date" : "2003 to 2021",
      "description" : "repairing of cameras and photo lens
and firmware"
      }
      ]
}
```

```
{
      "fullname": "Napoleon Arvanitis",
      "mail": "napar@gmail.com",
      "birthday": "02/05/1976",
      "location": "Agrinio, Greece",
      "Specialized": "Olympus, canon, samsung, philips,
Sony",
      "Experience": [
      {
      "company": "CameraFIX",
      "loc" :  "Agrinio Greece",
      "date": "2001 to 2021",
      "description" : "repairing of cameras and photo
lens.General Check of the smooth operation of the camera"
      }
      ]
}
```


Η συλλογή δεδομένων camera_specs.tar.gz περιέχει φακέλους κάθε ένας από τους οποίους έχει ένα μεγάλο αριθμό json αρχείων. Καθε json αρχείο περιέχει πληροφορίες και χαρακτηριστικά μιας κάμερας που είναι αναρτημένη στην συγκεκριμένη ιστοσελίδα που περιγράφεται από το όνομα του φακέλου. Κάθε json αρχείο έχει σαν όνομα έναν μοναδικό αριθμό. Υπάρχουν αρχεία με πολύ περισσότερες πληροφορίες ή και λιγότερες.


```
{
    "<page title>": "Nikon Coolpix S210 8 0 MP Digital Camera
Plum Extra Charger Battery Tripod | eBay",
    "brand": "Nikon",
    "condition": "Used: An item that has been used
previously. The item may have some signs of cosmetic wear,
but is fully\noperational and functions as intended. This
item may be a floor model or store return that has been used.
See the seller\u00e2\u0080\u0099s listing for full details
and description of any imperfections.\nSee all condition
definitions- opens in a new window or tab\n... Read moreabout
the condition",
    "megapixels": "8.0 MP",
    "model": "S210",
    "mpn": "26103",
    "optical zoom": "3x",
    "screen size": "2.5\"",
    "type": "Point & Shoot",
    "upc": "718122199020"
}
```

```
{
    "<page title>": "Canon EOS 7D Mark II Black SLR Digital
Camera Body Only (20.2 MP, CF/SD Card Slot) Price Comparison
at Buy.net",
    "analog video out": "Yes",
    "autofocus points": "65",
    "battery builtin": "No",
    "battery include": "Yes",
    "battery model supported": "LP-E6N / LP-E6",
    "battery rechargeable": "Yes",
    "brand name": "Canon",
    "builtin flash": "Yes",
    "exposure control": "Program AE Aperture Priority Shutter
Priority Manual",
    "gps": "Yes",
    "hd movie mode": "Yes",
    "iso sensitivity": "ISO 51200",
    "language support": "Japanese",
    "longest shutter speed": "30 Second",
    "maximum diopter adjustment": "1",
    "maximum frame rate": "60 fps",
    "minimum diopter adjustment": "-3",
    "number of batteries support": "1",
    "parent product type": "Digital Camera",
    "pictbridge": "Yes",
    "product line": "EOS",
    "product model": "7D Mark II",
    "self timer": "10 Second",
    "shortest shutter speed": "1/8000 Second",
    "viewfinder type": "SLR",
    "white balance modes": "Auto (AWB) Daylight Shade Cloudy,
Twilight, Sunset Tungsten Light White Fluorescent Light Flash
Custom (Custom WB) Color temperature"
}
```


Η τελευταία συλλογή δεδομένων είναι ένα .csv αρχείο το οποίο περιέχει συγκρίσεις διάφορων καμερών. Η πληροφορία που περιέχεται στο αναφερόμενο αρχείο, είναι η σύγκριση δύο φωτογραφικών μηχανών που έχουν ανακτηθεί από το σύνδεσμο που περιγράφεται στο left_spec_id και right_spec_id αντίστοιχα. Συνοδεύεται από το id της εκάστοτε κάμερας που έχει ανακτηθεί και τέλος το αποτέλεσμα της σύγκρισης αυτής περιγράφεται από το label, όπου στην περίπτωση όπου έχουμε ίδιες κάμερες με ίδια χαρακτηριστικά έχει την τιμή 1, αλλιώς 0.

Παρατίθεται δείγμα από τη μορφή του περιγραφόμενου αρχείου:

left_spec_id,right_spec_id,label
www.garricks.com.au//31,www.ebay.com//53278,1
www.ebay.com//58782,www.ebay.com//24817,0
www.ebay.com//58782,www.ebay.com//43019,0
www.ebay.com//42055,www.ebay.com//54403,0
www.ebay.com//44280,buy.net//6145,1
www.ebay.com//42074,www.ebay.com//47107,0

Όπως φαίνεται αρχικά με το κόμμα (,) χωρίζεται η πρώτη παράμετρος από την δεύτερη και η δεύτερη από την τρίτη. Βλέπουμε ότι η πρώτη παράμετρος (left_spec_id) αναφέρεται σε ποια σελίδα είναι αναρτημένη η πρώτη κάμερα (και κατ' επέκταση σε ποιο directory του αρχείου camera_specs.tar.gz θα βρεθεί π.χ. www.ebay.com, μετά τα "//" φαίνεται σε ποιο αρχείο json υπάρχουν οι πληροφορίες για αυτή (π.χ. στο 58782). Μετά βλέπουμε σε ποια ιστοσελίδα είναι αναρτημένη η δεύτερη (right_spec_id) και το αρχείο που υπάρχουν οι πληροφορίες για αυτή και τέλος βλέπουμε το label που είναι 0 ή 1 (ή flag όπως θα μπορούσε να χαρακτηριστεί) το οποίο είναι το αποτέλεσμα της σύγκρισης όπως αναφέρθηκε παραπάνω.

Κεφάλαιο 6. Επεξεργασία δεδομένων

Στο κεφάλαιο περιγράφεται η επεξεργασία των δεδομένων και τεκμηριώνονται κάποια από τα προγράμματα.

6.1 Αποθήκευση δεδομένων στην βάση δεδομένων

Ξεκινώντας με το dataset sigmod_medium_labelled_dataset.csv χρησιμοποιείται ένα απλό sql command copy into για αποθήκευση των δεδομένων στον πίνακα της βάσης δεδομένων.

Δημιουργήθηκε ένας πίνακα με τρεις στήλες, η καθεμία από τις οποίες αντιπροσωπεύει την κάθε παράμετρο. Στον πίνακα αυτό αποθηκεύεται το περιεχόμενο του αρχείου. Χρησιμοποιείται το κόμμα(,) για αλλαγή στήλης στον πίνακα και το '\n' όταν ξεκινάει μία νέα εγγραφή. Στο αρχείο πραγματοποιήθηκε η αφαίρεση της πρώτης γραμμής, μιας και δεν περιείχε πληροφορία, παρά μόνο το όνομα των παραμέτρων.

```
CREATE TABLE project.tbl_data(left_spec_id
varchar(52),right_spec_id varchar(52),label VARCHAR(5));

COPY INTO project.tbl_data FROM
'.../sigmod_medium_labelled_dataset.csv' USING DELIMITERS
',', E'\n', '"';

DELETE FROM project.tbl_data  where left_spec_id  =
'left_spec_id';
```

Json LOADERs

Σκοπός του Loader είναι να αντιγράψει ό,τι βρίσκεται στα Json αρχεία που είναι στους διάφορους καταλόγους του φακέλου στους πίνακες της βάσης. Διαβάζει από το φάκελο που παίρνει σαν όρισμα όλους τους φακέλους που αυτός περιέχει και όλα τα αρχεία μορφής .json που βρίσκονται σε αυτούς. Χρησιμοποιείται ένας Json Loader για κάθε πίνακα που δημιουργείται.

Αρχικά κατασκευάζεται ο json_cvloader που ασχολείται με τα στοιχεία των ειδικών τεχνικών επισκευής φωτογραφικών μηχανών. Ο json_cvloader, ο οποίος είναι υπεύθυνος για τη φόρτωση των προσωπικών στοιχείων από τα αρχεία των βιογραφικών, παίρνει σαν όρισμα το μονοπάτι όπου βρίσκεται ο φάκελος cv. Η έξοδος του συγκεκριμένου loader είναι η δημιουργία ενός πίνακα: του cv_tbl,ο οποίος συντηρεί τα στοιχεία των ειδικών.

Όπως φαίνεται στα json αρχεία, τα δεδομένα που περιέχουν είναι: το ονοματεπώνυμο του ειδικού, η ημερομηνία γέννησης, ο τόπος διαμονής και η

διεύθυνση ηλεκτρονικού ταχυδρομείου. Περνάει τα δεδομένα αυτά στις αντίστοιχες μεταβλητές και φορτώνονται στις αντίστοιχες στήλες του πίνακα cv_tbl.

```
CREATE OR REPLACE LOADER json_cvloader(directory STRING)
LANGUAGE PYTHON {
      import json
      import os
      fullname = ''
      email = ''
      birthday = ''
      location = ''
      for root,dirs,files in os.walk(directory):
            for file in files:
                  if file.endswith('.json'):
                  with open(os.path.join(root,file),'r') as f:
                  fullname = ''
                  email = ''
                  birthday= ''
                  location = ''
                  pairs = json.load(f)
                  for key , value in pairs.items():
                        if key =='fullname':
                              fullname= pairs[key]
                        elif key == 'birthday':
                              birthday= pairs[key]
                        elif key == 'location':
                              location= pairs[key]
                          elif key == 'mail':
                              email= pairs[key]
                  id =file
                  id = id.replace('.json','')
                        _emit.emit({'id' : id, 'fullname' :
fullname,'email' : email, 'birthday': birthday,'location' :
location,})
      f.close()
};
```

Έπειτα κατασκευάζεται ο json_specialityloader που χρησιμοποιείται για την δημιουργία του πίνακα speciality ο οποίος συντηρεί πληροφορίες σχετικά με την ειδικότητά των ειδικών επισκευής φωτογραφικών μηχανών. Κάθε ειδικός μπορεί να ειδικεύεται στην επισκευή διαφόρων μαρκών φωτογραφικών μηχανών. Ο json_specialityloader παίρνει σαν όρισμα το μονοπάτι που βρίσκεται ο φάκελος cv και η έξοδος του είναι η δημιουργία του πίνακα speciality, ο οποίος αποδίδει ένα μοναδικό id, το οποίο είναι το όνομα του αρχείου, και οι μάρκες στις οποίες ειδικεύεται ο κάθε τεχνικός.

```
CREATE OR REPLACE LOADER json_specialityloader(directory
STRING) LANGUAGE PYTHON {
      import json
      import os
      speciality = ''
      for root,dirs,files in os.walk(directory):
      for file in files:
            if file.endswith('.json'):
                  with open(os.path.join(root,file),'r') as f:
                  pairs = json.load(f)
                  for key , value in pairs.items():
                        if key=='Specialized':
                              specialities= pairs[key]
                              words = specialities.split(",")
                               id =file
                                    id = id.replace('.json','')
                                    speciality= ''
                                    for w in words:
                                          speciality=w
                                    _emit.emit({'id' : id,
'speciality' : speciality})
                                    speciality=''
      f.close()
};
```

Τέλος, ο json_exploader χρησιμοποιείται στη δημιουργία του πίνακα experience_tbl. Εκεί συντηρούνται οι πληροφορίες για την εργασιακή εμπειρία που έχουν οι ειδικοί στον κλάδο, σε ποιες εταιρείες έχουν δουλέψει, ποιο διάστημα απασχολήθηκαν από την κάθε εταιρεία, που εδρεύει η κάθε μία και μία σύντομη περιγραφή των καθηκόντων/ θέσης τους. Κάθε ειδικός μπορεί να έχει δουλέψει σε πολλές, μία ή και καμία εταιρεία.

Ο json_exploader παίρνει σαν όρισμα το μονοπάτι που βρίσκεται ο φάκελος cv και η έξοδος του είναι η δημιουργία του πίνακα experience_tbl, ο οποίος αποδίδει ένα μοναδικό id για τον κάθε ειδικό, το οποίο είναι το όνομα του αρχείου, και επιπλέον κάθε εταιρεία που δούλεψε, την τοποθεσία της, το διάστημα και την περιγραφή τη θέσης.

```
CREATE OR REPLACE LOADER json_exploader(directory STRING)
LANGUAGE PYTHON {
      import json
      import os
      for root,dirs,files in os.walk(directory):
      for file in files:
            if file.endswith('.json'):
                  with open(os.path.join(root,file),'r') as f:
                  pairs = json.load(f)
                  for key , value in pairs.items():
                        if key=='Experience':
                              for i in pairs['Experience']:
                                    x= i["company"]
                                    y= i["loc"]
                                    t= i["date"]
                                    w= i["description"]
                                    id =file
                  id = id.replace('.json','')
                                    _emit.emit({'id' : id,
'company' : x, 'companylocation' : y,'time_period' : t,
'description': w})
      f.close()
};
```

Για να εξεταστούν οι δυνατότητες που παρέχει η MonetDB στην διαχείριση μεγάλου όγκου δεδομένων σε πολύ μικρό χρόνο δημιουργείται ο json_loader που χρησιμοποιείται για την αποθήκευση της συλλογής δεδομένων με τα χαρακτηριστικά των φωτογραφικών μηχανών. Για την αποθήκευση των συγκεκριμένων πληροφοριών δημιουργείται ένας πίνακας. Εκχωρούνται τα δεδομένα που έχει κάθε αρχείο σε τρεις μεταβλητές, το page_title, το info και το specId. Το specId δημιουργείται παίρνοντας το όνομα του directory που βρισκόταν το .json αρχείο,, προσθέτοντας "//" και το id του json αρχείου και βγάζοντας την κατάληξη .json (αυτό δίνει και ένα μοναδικό κλειδί που χρησιμοποιείται αργότερα στον πινακά) . Με το _emit.emit({'specId' : specId,'page_title' : page_title, 'info' : info}) γίνεται εκχώρηση κάθε σειράς στον πίνακα με τις στήλες: specId, page_title και info.

```
CREATE OR REPLACE LOADER json_loader(directory STRING)
LANGUAGE PYTHON {

      import json
      import os

      page_title = ''
      info = ' '

      for root,dirs,files in os.walk(directory):
            for file in files:
                  if file.endswith('.json'):

                        with open(os.path.join(root,file),'r')
as f:

                              page_title = ''
                              info = ' '
                              pairs = json.load(f)
                              for key , value in pairs.items():
                                    if key =='<page title>':
                                          page_title = pairs[key]
                                    else:
                                          info = info +
str(pairs[key])

                  path_file = root.replace(directory+'/','')
                  specId = path_file+'//'+file
                  specId = specId.replace('.json','')
                  _emit.emit({'specId' : specId,'page_title' :
page_title, 'info' : info})
      f.close()
};
```

6.2 Σύνδεση πινάκων

Για την συλλογή δεδομένων με τα βιογραφικά δημιουργούνται οι πίνακες cv_tbl, speciality και experience_tbl. Γίνεται φόρτωση των δεδομένων από τα αρχεία και ορισμός των κλειδιών. Ο κύριος πίνακας είναι αυτός με τις πληροφορίες των ειδικών και εκεί ορίζεται το primary key που είναι το id. Στον πίνακα speciality γίνεται ορισμός του foreign key που είναι το id και αναφέρεται στον πίνακα cv_tvl. Όπως αναφέρθηκε κάθε ειδικός μπορεί να έχει διάφορες ειδικότητες. Έτσι η σχέση που προκύπτει είναι 1:N. Τέλος ορίζεται και στον πίνακα experience_tbl foreign

key, το οποίο είναι το id και αναφέρεται στον πίνακα cv_tbl. Και εδώ κάθε ειδικός μπορεί να έχει δουλέψει σε πολλές εταιρείες άρα η σχέση είναι 1:N.

```
ALTER TABLE project.cv_tbl ADD PRIMARY KEY ("id");

ALTER TABLE project.speciality
ADD FOREIGN KEY ("id") REFERENCES project.cv_tbl("id");

ALTER TABLE project.experience_tbl
ADD FOREIGN KEY ("id") REFERENCES project.cv_tbl("id");
```

Ένα παράδειγμα μιας εγγραφής είναι το παρακάτω. Έχουμε τη Γεωργία Χριστοπούλου που μένει στην Κόρινθο, έχει δουλέψει για τις εταιρείες The Repair Specialists και Intersys και ειδικεύεται στην επισκευή Olympus και Canon.

id	**fullname**	**Email**	**birthday**	**location**
23	Georgia Xristopoulou	xristg@gmail.com	28/12/1974	Korinthos, Greece

id	**company**	**companylocation**	**time_period**	**description**
23	The repair specialists	Korinthos Greece	2011 to now	Working with...
23	Intersys	Athens Greece	1999 to 2010	Repaired ...

id	**speciality**
23	Olympus
23	Canon

Όπως φαίνεται πολλαπλές εγγραφές του πίνακα experience μπορούν να αντιστοιχούν σε έναν ειδικό στον πίνακα cv_tbl. Αντίστοιχα πολλές εγγραφές του πίνακα speciality μπορούν να αντιστοιχούν στον πίνακας cv_tbl.

Εξετάζοντας τις συλλογές δεδομένων,

`camera_specs.tar.gz` και `sigmod_medium_labelled_dataset.csv`

που χρησιμοποιούνται στην μελέτη της διαχείρισης δεδομένων μεγάλου όγκου, φαίνεται ότι οι δύο πίνακες συνδέονται. Έτσι σχεδιάζεται η σύνδεση και στη βάση δεδομένων ώστε να είναι εμφανή και τα αποτελέσματα που προκύπτουν.

Αρχικά ορίζεται το πρωτεύον κλειδί στον πίνακα με τις κάμερες (tbl_camera_specs) που είναι το specId καθώς αυτό περιγράφει μοναδικά κάθε εγγραφή. Ύστερα γίνεται η σύνδεση του πίνακα tbl_data σε αυτόν με τις κάμερες με τη χρήση δύο foreign keys. Το ένα είναι το left_spec_id και το άλλο είναι το right_spec_id.

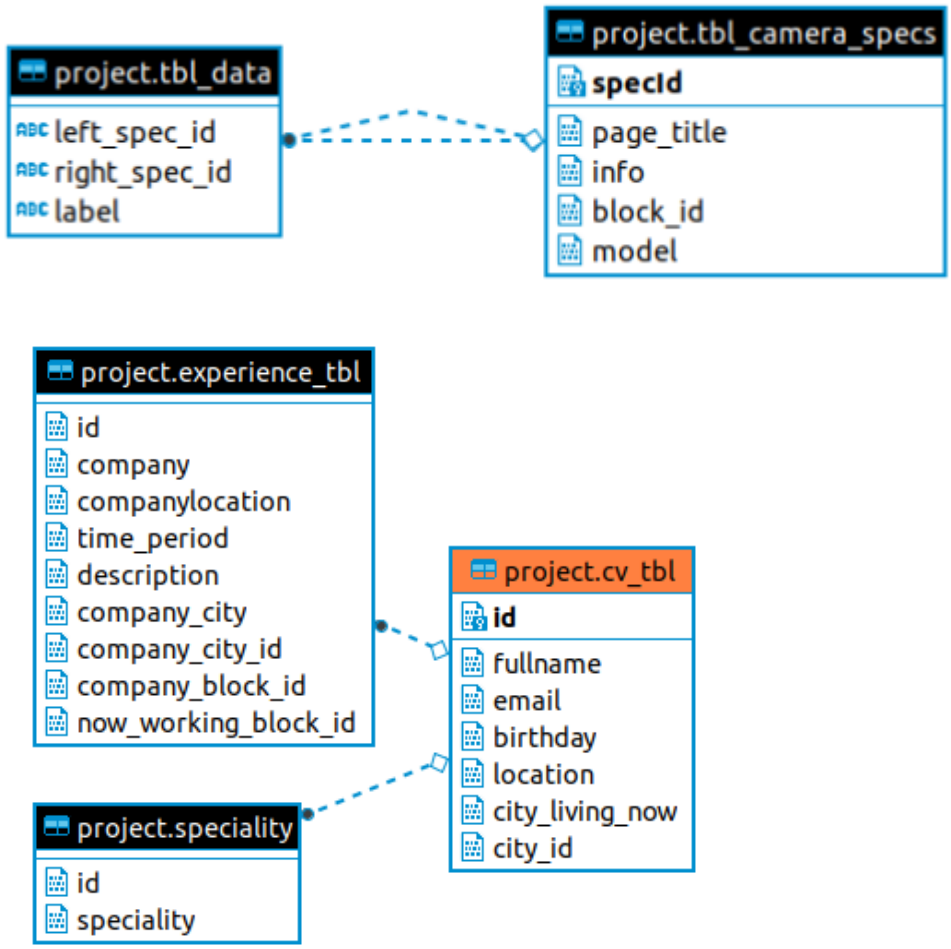
```
CREATE TABLE project.tbl_camera_specs FROM LOADER
json_loader('/home/2013_camera_specs');
ALTER TABLE tbl_camera_specs ADD PRIMARY KEY ("specId");

ALTER TABLE "tbl_data" ADD CONSTRAINT "project.tbl_data_fk1"
FOREIGN KEY ("left_spec_id") REFERENCES
"tbl_camera_specs"("specId");

ALTER TABLE "tbl_data" ADD CONSTRAINT "tbl_data_fk2"
FOREIGN KEY ("right_spec_id") REFERENCES
"tbl_camera_specs"("specId");
```

6.3 Σχήμα

Το relational schema της βάσης που δημιουργήθηκε είναι της μορφής :



6.4 Επεξεργασία των δεδομένων

Για την επεξεργασία των δεδομένων της βάσης δεδομένων δημιουργούνται κάποια UDFs(Pyspark UserDefindFunctions). UDFs είναι ένας εύκολος τρόπος να αλλαχθεί συνηθισμένος python κώδικας σε κάτι πιο κλιμακούμενο.

Δημιουργήθηκαν κάποια UDFs για την επεξεργασία στηλών στον πίνακα με τις φωτογραφικές κάμερες, κάποια για την επεξεργασία διάφορων χρήσιμων στηλών στους πίνακες cv_tbl, speciality και experience_tbl και κάποια που χρησιμοποιήθηκαν σε όλους τους πίνακες.

convert_lower

Η συνάρτηση convert_lower δεχεται σαν όρισμα μία συμβολοσειρά, μετατρέπει όσα γράμματα είναι κεφαλαία σε πεζά και μετά επιστρέφει την συμβολοσειρά. Χρησιμοποιείται για μορφοποίηση σε όλες τις στήλες των πινάκων που επιλέγονται για επεξεργασία.

```
CREATE OR REPLACE FUNCTION convert_lower(page_title STRING)
RETURNS STRING
LANGUAGE PYTHON {
      return numpy.array([x.lower() for x in page_title],
dtype=numpy.object)
};
```

replace_punctuation

Η συνάρτηση replace_punctuation δεχεται σαν όρισμα, δηλαδή σαν είσοδο, μία συμβολοσειρά την page_title. Σε αυτή την συμβολοσειρά εισόδου αλλάζει όποιο σημείο στίξης βρει απο αυτά που βρίσκονται στο punctuation με το κενό. Επιστρέφει σαν έξοδο την τελική συμβολοσειρά που προέκυψε. Χρησιμοποιείται και αυτή όπως και η convert_lower για μορφοποίηση σε όλες τις στήλες των πινάκων που επιλέγονται για επεξεργασία.

```
CREATE OR REPLACE FUNCTION replace_punctuation(page_title
STRING)
RETURNS STRING
LANGUAGE PYTHON {

      import string
      import numpy as np

      punctuation = [",", ":", ";", "!", "?", "(", ")", "[",
"]", "{", "}", "/", "|",    '"', "*"]
      punctuation_string = ''
      for p in punctuation:
              punctuation_string += p

      strip_pnctuation = str.maketrans(punctuation_string,'
','-')
      return np.array([i.translate(strip_pnctuation) for i in
page_title])
};
```

replace_err

Η συνάρτηση replace_err δεχεται σαν όρισμα, δηλαδή σαν είσοδο, μία συμβολοσειρά την page_title. Σε αυτή ψάχνει αν υπάρχουν κάποια συνήθη ορθογραφικά λάθη που γίνονται κατά την πληκτρολόγηση της μάρκας διαφορων φωτογραφικών μηχανών και που έχουν οριστεί στην alianses. Αν βρεθεί κάποιο από αυτά στην συμβολοσειρά το αλλάζει με αυτό που έχει θεωρηθεί ως σωστό. Όταν ολοκληρωθεί η διαδικασία για όλες τις λέξεις που υπάρχουν στην συμβολοσειρά επιστρέφει την νέα συμβολοσειρά. Τα λάθη που έχουν οριστεί στην alianses είναι τα πιο συνήθη όπως αναφέρθηκε. Μπορεί να γίνει και περαιτέρω επέκταση της και με άλλα. Χρησιμοποιείται στον πίνακα speciality και στον πίνακα με τις φωτογραφικές μηχανές.

```
CREATE OR REPLACE FUNCTION replace_err(page_title STRING)
RETURNS STRING
LANGUAGE PYTHON {

aliases =
{"cannon":"canon","canonpowershot":"canon","eos":"canon","use
dcanon":"canon","fugi":"fujifilm","fugifilm":"fujifilm",
"fuji":"fujifilm","fujufilm":"fujifilm","general":"ge","gopro
s":"gopro","hikvision3mp":"hikvision","hikvisionip":"hikvisio
n",
"bell+howell":"howell","howellwp7":"howell","minotla":"minolt
a","canon&nikon":"nikon","olympuss":"olympus","panosonic":"pa
nasonic",
"pentax":"ricoh","ssamsung":"samsung","repairsony":"sony","el
f":"elph","s480016mp":"s4800","vivicam":"v","plus":"+","1080p
":"","720p":""}

row_list = []
for row in page_title:
    words = row.split(" ")
    new_words = []
    for w in words:
        if w in aliases:
            new_words.append(aliases[w])
        else:
            new_words.append(w)
    row = ' '.join(new_words)
    row_list.append(row)
return row_list
};
```

currentworking

Η συνάρτηση currentworking δέχεται σαν είσοδο την συμβολοσειρά time_period. Έχει την currenttime όπου εκεί υπάρχουν λέξεις κλειδιά που υποδηλώνουν αν βρεθούν σε μια εγγραφή ότι ο ειδικός τη συγκεκριμένη περίοδο απασχολείται από αυτή την εταιρεία. Για να είναι πιο εύκολο σε κάποια ενδεχόμενη αναζήτηση να εμφανιστεί η τρέχουσα θέση κάποιου ειδικού, προστίθεται μια στήλη που έχει την τιμή "1" (ένα) αν αυτή την περίοδο δουλεύει σε αυτή την εταιρεία. Χρησιμοποιήθηκε στον πίνακα experience_tbl ώστε να είναι δυνατό να προσδιοριστεί που εργάζεται ο κάθε ειδικός τη συγκεκριμένη χρονική περίοδο.

```
CREATE OR REPLACE FUNCTION currentworking(time_period STRING)
RETURNS STRING
LANGUAGE PYTHON {
currenttime = ["2021", "today", "now"]
currently_column = []
for row in time_period:
     words = row.split(" ")
     currently= "no"
     for w in words:
          if w in currenttime:
               currently="1"
               break
     currently_column.append(currently)
return currently_column
};
```

branding

Η συνάρτηση branding παίρνει σαν είσοδο την συμβολοσειρά page_title. Έχει την brands όπου βρίσκονται οι μάρκες των καμερών. Αν βρεθεί στη συμβολοσειρά εισόδου μια λέξη που είναι ίδια με μία από αυτές που υπάρχουν στη λίστα brands την κρατάει και την επιστρέφει. Με αυτό τον τρόπο αναγνωρίζει τι μάρκα είναι η κάθε φωτογραφική μηχανή. Χρησιμοποιείται στον πίνακα με τις φωτογραφικές μηχανές.

```
CREATE OR REPLACE FUNCTION branding(page_title STRING)
RETURNS STRING
LANGUAGE PYTHON {
brands = ["aiptek", "apple", "argus", "benq", "canon",
"casio", "coleman", "contour", "dahua", "epson", "fujifilm",
"garmin",
"ge", "gopro", "hasselblad","hikvision","howell", "hp",
"intova", "jvc", "kodak", "leica", "lg", "lowepro","lytro",
"minolta",
"minox", "motorola", "mustek", "nikon","olympus",
"panasonic", "pentax", "philips", "polaroid", "ricoh",
```

```
"sakar", "samsung",
"sanyo", "sekonic", "sigma", "sony", "tamron",
"toshiba","vivitar", "vtech", "wespro", "yourdeal"]
brand_column = []
for row in page_title:
     words = row.split(" ")
     brand = "None"
     for w in words:
          if w in brands:
               brand = w
               break
     brand_column.append(brand)
return brand_column
};
```

location

Η συνάρτηση location παίρνει σαν είσοδο τη συμβολοσειρά location. Κάνει αναζήτηση στη συμβολοσειρά που του δίνεται για να βρει αν υπάρχει κάποια από τις πόλεις που έχουν οριστεί στο locations. Αν βρεθεί κάποια την επιστρέφει αλλιώς επιστρέφει το "None". Χρησιμοποιείται στους πίνακες cv_tbl και experience_tbl. Είναι χρήσιμη καθώς σε άλλα βιογραφικά μπορεί να αναφέρεται η πλήρης διεύθυνση, σε άλλα μόνο η πόλη και σε άλλα να παραλείπεται η πόλη. Η πλήρης διεύθυνση είναι περιττή. Το χρήσιμο στοιχείο κατά την αναζήτηση ενός ειδικού είναι η τοποθεσία του και συγκεκριμένα η πόλη. Στον πίνακα cv_tbl κάνει αναζήτηση ώστε να βρεθεί σε ποια πόλη διαμένει ο ειδικός. Στον πίνακα experience_tbl κάνει αναζήτηση για να γίνει γνωστό σε ποια πόλη εδρεύει η κάθε εταιρεία.

```
CREATE OR REPLACE FUNCTION location(location STRING)
RETURNS STRING
LANGUAGE PYTHON {
locations = ["athens", "tripoli", "larisa", "lamia",
"nauplio", "arta", "agrinio", "katerini", "kavala" ,
"hrakleio", "xania", "sparti", "mitilini", "ioannina",
"trikala", "kozani", "komotini", "korinthos", "xalkida",
"rodos", "drama", "veroia", "karditsa", "mitilini", "purgos",
"megara", "xios", "preveza", "serres", "florina", "naousa",
"thiva", "alexandroupoli", "xanthi" ,"thessaloniki",
"xalkida", "volos", "kalamata", "patra", "argos"]
location_column = []
for row in location:
     words = row.split(" ")
     city= "None"
     for w in words:
          if w in locations:
               city = w
```

```
                break
     location_column.append(city)
return location_column
};
```

blocking

Στην προσπάθεια να εφαρμοστεί blocking technique (τεχνική αποκλεισμού) δημιουργήθηκε μια συνάρτηση που αναθέτει ένα block id για κάμερες που ανήκουν στην ίδια μάρκα. Η συνάρτηση blocking παίρνει σαν όρισμα το brand_column που περιέχει τις μάρκες των φωτογραφικών μηχανών και σε κάθε μάρκα δίνει ένα μοναδικό block_id. Χρησιμοποιείται στον πίνακα με τις φωτογραφικές μηχανές.

```
CREATE OR REPLACE FUNCTION blocking(brand_column STRING)
RETURNS STRING
LANGUAGE PYTHON {

     brand_column_distinct = list(set(brand_column))

     block_ids = []
     counter_id = 0;
     for brand in brand_column_distinct:
          if brand == "None":
               block_ids.append("None")
          else:
          block_ids.append(counter_id)
               counter_id+=1
     block_id_dict = {brand_column_distinct[i]: block_ids[i]
for i in range(len(brand_column_distinct))}
     return [block_id_dict[brand] for brand in brand_column]
};
```

blockingcities

Η συνάρτηση blockingcities παίρνει σαν είσοδο τη συμβολοσειρά city που περιέχει τις πόλεις που διαμένουν οι ειδικοί τεχνικοί επισκευής φωτογραφικών μηχανών. Εφαρμόζει blocking technique (τεχνική αποκλεισμού) και αναθέτει ένα block_id για ειδικούς που διαμένουν στην ίδια πόλη. Σε κάθε πόλη δίνει ένα μοναδικό block_id ενώ αν κάπου δεν υπάρχει πόλη επιστρέφει το "None". Χρησιμοποιείται στον πίνακα cv_tbl ώστε να ομαδοποιηθούν οι πόλεις που διαμένουν οι ειδικοί.

```
CREATE OR REPLACE FUNCTION blockingcities(city STRING)
RETURNS STRING
LANGUAGE PYTHON {

     city_distinct = list(set(city))

     block_ids = []
     counter_id = 0;
     for cities in city_distinct:
          if cities == "None":
               block_ids.append("None")
          else:
          block_ids.append(counter_id)
               counter_id+=1
     block_id_dict = {city_distinct[i]: block_ids[i] for i
in range(len(city_distinct))}
     return [block_id_dict[cities] for cities in city]
};
```

Filtering

Στην προσπάθεια να φιλτραριστούν οι κάμερες με βάση το μοντέλο τους δημιουργείται η συνάρτηση modelling. Αυτή η συνάρτηση παίρνει σαν είσοδο τη συμβολοσειρά brands. Παρατηρούμε ότι τα μοντέλα σχεδόν πάντα αποτελούνται από αριθμούς που ακολουθούνται ή προηγούνται γραμμάτων. Άρα θα χρειαστεί να φτιάξουμε δύο patterns. Σε περίπτωση που αναγνωριστεί ότι μια λέξη έχει αυτή την μορφή τη θεωρεί ως το μοντέλο της κάμερας και την επιστρέφει. Χρησιμοποιείται στον πίνακα με τις κάμερες για να βρει το μοντέλο κάθε κάμερας.

```
CREATE OR REPLACE FUNCTION modelling(brands STRING)
RETURNS STRING
LANGUAGE PYTHON {
import re

pattern_1 = '[a-z]{2,4}\d{1,4}[a-z]{0,2}'
pattern_2 = '[a-z]{1,4}\d{1,4}[a-z]{0,2}'
models = []

for t in brands:
    model = 'None'
    words = t.split(" ")[:5]
    for word in words:
        if re.match(pattern_1, word):
```

```
            model = word
            models.append(model)
            break
        elif re.match(pattern_2, word):
            model = word
            models.append(model)
            break
    if model=='None':
        models.append('None')

return models
};
```

Παρακάτω θα δούμε πως γίνεται η κλήση των παραπάνω συναρτήσεων ώστε να γίνει η επεξεργασία διαφόρων στηλών των πινάκων.

Αρχικά παρακάτω φαίνεται πως χρησιμοποιήθηκαν οι συναρτήσεις για τους πίνακες στους οποίους είναι αποθηκευμένα τα δεδομένα των βιογραφικών.

```
UPDATE project.speciality
SET speciality=
replace_punctuation(convert_lower(speciality));

ALTER TABLE project.cv_tbl ADD COLUMN city_living_now STRING;
UPDATE project.cv_tbl
SET city_living_now =
location(replace_punctuation(convert_lower(location)));

ALTER TABLE project.cv_tbl ADD COLUMN city_id STRING;
UPDATE project.cv_tbl
SET city_id = blockingcities(city_living_now);

ALTER TABLE project.experience_tbl ADD COLUMN company_city
STRING;
UPDATE project.experience_tbl
SET company_city =
location(replace_punctuation(convert_lower(companylocation)))
;

ALTER TABLE project.experience_tbl ADD COLUMN company_city_id
STRING;
UPDATE project.experience_tbl
SET company_city_id = blockingcities(company_city);


ALTER TABLE project.experience_tbl ADD COLUMN
company_block_id STRING;
UPDATE project.experience_tbl
SET company_block_id =
blockingcities(replace_punctuation(convert_lower(company)));
```

```
ALTER TABLE project.experience_tbl ADD COLUMN
now_working_block_id STRING;
UPDATE project.experience_tbl
SET now_working_block_id =
currentworking(replace_punctuation(convert_lower(time_period)
));
```

Για τον πίνακα tbl_camera_specs πραγματοποιήθηκε η παρακάτω επεξεργασία με τη χρήση των συναρτήσεων.

```
ALTER TABLE project.tbl_camera_specs ADD COLUMN block_id
STRING;

UPDATE project.tbl_camera_specs
SET block_id =
blocking(branding(replace_err(replace_punctuation(convert_low
er(page_title)))));


ALTER TABLE project.tbl_camera_specs ADD COLUMN model STRING;
UPDATE project.tbl_camera_specs
SET model =
modelling(replace_punctuation(convert_lower(page_title)));
```

Με αυτό τον τρόπο θα έχουμε μια συλλογή από clusters με κάμερες που θα ταιριάζουν στη μάρκα και το μοντέλο.

Μπορούμε να δούμε πόσο αποτελεσματική ήταν η όλη επεξεργασία φτιάχνοντας δύο επιπλέον βοηθητικά tables, ένα για τις κάμερες που βρέθηκε το μοντέλο και η μάρκα και ένα για αυτές που δεν βρέθηκε κάτι από τα δύο.

```
CREATE TABLE tbl_cameras_matched AS
     SELECT * FROM project.tbl_camera_specs
          WHERE block_id <> 'None' AND model <> 'None';

CREATE TABLE tbl_cameras_unmatched AS
     SELECT * FROM project.tbl_camera_specs
          WHERE block_id = 'None' OR model = 'None';
```

Κεφάλαιο 7. Αποτελέσματα

Παρακάτω φαίνεται πώς έχει διαμορφωθεί η τελική εικόνα των πινάκων από τα βιογραφικά των ειδικών αφού καταχωρήθηκαν τα δεδομένα και μετά την επεξεργασία που έχει γίνει.

Ο πρώτος πίνακας είναι ο cv_tbl όπου συντηρούνται οι προσωπικές πληροφορίες των ειδικών αποτελείται από τις στήλες id που είναι το πρωτεύων κλειδί, την στήλη fullname που περιέχει το ονοματεπώνυμο του ειδικού, την στήλη email που έχει τα στοιχεία ηλεκτρονικής αλληλογραφίας του ειδικού, τη στήλη birthday που περιέχει την ημερομηνία γέννησης, τη στήλη location που αφορά τον τόπο διαμονής του, την στήλη city_living_now όπου φαίνεται σε ποια πόλη μένει και την στήλη city_id που έχει αποδοθεί ένα id για κάθε πόλη:

	id	fullname	email	birthday	location	city_living_now	city_id
1	34	Anastasia Florou	floroua@gmail.com	22/08/1984	Xania, Crete, Greece	xania	18
2	1	Ashley Miller	asmiller@gmail.com	25/8/1990	Athens, Greece	athens	16
3	2	Dimitrios Sotos	sotosd@gmail.com	25/12/1965	Athens, Greece	athens	16
4	15	Fotios Kotsakis	kotsakis@gmail.com	03/06/1960	Patra, Greece	patra	10
5	3	George Pappas	gpappas@gmail.com	28/11/1985	Xaidari, Athens, Greece	athens	16
6	23	Georgia Xristopoulou	xristg@gmail.com	28/12/1974	Korinthos, Greece	korinthos	9
7	17	Giorgos Andreou	angiorgos@gmail.com	03/06/1965	Trikala, Greece	trikala	5
8	38	Giorgos Asimakopoulos	asimakg@gmail.com	25/01/1980	Kalamata, Greece	kalamata	12
9	22	Giorgos Panoudakis	panoudg@gmail.com	28/10/1987	Ioannina, Greece	ioannina	6
10	13	Ioanna Metaksa	ioanna909@gmail.com	03/06/1989	Peristeri, Athens, Greece	athens	16
11	4	John Opan	opanj@gmail.com	18/9/1989	Thessaloniki, Greece	thessaloniki	11
12	26	Karampas Arhs	kararhs@gmail.com	20/11/1968	Athens, Greece	athens	16
13	39	Konstantinos Logothetis	logkon@gmail.com	17/02/1983	Athens, Greece	athens	16
14	45	Manolis Kapetanios	mkapet@gmail.com	12/08/1979	Katerini, Greece	katerini	13
15	27	Marialena Exarchou	exarcoum@gmail.gr	03/10/1988	Thessaloniki, Greece	thessaloniki	11
16	35	Marina Apostolou	aposm@gmail.com	22/07/1986	Thiva, Greece	thiva	1
17	36	Marinos Triantos	trianm@gmail.com	22/06/1980	Thessaloniki, Greece	thessaloniki	11
18	5	Marios Giannakakos	gmarios@gmail.com	25/2/1972	Argos, Greece	argos	15
19	50	Marios Ksenos	kswenosm@gmail.com	13/10/1986	Thessaloniki, Greece	thessaloniki	11
20	31	Markos Kampeas	kampm@gmail.com	18/07/1990	Xalkida, Greece	xalkida	7
21	30	Markos Oikonomou	moik@gmail.com	8/3/1992	Tripoli, Greece	tripoli	17
22	44	Napoleon Arvanitis	napar@gmail.com	02/05/1975	Agrinio, Greece	agrinio	3
23	43	Napoleon Arvanitis	napar@gmail.com	02/05/1976	Agrinio, Greece	agrinio	3
24	12	Nikos Apostolakhs	apostnik@gmail.com	03/08/1973	Athens, Greece	athens	16
25	19	Nikos Giannakopoulos	gianaknik@gmail.com	09/07/1974	Kozani, Greece	kozani	14

Αναλύοντας τα δεδομένα με τις προσωπικές πληροφορίες των ειδικών επισκευής φωτογραφικών μηχανών που βρίσκονται στον πίνακα cv_tbl προκύπτουν τα ποσοστά των ειδικών επισκευής φωτογραφικών μηχανών ανά πόλη. Αυτά φαίνονται παρακάτω:

Έπειτα δημιουργήθηκε ο πίνακας experience_tbl ο οποίος φαίνεται παρακάτω και περιέχει τα στοιχεία για την εργασιακή εμπειρία των ειδικών. Σε αυτόν, όπως βλέπουμε, έχουμε τη στήλη id που είναι το foreign key, τη στήλη company που είναι η εταιρεία που έχει δουλέψει ή δουλεύει, τη στήλη companylocation που περιέχει την τοποθεσία που εδρεύει η εταιρεία, τη στήλη time_period που αναφέρεται στην περίοδο απασχόλησης σε κάθε εταιρεία, τη στήλη description όπου περιγράφονται οι αρμοδιότητές του και τις στήλες company_city_id και now_working_block που δημιουργήθηκαν κατά την επεξεργασία των δεδομένων.

Η στήλη company_city_id είναι το μοναδικό αναγνωριστικό κάθε πόλης.

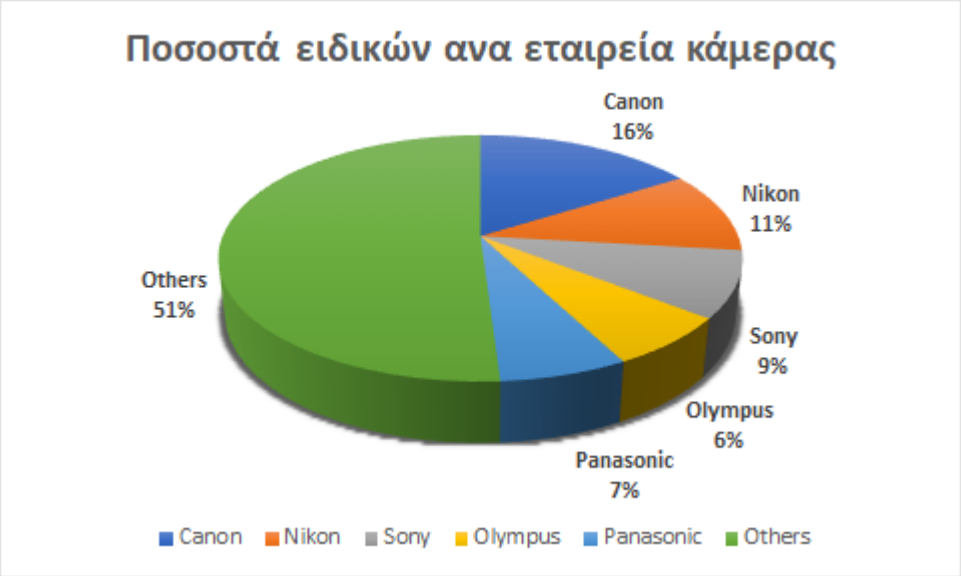
Η στήλη now_working_block υποδηλώνει για κάθε εγγραφή αν αποτελεί την τωρινή εταιρεία απασχόλησης. Παρακάτω παρατίθεται δείγμα από τα δεδομένα που περιέχονται στον πίνακα.

	id	company	companylocation	time_period	description	company_city_id	now_working_block
1	40	Camera Technicians	Thessaloniki Greece	2004 to 2006	repairing of came	11	no
2	40	camera repair	Thessaloniki, Greece	2006 to 2019	repairing of came	11	no
3	24	EasyRepair	Athens Greece	2004 to 2006	Reraired many dif	17	no
4	24	The Repair Specialists	Athens Greece	2007 to 2010	Specialized in rep	17	no
5	24	The Repair Specialists	Hrakleio of Crete, Gree	2011 to now	working with vari	4	1
6	41	The repair Specialists	Sparti Greece	2006 to 2014	repairing of came	8	no
7	41	Camera Repair Exper	Sparti, Greece	2016 to 2021	repairing of came	8	1
8	49	Camera EasyRepair s	Thessaloniki Greece	2009 to 2011	gained experience	11	no
9	49	Canon camera repair	Thessaloniki, Greece	2011 to 20013	Fixing various can	11	no
10	49	Camera EasyRepair s	Thessaloniki Greece	2015 to 2020	experienced in se	11	no
11	12	fotoklik	Athens Greece	1999 to 2005	gained experience	17	no
12	12	camera techninians	Athens, Greece	2006 to 2013	Fixing various can	17	no
13	12	camera masters	Athens Greece	2013 to 2021	experienced in se	17	1
14	11	digital lab service	Athens Greece	2000 to 2008	Worked with many	17	no
15	11	service video	Athens, Greece	2008 to 2015	Fixing various can	17	no
16	11	fotoklik	Athens Greece	2015 to 2021	experienced in se	17	1
17	15	Fotis Kotsakis camer	Patra Greece	1975 to 2021	repair every brand	10	1
18	19	fotoklik	Athens, Greece	1998 to 1999	repairing of came	17	no
19	19	COMPULAND	Kozani, Greece	1999 to 2021	repairing of came	14	1
20	39	ifixit	Athens Greece	2007 until 2013	Fixing canon, fujif	17	no
21	39	Camera Technicians	Athens, Greece	2015 until 2020	worked with casic	17	no
22	14	Sony Service	Patra Greece	2005 to 2012	repairing of came	10	no
23	14	fotis kotsakhs camer	Patra, Greece	2012 to 2021	repairing of came	10	1

Ο τελευταίος πίνακας που σχετίζεται με τα βιογραφικά των ειδικών είναι ο πίνακας speciality που αποτελείται από τη στήλη id που είναι το foreign key και τη στήλη speciality που περιέχει τις μάρκες στις οποίες ειδικεύεται ο κάθε ειδικός. Δείγμα των δεδομένων του πίνακα speciality:

	id	speciality
1	40	canon
2	40	sony
3	40	apple
4	40	intova
5	24	olympus
6	24	canon
7	24	nikon
8	41	canon
9	41	hp
10	41	contour
11	49	canon
12	49	ricoh
13	49	howell
14	49	hp
15	12	canon

Αναλύοντας τα δεδομένα του πίνακα speciality φαίνονται τα ποσοστά εξειδίκευσης των ειδικών επισκευής ανά φωτογραφική μηχανή. Το μεγαλύτερο ποσοστό το κατέχει η Canon με ποσοστό δεκαέξι τοις εκατό.



Όσον αφορά τα δεδομένα με τις φωτογραφικές μηχανές, στην παρακάτω εικόνα φαίνεται πως έχει διαμορφωθεί ο πίνακας project.tbl_camera_specs μετά τις αλλαγές που έγιναν κατά την επεξεργασία. Όπως φαίνεται υπάρχει αρχικά το πρωτεύον κλειδί που είναι το specid, μετά μία στήλη με τους τίτλους των σελίδων, μετά οι πληροφορίες που αναφέρονται στην κάθε σελίδα για την εκάστοτε φωτογραφική μηχανή και μετά ακολουθούν δύο στήλες οι οποίες δημιουργήθηκαν κατά την επεξεργασία των δεδομένων, δηλαδή οι στήλες block_id και model.

Το block_id περιγράφει την μάρκα της κάθε φωτογραφικής μηχανής. Για παράδειγμα το block_id 36 δηλώνει ότι η φωτογραφική μηχανή είναι Canon ενώ το block_id 18 δηλώνει ότι η μηχανή είναι Nikon.

Το model δείχνει τι μοντέλο είναι η κάθε φωτογραφική μηχανή. Σε κάποιες κάμερες έχει βρεθεί επιτυχώς ενώ σε κάποιες άλλες όχι.

	specId	page_title	info	block_id	model
1	www.pcconnection.com//12419	Buy Canon PowerShot SX600 HS, 16MI	450 mm25 mm5.80 ozLithium ionPoin	36	sx600
2	www.pcconnection.com//12179	Buy Nikon COOLPIX AW120 Waterpro	120.00 mm24.00 mmCompactCamoul	18	aw120
3	www.pcconnection.com//4463	Buy Olympus Stylus SH-1 Digital Came	CompactSilver3 inPop-up flash16 me	26	sh1
4	www.pcconnection.com//12401	Buy Pentax WG-4 Digital Camera, 16MI	100.00 mm25.00 mmCompactSilver3 i	35	wg4
5	www.pcconnection.com//4422	Buy Olympus E-P5 PEN Mirrorless Digi	Mirrorless systemWhite3 inPop-up fla	26	ep5
6	www.pcconnection.com//12196	Buy Panasonic DMC-LX7 Digital Came	CompactBlack3 inPop-up flash10.1 m	43	None
7	www.pcconnection.com//12443	Buy Sony Cyber-shot Digital Camera R	100.00 mm28.00 mm350 image(s)Lith	24	None
8	www.pcconnection.com//12370	Buy Pentax K-S1 Camera, Blue, with Le	This product is subject to our return p	35	ks1
9	www.pcconnection.com//12382	Buy Sony Cyber-Shot WX80 Digital Car	CompactWhite2.7 inBuilt-in flash16.2	24	wx80
10	www.pcconnection.com//12437	Buy Sony DSC-TX30 Camera Cameras -	CompactPink3.3 inBuilt-in flash18.2 m	24	None
11	www.pcconnection.com//12378	Buy Fujifilm FinePix JX680 Digital Cam	CompactRed3 inBuilt-in flash16 mega	39	jx680
12	www.pcconnection.com//12374	Buy Canon Canon EOS Rebel T4i Digita	Lithium ionSLR cameraBlackCamera,	36	None
13	www.pcconnection.com//12430	Buy Canon PowerShot SX700 HS Digita	750 mm25 mm9.50 ozLithium ionPoin	36	sx700
14	www.pcconnection.com//4437	Buy Nikon WR-R10 Wireless Remote T	This product is subject to our return p	18	wrr10
15	www.pcconnection.com//12450	Buy Nikon COOLPIX AW120 Waterpro	120.00 mm24.00 mmCompactOrange	18	aw120
16	www.pcconnection.com//12380	Buy Sony DSC-TX30 Camera Cameras -	CompactOrange3.3 inBuilt-in flash18.:	24	None
17	www.pcconnection.com//4461	Buy Olympus Stylus SH-1 Digital Came	CompactBlack3 inPop-up flash16 me	26	sh1
18	www.pcconnection.com//12165	Buy Pentax Q7 Compact Mirrorless Ca	Mirrorless systemSilver3 inPop-up fla	35	q7
19	www.pcconnection.com//12186	Buy Sony Cyber-shot DSC-H300 Digital	175 minute(s)CompactBlack3 inPop-u	24	dsch300
20	www.pcconnection.com//4439	Buy Olympus E-P5 PEN Mirrorless Digi	Mirrorless systemBlack3 inPop-up fla	26	ep5
21	www.pcconnection.com//4453	Buy Canon EOS 5D Mark III DSLR Came	30.30 ozLithium ionSLR cameraBlackN	36	None
22	www.pcconnection.com//12402	Buy Canon PowerShot ELPH 150 IS Dig	240 mm24 mm5.00 ozLithium ionPoin	36	None
23	www.pcconnection.com//12415	Buy Fujifilm FinePix SL1000 Digital Car	This product is subject to our return p	39	sl1000
24	www.pcconnection.com//4408	Buy Nikon J3 Interchangable Lens Digi	Mirrorless systemWhite3 inPop-up fla	18	j3
25	www.pcconnection.com//12372	Buy Canon PowerShot ELPH 150 IS, 20	240 mm24 mm4.40 ozLithium ionPoin	36	None
26	www.pcconnection.com//12369	Buy Pentax Ricoh GR Digital Camera, 1	CompactBlack3 inPop-up flash16.2 m	35	None
27	www.pcconnection.com//12376	Buy Sony DSC-HX300 Camera - Black C	CompactBlack3 inPop-up flash20.4 m	24	None
28	www.pcconnection.com//12418	Buy Nikon Nikon 1 V2 Digital Camera v	Mirrorless systemWhite3 inPop-up fla	18	v2
29	www.pcconnection.com//12449	Buy Sony Smartphone Attachable Len	Smartphone attachableBlack18.2 me	24	None

Διαπιστώνουμε ότι έχει βρεθεί η μάρκα σε 24694 από τις 29787 φωτογραφικές μηχανές και το μοντέλο σε 16623 από τις 29787. Μιλώντας με ποσοστά βρέθηκε η μάρκα στο ογδόντα τρία τοις εκατό των εγγραφών και το μοντέλο στο πενήντα έξι τοις εκατό.

Για την αξιολόγηση της συνάρτησης που βρίσκει το μοντέλο των φωτογραφικών μηχανών και της συνάρτησης που βρίσκει τη μάρκα των φωτογραφικών μηχανών δημιουργούνται δύο πίνακες. Στον ένα πίνακα αποθηκεύονται οι φωτογραφικές μηχανές για τις οποίες έχει βρεθεί και η μάρκα και το μοντέλο της φωτογραφικής μηχανής. Στον άλλο πίνακα αποθηκεύονται αυτές που δεν έχει βρεθεί ένα από τα δύο ή και τα δύο.

Παρατηρώντας το πλήθος των εγγραφών στους βοηθητικούς πίνακες που χρησιμοποιήσαμε διαπιστώνουμε ότι έχουν βρεθεί και μάρκα και μοντέλο σε πάνω από τις μισές εγγραφές το οποίο είναι ένα πολύ θετικό δείγμα καθώς δεν πρέπει να ξεχνάμε ότι σε κάποιους τίτλους μπορεί να μην αναφέρεται κάτι από τα δύο ή μπορεί να μην αναφέρεται σε φωτογραφικές μηχανές αλλά σε εξάρτημα τους. Επιπλέον παρατηρείται ότι στις περισσότερες φωτογραφικές μηχανές που βρέθηκε η μάρκα έχει βρεθεί και το μοντέλο.

Τέλος, παρατίθενται ενδεικτικά οι χρόνοι εκτέλεσης κάποιων βασικών διαδικασιών για τα δεδομένα των φωτογραφικών μηχανών, όπως η φόρτωση των αρχικών δεδομένων, η αποθήκευσή τους καθώς και η επεξεργασία τους.

Execution Log

Time	Type	Text	Duration (ms)	Rows	Result
Jul-08 20:37:44	SQL / User scr	SELECT COUNT(*) FROM tbl_cameras_unmatched	6	1	Success
Jul-08 20:37:43	SQL / User scr	SELECT * FROM project.tbl_data	147	200	Success
Jul-08 20:37:43	SQL / User scr	select * from project.tbl_camera_specs	279	200	Success
Jul-08 20:37:43	SQL / User scr	SELECT COUNT(*) FROM tbl_cameras_matched	6	1	Success
Jul-08 20:37:39	SQL / User scr	CREATE TABLE tbl_cameras_unmatched AS¶ SELECT * FRO	3,569		Success
Jul-08 20:37:38	SQL / User scr	----------CREATE TABLE----------¶----------	498		Success
Jul-08 20:37:36	SQL / User scr	UPDATE project.tbl_camera_specs¶SET model = modelling	1,986	29787	Success
Jul-08 20:37:36	SQL / User scr	ALTER TABLE project.tbl_camera_specs ADD COLUMN mod	97		Success
Jul-08 20:37:31	SQL / User scr	UPDATE project.tbl_camera_specs ¶SET block_id = blockin	5,522	29787	Success
Jul-08 20:37:31	SQL / User scr	----------ALTER TABLE----------¶----------	110		Success
Jul-08 20:37:31	SQL / User scr	----------6th UDF----------¶----------	13		Success
Jul-08 20:37:31	SQL / User scr	----------5th UDF----------¶----------	14		Success
Jul-08 20:37:31	SQL / User scr	----------4th UDF----------¶----------	63		Success
Jul-08 20:37:31	SQL / User scr	----------3rd UDF----------¶----------	21		Success
Jul-08 20:37:31	SQL / User scr	----------2nd UDF----------¶----------	56		Success
Jul-08 20:37:31	SQL / User scr	----------1st UDF----------¶----------	14		Success
Jul-08 20:37:30	SQL / User scr	ALTER TABLE "tbl_data" ADD CONSTRAINT "tbl_data_fk2" ¶	119		Success
Jul-08 20:37:30	SQL / User scr	----------FOREIGN KEYS----------¶ALTER TABLE	260		Success
Jul-08 20:37:25	SQL / User scr	SELECT *¶FROM project.tbl_camera_specs AS T¶INNER JO	5,252	200	Success
Jul-08 20:37:25	SQL / User scr	ALTER TABLE tbl_camera_specs ADD PRIMARY KEY ("specIc	93		Success
Jul-08 20:37:16	SQL / User scr	----------CREATE TABLE TBL_CAMERA_SEPCS----------	8,341		Success
Jul-08 20:37:16	SQL / User scr	---------- JSON LOADER ----------¶--Retri	70		Success
Jul-08 20:37:16	SQL / User scr	-- The first row of our csv file is not needed, thus we delete it	46	1	Success
Jul-08 20:37:16	SQL / User scr	COPY INTO project.tbl_data FROM '/home/vasiliki/Downlo	313	46666	Success
Jul-08 20:37:15	SQL / User scr	----------CREATE TABLE TBL_TRAINING_DATA CSV F	627		Success
Jul-08 20:37:15	SQL / User scr	drop table tbl_cameras_unmatched	56		Success
Jul-08 20:37:14	SQL / User scr	drop table tbl_cameras_matched	864		Success
Jul-08 20:37:14	SQL / User scr	drop table tbl_camera_specs	38		Success
Jul-08 20:37:14	SQL / User scr	drop table tbl_data	48		Success

Name	Value
Queries	29
Updated Rows	106241
Execute time (ms)	27479
Fetch time (ms)	12
Total time (ms)	27491
Finish time	2021-07-08 20:37:44.918

Παρατηρούμε πως οι χρόνοι εκτέλεσης των διαδικασιών είναι πολύ χαμηλοί αναλογικά με τον όγκο των δεδομένων που έχουμε στη διάθεσή μας.

Δεν έχουν παρατεθεί οι χρόνοι για τα δεδομένα των τεχνικών, καθώς ο όγκος τους δεν είναι αξιοσημείωτος, και ο χρόνος εκτέλεσης των διαδικασιών είναι χαμηλός, όπως θα περίμενε κανείς.

Κεφάλαιο 8. Συμπεράσματα-προτάσεις

Οι σημερινές επιχειρήσεις και οργανισμοί αναζητούν και διαχειρίζονται τη γνώση καθώς είναι ζωτικής σημασίας για την επιβίωσή τους. Η δημιουργία και η διάδοση της γνώσης γίνονται όλο και πιο σημαντικοί παράγοντες ανταγωνισμού μεταξύ των οργανισμών κάνοντας τους οργανισμούς να στραφούν στη δημιουργία διαφόρων συστημάτων διαχείρισης γνώσης.

Το σύστημα που δημιουργήθηκε αποδεικνύει ότι είναι εφικτή τεχνικά η δημιουργία ενός Συστήματος Μητρώου Ειδικών που μπορεί να βοηθήσει στην εύρεση ειδικών και τελικά να βοηθήσει τον οργανισμό.

Μελλοντικές επεκτάσεις που μπορούν να γίνουν είναι αρχικά να δημιουργηθεί εφαρμογή ώστε τα δεδομένα και οι πληροφορίες να είναι προσβάσιμα από όλους και ειδικά από αυτούς που πρόκειται να χρησιμοποιήσουν ένα τέτοιο σύστημα. Στην εφαρμογή ο κάθε ειδικός θα δημιουργεί ένα προσωπικό προφίλ και θα συμπληρώνει τα στοιχεία του, την ειδικότητά του, την εργασιακή του εμπειρία και πληροφορίες για τα ενδιαφέροντά του. Το περιβάλλον θα πρέπει να είναι φιλικό προς τον χρήστη τόσο όταν δημιουργεί το προφίλ του όσο και στην αναζήτηση για την εύρεση του κατάλληλου ειδικού. Επιπλέον μπορεί να προστεθεί στην εφαρμογή λειτουργία άμεσης επικοινωνίας (instant chat) με τους ειδικούς και αποστολή αυτοματοποιημένων μηνυμάτων ή μηνυμάτων ηλεκτρονικού ταχυδρομείου ώστε να υπενθυμίζουν στους χρήστες να έχουν ενημερωμένες τις πληροφορίες τους. Το τελευταίο συγκεκριμένα θα βοηθήσει στο να μην υπάρχουν μεγάλες αποκλίσεις σε αυτά που αναφέρονται και σε αυτά που ισχύουν στην πραγματικότητα και θα βοηθήσει να είναι λειτουργικό το Σύστημα Μητρώου Ειδικών.

Βιβλιογραφία

Κεφάλαια Βιβλίων

1) Dalkir. 2011. Knowledge Management in Theory and Practice Second Edition, The MIT Press

Βασική πηγή για το θεωρητικό μέρος της διπλωματικής. Χρησιμοποιήθηκαν κυρίως τα κεφάλαια 1, 5.

2) Davenport, T., and L. Prusak. 1998. Working knowledge: How Organizations Manage What They Know, Boston, MA : Harvard Business School Press.
Χρησιμοποιήθηκε η παρακάτω σύνοψη:

https://www.researchgate.net/publication/229099904_Working_Knowledge_How_Organizations_Manage_What_They_Know

3) Davenport, T. 2005. Thinking for a living, how to get better performance and results from knowledge workers, Boston, MA : Harvard Business School Press.

Χρησιμοποιήθηκε η παρακάτω σύνοψη:

https://www.researchgate.net/publication/248078273_Thinking_for_A_Living_How_to_Get_Better_Performance_and_Results_from_Knowledge_Workers

4) Klein, D. 1998. The strategic management of intellectual capital. Oxford, UK : Butterworth-Heineman, Oxford .

Χρησιμοποιήθηκε η παρακάτω σύνοψη:

https://www.taylorfrancis.com/books/mono/10.4324/9780080517926/strategic-management-intellectual-capital-david-klein

5) Pasternack, B. , and A. Viscio. 1998. The centerless corporation. Simon and Schuster

Χρησιμοποιήθηκε η παρακάτω σύνοψη:

https://www.strategy-business.com/article/19182

6) Stewart, T. 2007. The wealth of knowledge: Intellectual capital and the twenty-first century organization.

books.google.com

7) Μαρινάγη. Α. και Χ. Σκουρλάς. 2021. Διαχείρισης Γνώσης, (υπό έκδοση) Κάλλιπος

Άρθρα

1) Barth, S. 2000. Heeding the sage of the knowledge age. CRM Magazine. May.

2) Collison, C. 2005. Knowledge Management - Creating a Sustainable Yellow Pages System

https://ezinearticles.com/?Knowledge-Management---Creating-a-Sustainable-Yellow-Pages-System&id=12271

3) Drucker, P. 1994. The social age of transformation. Atlantic Monthly, http://www.providersedge.com/docs/leadership_articles/Age_of_Social_Transformation.pdf

4) Lamont, J. 2003. Expertise location and the learning organization. KMWorld Magazine 12 (1).

5) Pfeffer, J., and R. Sutton. 2001. The Knowing-Doing Gap: How Smart Companies Turn Knowledge into Action, Supply Chain Management. Volume 6 Issue 3

6) Nonaka, I. 2007. The knowledge-creating company. Harvard Business Review, pp. 162-171

7) Ruggles, R. 1996. Knowledge Management Tools: An introduction, *in the Knowledge Management Tools.* Butterworth-Heinemann, pp. 1-10

Πηγές στο διαδίκτυο

Documentation της MonetDB https://www.monetdb.org/Documentation

Πληροφορίες για το εργαλείου MonetDB https://en.wikipedia.org/wiki/MonetDB

Πληροφορίες για την γλώσσα προγραμματισμού Python https://en.wikipedia.org/wiki/Python_(programming_language)

Πληροφορίες για τα Συστήματα Μητρώου Ειδικών, Yellow Pages/Expertise Locator Systems/Staff Directories https://kstoolkit.org/Yellow+pages

Πληροφορίες για τα Συστήματα Μητρώου Ειδικών. Garfield S. May 24, 2017 Expertise Locators and Ask the Expert https://stangarfield.medium.com/expertise-locators-and-ask-the-expert-f273db1e227c

Πληροφορίες για τα συστήματα μητρώου ειδικών. Expertise Locator / Who's Who https://sites.google.com/site/apokmtools/home/8-0-km-tools-manual/expert-locator--whos-who

Πληροφορίες για τα συστήματα μητρώου ειδικών. Kota K. Knowledge Management Part 3 – The Knowledge Yellow Pages http://blog.adminitrack.com/blog/knowledge-yellow-pages/

Τα datasets είναι διαθέσιμα στην ιστοσελίδα:

http://www.inf.uniroma3.it/db/sigmod2020contest/task.html

Παράρτημα Datasets

Αναφορικά με τα δύο γνωστά datasets `camera_specs.tar.gz` και

`sigmod_medium_labelled_dataset.csv` που χρησιμοποιήθηκαν παραθέτουμε τη σχετική ιστοσελίδα:

Programming Contest 2020 ACM SIGMOD

Task Details

The task consists of identifying which product specifications (in short, specs) from multiple e-commerce websites represent the same real-world product.

You are provided with **a dataset including ~30k specs in JSON format**, each spec containing a list of (attribute_name, attribute_value) pairs extracted from a different web page, collected across 24 different web sources. We will refer to this dataset as dataset **X**.

- Each spec is stored as a file, and files are organized into directories, each directory corresponding to a different web source (e.g., www.alibaba.com).
- All specs refer to cameras and include information about the camera model (e.g. Canon EOS 5D Mark II) and, possibly, accessories (e.g. lens kit, bag, tripod). Accessories do not contribute to product identification: for instance, a Canon EOS 5D Mark II that is sold as a bundle with a bag represents the same core product as a Canon EOS 5D Mark II that is sold alone.

```
{
  "<page title>": "Samsung Smart WB50F Digital Camera White
Price in India with Offers & Full Specifications |
PriceDekho.com",
  "brand": "Samsung",
  "dimension": "101 x 68 x 27.1 mm",
  "display": "LCD 3 Inches",
  "pixels": "Optical Sensor Resolution (in MegaPixel)\n16.2
MP"
  "battery": "Li-Ion"
}
```

Note that, while the page title attribute is always present, all other attribute names can vary (even within the same web source). Note also that two attributes with the same name (homonyms) might have different semantics (e.g. "battery" that can refer to "battery type", like "AAA", or "battery chemistry", like "Li-Ion"), and that two attributes with the same semantics (synonyms) might have different names (e.g., "resolution" and "pixels").

You are also provided with **a labelled dataset in CSV format**, containing three columns: "left_spec_id", "right_spec_id" and "label". We will refer to this dataset as dataset **W** (which includes the previously released labelled dataset, referred to as dataset **Y**).

- The "spec_id" is a global identifier for a spec and consists of a relative path of the spec file. Note that instead of "/" the spec_id uses a special character "//" and that there is no extension. For instance, the spec_id "www.ebay.com//1000" refers to the 1000.json file inside the www.ebay.com directory. All "spec_id" in the labelled dataset **W** refer to product specs in dataset **X**. Thus, the dataset **W** provides labels for a subset of the product pairs in the Cartersian product of the specs dataset **X** with itself.
- Each row of the labelled dataset represents a pair of specifications. Label=1 means that the left spec and the right spec refer to the same real-world product (in short, that they are matching). Label=0 means that the left spec and the right spec refer to different real-world products (in short, that they are non-matching).

```
left_spec_id, right_spec_id, label
www.ebay.com//1, www.ebay.com//2, 1
www.ebay.com//3, buy.net//10, 0
```

Note that there might be matching pairs even within the same web source, and that the labelled dataset **W** is transitively closed (i.e., if A matches with B and B matches with C, then A matches with C).

More details about the datasets can be found in the dedicated "Datasets" section.

Your goal is to find all pairs of product specs in dataset **X** that match, that is, refer to the same real-world product. Your output must be stored in a CSV file containing only the matching spec pairs found by your system. The CSV file must have two columns: "left_spec_id" and "right_spec_id": each row in this CSV file consists of just two ids, separated by comma.

```
left_spec_id, right_spec_id
www.ebay.com//10, www.ebay.com//20
www.ebay.com//10, buy.net//100
```

An example CSV file is also included in the Quick Start Package (see "Submitting" section).

Dataset X	Specs Dataset	8.5 Mb (compressed)
Dataset Y	Labelled Dataset (Medium)	2.1 Mb
Dataset W	Labelled Dataset (Large)	13.0 Mb

Note that the labelled dataset W (which includes the previously released labelled dataset, referred to as dataset Y) provided to participants is disjoint from the held-out dataset used in the evaluation process. More details about the evaluation process can be found in the dedicated "Evaluation Process" section.