Σημασιολογικές Τεχνολογίες και Τεχνολογίες Πληροφοριών

έννοιες, αναδρομή, τάσεις

Άρθρο του ***Δημήτρη Φιλίππου***
Technical Director, I²QS PC.
BICSI Country Chair, IEEE Senior Member,
CSI Professional Engineer Member,
ETHERNET ALLIANCE Consulting Member,
ELOT Technical Committee TC93 Member,
CENELEC TC209 & TC215 Delegate of Greek National Committee,
CENELEC TC215 WG1, WG2 & WG3 Member.
e-mail: dfilippou@i2qs.com

ΜΕΡΟΣ Α΄

Οι σημασιολογικές τεχνολογίες βελτιώνουν σημαντικά την ικανότητα κατανόησης της μηχανής και τη διαλειτουργικότητα των δεδομένων. Οι τέσσερις διακριτοί τομείς των αναδυόμενων Τεχνολογιών Πληροφοριών για τη λήψη δεδομένων από τους ανθρώπους και τις μηχανές: τα Συνδεδεμένα Δεδομένα (Linked Data), τα Μέγα Δεδομένα (Big Data), το Διαδίκτυο των Πραγμάτων (Internet of Things - IoT) και τα Μέσα Κοινωνικής Δικτύωσης (Social Media) συσχετίζονται μεταξύ τους, με την τεχνολογία των Συνδεδεμένων Δεδομένων να παίζει τον ρόλο του ολοκληρωτή για την σημασιολογική διαλειτουργικότητα και διασύνδεση των δεδομένων και των τεχνολογιών.

1. ΕΙΣΑΓΩΓΗ

Σήμερα, οι άνθρωποι και οι μηχανές που συνδέονται στο Διαδίκτυο παράγουν ένα τεράστιο όγκο δεδομένων που πρέπει να συλλεχθούν, να αναλυθούν και ουσιαστικά να χρησιμοποιηθούν. Οι άνθρωποι δημιουργούν ογκώδη δεδομένα χρησιμοποιώντας διαφορετικές τεχνολογίες μέσων κοινωνικής δικτύωσης. Οι μηχανές (π.χ. αισθητήρες και συσκευές) δημιουργούν δεδομένα που μεταδίδονται σε υπολογιστές μέσω του Διαδικτύου.

Για την διαχείριση της συνεχώς αυξανόμενης ποσότητας των διαφορετικών τύπων δεδομένων που προέρχονται από διαφορετικές πηγές προέκυψαν διάφορες νέες τεχνολογίες. Για παράδειγμα, οι τεχνολογίες των Μέγα Δεδομένων συμβάλλουν στην επίλυση προβλημάτων που σχετίζονται με τον όγκο, την ποικιλία, τη μεταβλητότητα, την ταχύτητα και την εγκυρότητα των δεδομένων. Η ανάλυση των Μέγα Δεδομένων προσπαθεί να βρει κρυμμένες σχηματομορφές (Pattern) στα δεδομένα για λόγους λήψης αποφάσεων. Ωστόσο, υπάρχει ανάγκη για τεχνολογίες που θα μπορούσαν να λύσουν

σημασιολογικά προβλήματα διαλειτουργικότητας των δεδομένων που παράγονται από τους ανθρώπους και τις μηχανές. Σε γενικές γραμμές, τα προβλήματα αυτά θα μπορούσαν να επιλυθούν με σημασιολογικές τεχνολογίες, όπως οι τεχνολογίες των Συνδεδεμένων Δεδομένων και οι οντολογίες.

Τα αποτελέσματα ενός μεγάλου αριθμού μελετών σχετικά με τις τάσεις έρευνας και τεχνολογίας δείχνουν ότι λόγω της εξαιρετικά γρήγορης σύνδεσης στο Διαδίκτυο το μέλλον ανήκει στον ιστό των συνδεδεμένων δεδομένων και συσκευών που θα επιτρέπει σε πραγματικό χρόνο τη συγκέντρωση και ανάλυση ενός υπερβολικά μεγάλου όγκου δεδομένων. Για παράδειγμα, η Cisco εξετάζει τις τάσεις αυτές από την άποψη των δικτύων και αποκαλεί το φαινόμενο αυτό ως το Διαδίκτυο των Πάντων (Internet of Everything - IoE), εφόσον όλο και περισσότεροι άνθρωποι, τοποθεσίες και πράγματα συνδέονται σε IP δίκτυα.

Κύριο ερώτημα αυτού του άρθρου είναι το εξής: Ποιος θα είναι ο ρόλος της τεχνολογίας των Συνδεδεμένων Δεδομένων στον νέο αναδυόμενο διαδικτυακό ιστό, τον ιστό των πάντων;

Προκειμένου να απαντηθεί το ερώτημα αυτό, πρώτα θα εξετασθούν και θα αξιολογηθούν εν συντομία οι αναδυόμενοι τομείς της τεχνολογίας των Συνδεδεμένων Δεδομένων, Μέγα Δεδομένων, των IoT και των μέσων κοινωνικής δικτύωσης. Για καθέναν από αυτούς τους τομείς, προσεγγίζεται η έννοια, γίνεται ανασκόπηση των εξελίξεων τους και αναφορά των τάσεων. Επιπλέον, παρουσιάζεται πώς οι τεχνολογίες αυτές σχετίζονται μεταξύ τους και ποιός ο ρόλος της τεχνολογίας των Συνδεδεμένων Δεδομένων.

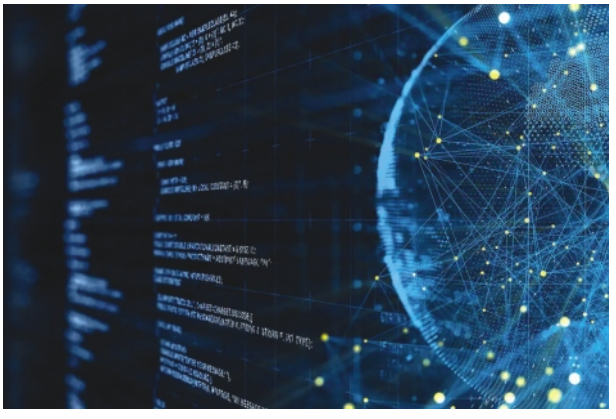
2. ΤΕΧΝΟΛΟΓΙΕΣ ΣΥΝΔΕΔΕΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

2.1. Έννοια

Η τεχνολογία των Συνδεδεμένων Δεδομένων είναι μία από τις σημασιολογικές τεχνολογίες. Τα Συνδεδεμένα Δεδομένα είναι κατανεμημένα, συνδεδεμένα και σημασιολογικά διαλειτουργικά σύνολα δεδομένων που αναπαρίστανται με μια ομοιόμορφη δομημένη μορφή (π.χ. το Πλαίσιο Περιγραφής Πόρων (Resource Description Framework - RDF)), δημοσιεύονται στον Ιστό και είναι προσβάσιμα μέσω ερωτησιακών ακροσημείων (Endpoint) (π.χ. ακροσημεία SPARQL). Η τεχνολογία των Συνδεδεμένων Δεδομένων αναπτύχθηκε πάνω στις υπάρχουσες συμβατικές τεχνολογίες ιστού. Βασίζεται στην ελάχιστη απαιτούμενη συναίνεση διαλειτουργικότητας για την αναπαράσταση των δεδομένων στον ιστό, χρησιμοποιώντας Uniform Resource Identifier (URI)[a] και Resource Description Framework (RDF)[b] και επιτρέποντας την πρόσβαση σε δεδομένα μέσω HTTP. Τα URI χρησιμοποιούνται για την ανταλλαγή πληροφοριών μεταξύ υπολογιστών για την υποστήριξη της δυνατότητας ανάγνωσης της μηχανής σε αντίθεση με την προβολή ιστοσελίδων στον άνθρωπο. Αυτό με τη σειρά του επιτρέπει την σύνδεση των δεδομένων από διαφορετικές πηγές και την αναζήτηση των δεδομένων αυτών.

Ένα από τα πιο διάσημα έργα των Συνδεδεμένων Δεδομένων είναι το έργο DBpedia[c], το οποίο δημοσιεύει τις πληροφορίες που εξάγονται από το Wikipedia[d] στον Ιστό σε RDF μορφή.

Ο T. Berners-Lee εισήγαγε το 2006[e] τέσσερις



a. http://www.w3.org/TR/uri-clarification
b. http://www.w3.org/RDF
c. http://dbpedia.org
d. http://www.wikipedia.org
e. http://www.w3.org/DesignIssues/LinkedData.html

αρχές των Συνδεδεμένων Δεδομένων. Οι αρχές αυτές δεν ταυτίζουν τα Συνδεδεμένα Δεδομένα με τα ανοικτά δεδομένα. Η σχέση μεταξύ των Συνδεδεμένων Δεδομένων και των ανοικτών δεδομένων εισήχθη από το έργο των Ανοικτών Συνδεδεμένων Δεδομένων του W3C και αργότερα ενισχύθηκε από το έργο LOD2[f] της Ευρωπαϊκής Ένωσης. Έτσι, τα Συνδεδεμένα Δεδομένα θα μπορούσαν να είναι ανοιχτά (π.χ. δημόσια δεδομένα) ή κλειστά (π.χ. επιχειρηματικά δεδομένα).

Τα πιο σημαντικά πλεονεκτήματα που προσφέρει η τεχνολογία των Συνδεδεμένων Δεδομένων είναι η δυνατότητα συγκέντρωσης των δεδομένων που βρίσκονται σε διαφορετικά μέρη χωρίς την αποθήκευση τους και μια πιο άμεση και ικανή ανάλυση των δεδομένων αυτών σε σύγκριση με τις παραδοσιακές μεθόδους.

2.2. Ανασκόπηση

Η βάση για την τεχνολογία των Συνδεδεμένων Δεδομένων καθιερώθηκε από το W3C, δημοσιεύοντας την RDF σύσταση το 2004 και την RDF ερωτησιακή γλώσσα SPARQL, ως πρότυπο του W3C το 2008[g]. Το 2006, ο T. Berners-Lee δημοσίευσε τις αρχές των Συνδεδεμένων Δεδομένων για τα ανοιχτά δεδομένα ως τα «*5 Αστέρια των Συνδεδεμένων Ανοικτών Δεδομένων*[h]». Η ανάπτυξη των Συνδεδεμένων Δεδομένων συσχετίστηκε στενά με τα έργα των ανοικτών κυβερνητικών δεδομένων στις ΗΠΑ και το Ηνωμένο Βασίλειο το 2009. Αντίστοιχα, τα FP7 έργα της Ευρωπαϊκής Ένωσης LOD2 και LATC[i] χρηματοδοτήθηκαν το 2010.

Το W3C κατέβαλλε μεγάλη προσπάθεια για την ανάπτυξη των τεχνολογιών των Συνδεδεμένων Δεδομένων, δημοσιεύοντας βέλτιστες πρακτικές και οδηγίες για πλατφόρμες συνδεδεμένων δεδομένων [W3C (2013a)] και την σύσταση του SPARQL 1.1 [W3C (2013b)].

Επιπλέον, το W3C κατάφερε να τυποποιήσει ορισμένα σημαντικά λεξιλόγια (οντολογίες), όπως των Ανθρώπων (People)[j] και ενός οργανισμού (Organization)[k]. Η οντολογία των ανθρώπων χρησιμοποιεί την οντολογία Φίλος-ενός-Φίλου (Friend-of-a-Friend - FOAF)[l] για την περιγραφή της έννοιας ενός *ατόμου (Person)* ως μια οντότητα του τύπου *foaf: Person*. Η οντολογία των ανθρώπων περιλαμβάνει περισσότερους όρους από το λεξιλόγιο FOAF για την περιγραφή των ανθρώπων (π.χ. τις διευθύνσεις τους, τις συνδέσεις τους με οργανισμούς, κλπ).

Συμπερασματικά έχει δημιουργηθεί η τεχνολογική βάση για την τεχνολογία των Συνδεδεμένων Δεδομένων. Ωστόσο, υπάρχει έλλειψη εργαλείων που να υποστηρίζουν όλα τα στάδια του κύκλου ζωής των Συνδεδεμένων Δεδομένων, αποτρέποντας έτσι την ταχύτερη ανάπτυξη των Συνδεδεμένων Δεδομένων. Ένα άλλο ζήτημα είναι η απόδοση των εργαλείων αυτών στην περίπτωση των μεγάλης κλίμακας δεξαμενών δεδομένων.

Επίσης, ένα μικρό μόνο μέρος του συνόλου των Συνδεδεμένων Δεδομένων συνδέεται μέσω συνδέσμων και ένας από τους λόγους για αυτό, είναι η έλλειψη κοινώς αποδεκτών και ανοιχτών ταξονομιών (ή οντολογιών).

Τα πρότυπα του W3C όπως η οντολογία των ανθρώπων ή ενός οργανισμού μπορούν να χρησιμοποιηθούν, αλλά η χρήση τους περιορίζεται σε ορισμένα σύνολα δεδομένων και εφαρμογών. Η κατάσταση είναι καλύτερη σε συγκεκριμένα πληροφοριακά συστήματα, επειδή οι οντολογίες που απαιτούνται για την επιμέλεια των δεδομένων και την διασύνδεση τους αναπτύσσονται μαζί με το σύστημα. Επομένως, η μηχανική της οντολογίας [Gómez-Pérez et al. (2004)] ως πεδίο σημασιολογικών τεχνολογιών είναι πολύ σημαντική, προκειμένου να υποστηριχθεί η υλοποίηση των Συνδεδεμένων Δεδομέ-

f. http://lod2.eu
g. http://www.w3.org/TR/rdf-sparql-query
h. http://5stardata.info
i. http://latc-project.eu
j. http://www.w3.org/TR/vocab-people
k. http://www.w3.org/TR/vocab-org
l. http://xmlns.com/foaf/0.1

νων και η ανάπτυξη των εφαρμογών τους.

Από την άλλη πλευρά, είναι δύσκολο να αναπτυχθούν τυποποιημένες κοινές οντολογίες, εφόσον είναι πολύ επώδυνη ακόμη και η δημιουργία κοινώς αποδεκτών τομέων οντολογίας. Οι λόγοι είναι η πολυπλοκότητα του έργου, η γνώση του και η κατανάλωση χρόνου, καθώς και η απαιτούμενη δέσμευση για τη δημιουργία οντολογιών. Μία από τις λύσεις είναι η εκμάθηση των οντολογιών από μια μηχανή [Maedche and Staab (2001), Haav (2006)].

Οι εξελίξεις των τεχνολογιών των Συνδεδεμένων Δεδομένων σχετίζονται με την υιοθέτηση της σημασιολογικής τεχνολογίας, την έκρηξη των Μέγα Δεδομένων και των ανοιχτών δεδομένων. Επιπλέον, το IoT και η ανάλυση των μέσων κοινωνικής δικτύωσης επηρεάζουν τις τάσεις ανάπτυξης των Συνδεδεμένων Δεδομένων.

Σύμφωνα με τη μελέτη του Gartneri «*Ο Κύκλος Προώθησης των Μέγα Δεδομένων*» [Gartner (2012a)], οι σημασιολογικές τεχνολογίες πρόκειται να φτάσουν στο υψηλότερο επίπεδο της «*Κορυφής των Υψηλών Προσδοκιών*» σε περισσότερα από 10 χρόνια.

Η κορυφαία πρόβλεψη της Gartner για το 2014 [Gartner (2013c)] έφερε στο προσκήνιο το IoT ως γέφυρα μεταξύ μηχανών και ανθρώπων και νευροϋπολογιστικής (βλέπε επίσης Ενότητα 4). Η νευροϋπολογιστική μπορεί να γίνει μια από τις επεκτάσεις ή τις αντικαταστάσεις των τεχνολογιών των Συνδεδεμένων Δεδομένων.

2.3. Τάσεις

Τα Συνδεδεμένα Δεδομένα από τη φύση τους είναι ανοιχτά δεδομένα και επομένως τα ανοιχτά δημόσια δεδομένα αποτελούν μια καλή τράπεζα δοκιμών για τις τεχνολογίες των Συνδεδεμένων Δεδομένων. Αυτός είναι ένας από τους ενθαρρυντικούς παράγοντες των πρωτοβουλιών των συνδεδεμένων ανοιχτών δεδομένων. Η επιτυχής ανάπτυξη των Συνδεδεμένων Δεδομένων σχετίζεται με το λεγόμενο πρόβλημα της ενεργοποίησης (π.χ. η εκθετική ανάπτυξη των Συνδεδεμένων Δεδομένων απαιτεί την ύπαρξη μιας κρίσιμης ποσότητας Συνδεδεμένων Δεδομένων). Οι συλλογικές ενέργειες για τα συνδεδεμένα ανοιχτά δεδομένα βοηθούν στη δημιουργία αυτής της αρχικής ποσότητας δεδομένων.

Η τεχνολογία των Συνδεδεμένων Δεδομένων χρησιμοποιείται κυρίως για την ολοκλήρωση των δεδομένων, καθόσον η σύνδεση των δεδομένων που βασίζεται σε ένα συνεπές λεξιλόγιο (οντολογία) μειώνει το κόστος ολοκλήρωσης των δεδομένων και δημιουργεί ευκαιρίες για την ανάπτυξη εφαρμογών.

Στον ιδιωτικό τομέα, οι εταιρείες χρησιμοποιούν την τεχνολογία αυτήν για τη σύνδεση των κλειστών συνδεδεμένων δεδομένων τους με τα (συνδεδεμένα) ανοικτά δεδομένα ή τα αποτελέσματα εξόρυξης των ανοικτών δεδομένων. Για παράδειγμα, η Fujitsu Europe συνδυάζει τα ανοιχτά και κλειστά δεδομένα για τη δημιουργία μιας εφαρμογής υγειονομικής περίθαλψης, χρησιμοποιώντας αισθητήρες. Η τεχνολογία των Συνδεδεμένων Δεδομένων χρησιμοποιείται εντατικά για την ανάπτυξη σημασιολογικών συστημάτων διαχείρισης περιεχομένου και εφαρμογών πολυμέσων. Για παράδειγμα, το BBC είναι μία από τις πρώτες εταιρείες που άρχισαν να χρησιμοποιούν την τεχνολογία των Συνδεδεμένων Δεδομένων με την κατασκευή υπηρεσίας δεδομένων για τους Ολυμπιακούς Αγώνες του 2012[m]. Η Garlik από τον χρηματοπιστωτικό τομέα χρησιμοποιεί εκτεταμένα κλειστά και ιδιαιτέρως ασφαλή συνδεδεμένα δεδομένα. Το Sindicetech βοηθά τις εταιρείες να δημιουργήσουν κλειστά νέφη συνδεδεμένων δεδομένων. Οι εταιρείες αυτές έχουν πελάτες, όπως η εκδοτική εταιρεία Elsivier, η φαρμακευτική εταιρεία Astra-Zeneca[n], κλπ.

Αν και πολλές επιχειρήσεις του ιδιωτικού τομέα σε όλο τον κόσμο υιοθέτησαν την τεχνο-

m. http://www.bbc.co.uk/blogs/internet/posts/olympic_data_xml_latency

n. http://semanticweb.com/sindicetech-helps-enterprises-build-private-linked-data-clouds_b30454

λογία των συνδεδεμένων δεδομένων ως μια αποτελεσματική υποδομή ολοκλήρωσης δεδομένων, υπήρξαν και οι απαισιόδοξες απόψεις, και εξαιτείας αυτών ορισμένες εταιρείες πληροφορικής (ICT) δεν υιοθέτησαν την τεχνολογία των Συνδεδεμένων Δεδομένων, καθώς και των Μέγα Δεδομένων[o]. Οι μεγάλοι προμηθευτές λογισμικού, όπως η IBM, η Microsoft, κλπ. δεν είχαν ενδιαφερθεί επαρκώς για τα Συνδεδεμένα Δεδομένα. Τα Συνδεδεμένα Δεδομένα αποτελούν μια συγκεκριμένη τεχνολογική πλατφόρμα για αυτούς και δεν προέβλεπαν κερδοφόρα αποτελέσματα συνδυασμένα με τα δικά τους προϊόντα. Ωστόσο, η Oracle παρέχει λύσεις RDF βάσης δεδομένων. Ταυτόχρονα, η IBM και άλλοι έχουν ενσωματώσει εργαλεία Μέγα Δεδομένων ανοιχτού κώδικα, όπως το Hadoop και το MapReduce με τα δικά τους προϊόντα.

Από την άλλη πλευρά, έχουν εμφανιστεί εξειδικευμένες εταιρείες που προσφέρουν λύσεις με βάση την τεχνολογία των Συνδεδεμένων Δεδομένων, όπως το MarkLogic, το OpenLinks Software, το Franz, κλπ.

Στο δημόσιο τομέα, η τεχνολογία των Συνδεδεμένων Δεδομένων χρησιμοποιείται για τη δημιουργία συστημάτων πληροφοριών που ενσωματώνουν δεδομένα από κλειστές, ανοιχτές ή και από τις δύο πηγές. Τα κλειστά Συνδεδεμένα Δεδομένα χρησιμοποιούνται σε τομείς εθνικής ασφάλειας και άμυνας. Για παράδειγμα, οι ΗΠΑ αναδόμησαν την πύλη[p] υγειονομικής περίθαλψης που δημοσιεύει εν μέρει ανοιχτά δεδομένα, σύμφωνα με τις αρχές της τεχνολογίας των Συνδεδεμένων Δεδομένων. Η πύλη στατιστικών δεδομένων "The Data Web"[q] χρησιμοποιεί, μεταξύ των άλλων τεχνολογιών, και τις αρχές των Συνδεδεμένων Δεδομένων για την δημοσίευση και την ενοποίηση τους. Υπάρχουν διαθέσιμα ανοιχτά σύνολα δεδομένων για επαναχρησιμοποίηση από εταιρείες, όπως το νέ-

φος Συνδεδεμένων Δεδομένων της OpenCorporates[r] που περιλαμβάνει δεδομένα για πάνω από 50.000 επιχειρήσεις σε όλο τον κόσμο. Ένα άλλο παράδειγμα είναι η Product Ontology[s] που ταξινομεί και περιλαμβάνει πληροφορίες για περισσότερα από 1 εκατομμύρια προϊόντα. Επίσης, πολλές δημόσιες πύλες ανοιχτών δεδομένων δημοσιεύουν Συνδεδεμένα Ανοικτά Δεδομένα (π.χ. σε ΗΠΑ, Ηνωμένο Βασίλειο, Ευρωπαϊκή Ένωση, κλπ.).

ΤΕΧΝΟΛΟΓΙΕΣ ΜΕΓΑ ΔΕΔΟΜΕΝΩΝ

3.1. Έννοια

Τα Μέγα Δεδομένα χαρακτηρίζονται από τις ακόλουθες ιδιότητες: όγκος, ποικιλία, μεταβλητότητα, ταχύτητα, πολυπλοκότητα και εγκυρότητα (Veracity). Γενικά, τα δεδομένα θεωρούνται ως Μέγα Δεδομένα, όταν ο συνολικός τους όγκος είναι πολύ μεγάλος (π.χ. μετριέται σε petabytes) και τα δεδομένα είναι τόσο περίπλοκα που οι παραδοσιακές τεχνικές βάσεων δεδομένων δεν είναι αποτελεσματικά εφαρμόσιμες για την επεξεργασία των δεδομένων αυτών. Η ποικιλία τους αναφέρεται σε διαφορετικές μορφές των Μέγα Δεδομένων, για παράδειγμα, δομημένης και μη δομημένης μορφής. Η μεταβλητότητα χαρακτηρίζει τα Μέγα Δεδομένα από την άποψη της ανάλυσης τους, υποδεικνύοντας ποικίλες επιλογές για την ανάλυση και την ερμηνεία των αποτελεσμάτων της

o. http://blog.semantic-web.at/2013/06/04/theres-money-in-linked-data
p. www.healthcare.gov
q. http://thedataweb.rm.census.gov/TDW.html
r. http://opencorporates.com
s. http://www.productontology.org

ανάλυσης. Η ταχύτητα περιγράφει το πόσο γρήγορα δημιουργούνται και παραδίδονται τα δεδομένα για σκοπούς ανάλυσης. Συνήθως, η παράδοση σε πραγματικό χρόνο ή σχεδόν σε πραγματικό χρόνο λαμβάνεται υπόψη στην περίπτωση των Μέγα Δεδομένων. Η πολυπλοκότητα είναι ένα σημαντικό χαρακτηριστικό των Μέγα Δεδομένων, εφόσον τα σύνολα των δεδομένων μπορεί να είναι πολύ πολύπλοκα συνδέοντας δεδομένα από πολλές διαφορετικές πηγές. Η εγκυρότητα αναφέρεται σε συγκεκριμένα χαρακτηριστικά των δεδομένων, όπως η αξιοπιστία, η έλλειψη σφαλμάτων, η συνέπεια κλπ.

Οι τεχνολογίες των Μέγα Δεδομένων μπορούν να διαχωριστούν σε δύο ομάδες:

- Τις τεχνολογίες επιμέλειας (Curation) και αποθήκευσης των Μέγα Δεδομένων
- Την ανάλυση των Μέγα Δεδομένων, η οποία χρησιμοποιείται για την εύρεση κρυφών μοτίβων και συσχετισμών από δεδομένα που συλλέχθηκαν προκειμένου να χρησιμοποιηθεί η αποκτηθείσα γνώση για την καλύτερη λήψη αποφάσεων.

Οι τεχνολογίες αυτές περιλαμβάνουν την υψηλής απόδοσης εξόρυξη δεδομένων, την προγνωστική ανάλυση, την εξόρυξη κειμένων, την πρόβλεψη και την βελτιστοποίηση.

3.2. Ανασκόπηση

Το 2010, η Gartner [Gartner (2010)] προέβλεπε ότι ο όγκος των εταιρικών δεδομένων οποιασδήποτε μορφής θα αυξηθεί περίπου 650% τα επόμενα 10 χρόνια. Η αύξηση αυτή του όγκου των δεδομένων αποτέλεσε πρόκληση για τις τεχνολογίες επεξεργασίας δεδομένων.

Σε σύγκριση με τις σχεσιακές βάσεις δεδομένων, τα Μέγα Δεδομένα δεν είναι τόσο καλά και ομοιόμορφα δομημένα. Τα Μέγα Δεδομένα δημιουργούνται από μηχανές (π.χ. αισθητήρες) ή από άτομα με πολύ διαφορετικό υπόβαθρο και εμπειρία (π.χ. μέσω των μέσων κοινωνικής δικτύωσης). Αυτό θέτει νέες απαιτήσεις για τις τεχνολογίες επεξεργασίας των Μέγα Δεδομένων, τα αναλυτικά εργαλεία, τους αλγόριθμους εξόρυξης δεδομένων και τις τεχνικές οπτικοποίησης δεδομένων.

Σήμερα, τα Μέγα Δεδομένα υπάρχουν σχεδόν σε όλους τους τομείς (π.χ. στον τραπεζικό, στην υγειονομική περίθαλψη, στην ασφάλιση, στον κατασκευαστικό, στο μάρκετινγκ, στις μεταφορές, κλπ.). Αυτό με τη σειρά του δημιουργεί ζήτηση στην αγορά για τεχνολογίες επεξεργασίας και ανάλυσης Μέγα Δεδομένων που έχουν θετική επίδραση στην ανάπτυξη του αντίστοιχου τεχνολογικού τομέα. Μία από τις κύριες προκλήσεις της τεχνολογίας της πληροφορίας από αυτή την άποψη ήταν η διαχείριση του όγκου δεδομένων και η δημιουργία αντίστοιχων επεκτάσιμων αρχιτεκτονικών επεξεργασίας δεδομένων, εφόσον οι παραδοσιακές τεχνολογίες βάσεων δεδομένων δεν ήταν πλέον κατάλληλες για το χειρισμό αυτού του νέου όγκου δεδομένων. Ένα από τα άλυτα προβλήματα που σχετίζονται με τα Μέγα Δεδομένα είναι η σημασιολογική διαλειτουργικότητα, καθώς αυξάνεται η ποικιλία και η μεταβλητότητα των δεδομένων. Η σημασιολογική διαλειτουργικότητα των Μέγα Δεδομένων αποτελεί ένα από τα πιο σημαντικά προβλήματα της αποτελεσματικής διαχείρισης των Μέγα Δεδομένων [Haav and Küngas (2014)]. Τα Μέγα Δεδομένα είναι ως επί το πλείστον μη δομημένα δεδομένα. Επομένως, η διαθεσιμότητα (σημασιολογικών) μεταδεδομένων είναι ζωτικής σημασίας. Οι τεχνολογίες των Συνδεδεμένων Δεδομένων θα μπορούσαν να βοηθήσουν στη σύνδεση διαφορετικών συνόλων Μέγα Δεδομένων και στην παροχή ευρύτερης θέασης των δεδομένων αυτών.

Από την άλλη πλευρά, η αναλυτική ικανότητα των Μέγα Δεδομένων είναι πολύ σημαντική, καθώς δίνει επιχειρηματική αξία στα Μέγα Δεδομένα εξορύσσοντας βαθύτερη γνώση από τα υφιστάμενα δεδομένα. Ο T. Pellegrin [Pellegrin (2014)] εξετάζει τη μακροοικονομική και μικροοικονομική αξία των Μέγα Δεδομένων. Υποστηρίζει ότι σε μακροοικονομικό επίπεδο τα Μέγα Δεδομένα είναι μάλλον πολιτικά παρά τεχνολογικά θέματα παρόμοια με τα ανοιχτά

δεδομένα που ξεκίνησαν ως αποτέλεσμα της πολιτικής των ΗΠΑ το 2009. Μετά από αυτό, η πολιτική των ανοιχτών δεδομένων μεταφέρθηκε στην Ευρωπαϊκή Ένωση, όπου έγινε το θεμέλιο της Ευρωπαϊκής Ψηφιακής Ατζέντας. Το πολιτικό κίνητρο των ανοιχτών δεδομένων ήταν να τονώσει την οικονομία, ενώ ο ρόλος των Μέγα Δεδομένων είναι να ενθαρρύνει τη διαδικασία λήψης αποφάσεων με βάση την ανάλυση των δεδομένων και τη χρήση μη δομημένων δεδομένων Ιστού για εθνικές κρίσιμες εφαρμογές. Ως εκ τούτου, οι ΗΠΑ ανακοίνωσαν την "Πρωτοβουλία των Μέγα Δεδομένων" το 2012 προκειμένου να εκφράσουν το ενδιαφέρον τους για έρευνα, ανάπτυξη και εφαρμογή της τεχνολογίας των Μέγα Δεδομένων [White-house (2012)]. Οι αντίστοιχες πρωτοβουλίες της Ευρωπαϊκής Ένωσης περιλαμβάνουν κλήσεις και έργα στα προγράμματα πλαίσιο της Ευρωπαϊκής Ένωσης (FP7 και Horizon 2020) που σχετίζονται με την τεχνολογία των Μέγα Δεδομένων και επίσης ευρύτερα φόρουμ, όπως το BIG Project[t].

Οι βασικές τεχνολογίες ανοιχτού κώδικα για την επεξεργασία των Μέγα Δεδομένων, όπως η Hadoop[u], η MapReduce[v] και οι συναφείς εφαρμογές, όπως η Cloudera[w] και η Hive[x] και η υιοθέτηση αυτών από μεγάλες εταιρείες λογισμικού, όπως η Oracle και η IBM συνέβαλαν στην υιοθέτηση των τεχνολογιών των Μέγα Δεδομένων.

Βασική τεχνολογία για την ανάλυση των Μέγα

t. http://www.big-project.eu
u. http://hadoop.apache.org
v. http://research.google.com/archive/mapreduce.html
w. http://www.cloudera.com
x. http://hive.apache.org

Δεδομένων είναι η παράλληλη ή η NoSQL βάση δεδομένων που εκτείνεται με μια Hadoop σύνδεση. Η Hadoop χρησιμοποιείται για την επεξεργασία αδόμητων Μέγα Δεδομένων. Για παράδειγμα, η Amazon χρησιμοποιεί την Hadoop. Η Hadoop χρησιμοποιείται ευρέως λόγω της κλιμακοσιμότητας της, αναφορικά τόσο με τον αυξανόμενο όγκο δεδομένων, όσο και το επεκτεινόμενο δίκτυο με νέους κόμβους. Το τελευταίο δημιουργεί καλές ευκαιρίες για την ανάπτυξη εφαρμογών που ικανοποιούν τις ανάγκες των εταιρειών. Μια άλλη σημαντική πτυχή της Hadoop είναι ότι αποτελεί λογισμικό ανοιχτού κώδικα.

Σύμφωνα με την έρευνα της Gartner [Gartner (2013b)] μεταξύ 720 εταιρειών, το 64% των οργανισμών σχεδίαζε έργα Μέγα Δεδομένων για το 2013. Τα μέσα μαζικής ενημέρωσης, οι εταιρείες επικοινωνίας και οι τράπεζες είχαν κατακτήσει την πρώτη θέση της λίστας. Πάνω από το ένα τρίτο των εταιρειών μέσων μαζικής ενημέρωσης και επικοινωνιών δήλωσαν ότι έχουν ήδη επενδύσει σε έργα ανάλυσης των Μέγα Δεδομένων. Ως εκ τούτου, η Gartner αποκάλεσε το έτος 2013, ως το έτος πειραματισμού και αρχής υιοθέτησης των Μέγα Δεδομένων.

Το UK Public Sector 2020 Databerg Report επικεντρώνεται στον τρόπο με τον οποίο ο δημόσιος τομέας χειρίζεται τα δεδομένα και κατά πόσον τα έχει μετατρέψει σε χρήσιμη μορφή.

Το 30% των δεδομένων που αποθηκεύονται από τους οργανισμούς του δημόσιου τομέα θεωρείται καθαρό και έχει γνωστή αξία - με άλλα λόγια, γνωρίζουν τι περιέχει και είναι χρήσιμο. Και ενώ δεν φαίνεται να είναι ιδιαίτερα μεγάλο ποσοστό, είναι συγκριτικά μεγαλύτερο του αντίστοιχου 15% των εταιρειών του ιδιωτικού τομέα. Το ποσοστό επίσης των περιττών, ξεπερασμένων και ασήμαντων (Redudant, Obsolete, Trivial - ROT) δεδομένων - ουσιαστικά σκουπίδια - είναι επίσης χαμηλότερο σε σύγκριση με του ιδιωτικού τομέα, αντιπροσωπεύοντας το 20% όλων των δεδομένων του δημόσιου τομέα, σε σύγκριση με το 35% των ιδιωτικών εταιρειών.

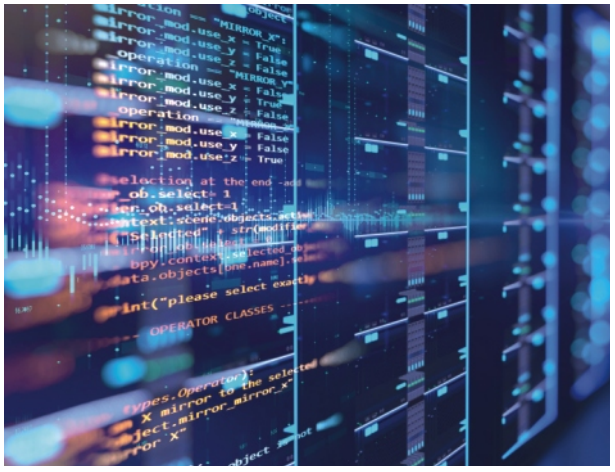
3.3. Τάσεις

Η υιοθέτηση των τεχνολογιών των Μέγα Δεδομένων και το οικονομικό της όφελος στις επιχειρήσεις και στο δημόσιο τομέα αυξάνεται ανάλογα με το ρυθμό της ικανότητας χρήσης αυτών των τεχνολογιών για την επίλυση σύνθετων προβλημάτων. Σε σχέση με αυτό, το εμπόδιο δεν είναι η τεχνολογία, αλλά η αναλυτική ικανότητα του προσωπικού επεξεργασίας δεδομένων. Η κατάσταση αυτή δημιουργεί την ανάγκη για αναλυτές δεδομένων ή επιστήμονες δεδομένων στην αγορά εργασίας.

Η ύπαρξη υπεύθυνου δεδομένων (Chief Data Officer - CDO) στις επιχειρήσεις είναι ανάγκη που επιβάλλει η ευρύτερη υιοθέτηση των τεχνολογιών δεδομένων σε όλους τους τομείς. Δεδομένου ότι υπάρχει τεράστια ζήτηση για επιστήμονες δεδομένων, ορισμένοι οργανισμοί στρέφονται στον αυτοματισμό για να μειώσουν το κόστος, με αποτέλεσμα η βοήθεια της τεχνητής νοημοσύνης (Artificial Intelligence - AI), να επιτρέπει στους επιστήμονες αυτούς να προχωρήσουν σε πιο προηγμένες διεργασίες, ώστε να είναι πιο παραγωγικοί προσθέτοντας περισσότερη αξία στην εργασία τους. Περιζήτητοι μάλιστα είναι οι επιστήμονες δεδομένων που είναι ειδικευμένοι στη μηχανική μάθηση, τις βάσεις δεδομένων, την προετοιμασία δεδομένων, την κωδικοποίηση, τις στατιστικές, την οπτικοποίηση δεδομένων κ.α. Στο μέλλον βέβαια οι ευθύνες θα χωριστούν σε πιο εξειδικευμένους ρόλους, καθώς είναι ανέφικτο ένα άτομο να καλύψει όλες τις ανάγκες μιας εταιρείας.

Οι περιοχές χρήσης της ανάλυσης των Μέγα Δεδομένων περιλαμβάνουν τον προγραμματισμό των επιχειρηματικών πόρων, τις δραστηριότητες έρευνας και ανάπτυξης, το μάρκετινγκ και την διαχείριση κινδύνων. Πολλές εταιρείες μέσων κοινωνικής δικτύωσης, όπως η Google, το Twitter, το Facebook, το LinkedIN, κλπ., εφαρμόζουν τη διαχείριση και ανάλυση των Μέγα Δεδομένων για τη βελτίωση της πελατειακής τους βάσης.

Ένας αριθμός προμηθευτών λογισμικού διαχεί-



ρισης βάσεων δεδομένων και αποθήκης δεδομένων (π.χ. η IBM, η Oracle, η Teradata, η SAP EMC, η HP, η Amazon, η MS, η Google κλπ.) έχουν ενσωματώσει λειτουργίες επεξεργασίας Μέγα Δεδομένων, ουσιαστικά ανοιχτού κώδικα Hadoop και MapReduce, στα προϊόντα τους.

Από την άλλη πλευρά, υπάρχουν εταιρείες, όπως η HaDapt, η Platfora, η YarcData, η SiSense, η Space-Time Insights, η Zettaset, κλπ., που προσανατολίζονται ειδικά στις τεχνολογίες και στην ανάλυση των Μέγα Δεδομένων.

Η IBM Research έχει δημιουργήσει το Accelerated Discovery Lab [IBM Research (2013)] για την έρευνα και ανάπτυξη των Μέγα Δεδομένων στο Σαν Χοσέ. Το εργαστήριο διαχειρίζεται έργα Μέγα Δεδομένων σε τομείς όπως η υγειονομική περίθαλψη, η φαρμακευτική, τα συστήματα γεωπληροφορικής, η διαχείριση υδατικών πόρων, τα Μέσα Κοινωνικής Δικτύωσης, κλπ.

Σύμφωνα με τη μελέτη του Ινστιτούτου SAS, "Τα Μέγα Δεδομένα σε Μέγα Εταιρείες", είναι σαφές ότι οι μεγάλες βιομηχανικές εταιρείες είχαν αρχίσει να χρησιμοποιούν την επεξεργασία και ανάλυση των Μέγα Δεδομένων, ακόμη και τότε που είχαν ανάγκη από επιπλέον χρόνο, νέες τεχνολογικές προσεγγίσεις, νέες οργανωτικές και διαχειριστικές δομές και νέες δεξιότητες. Για παράδειγμα, μεγάλες εταιρείες, όπως η UPS, η General Electric, κλπ., χρησιμοποίησαν τις τεχνολογίες των Μέγα Δεδομένων για την βελτιστοποίηση της διαδικασίας παραγωγής τους, συλλέγοντας δεδομένα με χρήση αισθητήρων.

Η Κ.Υ.Π. των ΗΠΑ (CIA) ενδιαφέρεται για τις τεχνολογίες των Μέγα Δεδομένων για τη συλλογή και ανάλυση δεδομένων πληροφοριών. Η αυτόματη επιμέλεια των δεδομένων και η ανάλυση τους παρέχουν ευκαιρίες εύρεσης κρυφών μοτίβων σε δεδομένα και σχέσεις μεταξύ δεδομένων και συμβάντων.

Σύμφωνα με την International Data Corporation, το 90% των μη δομημένων δεδομένων δεν αναλύεται ποτέ. Αυτά τα δεδομένα είναι γνωστά ως Σκοτεινά (Dark) δεδομένα. Τα σκοτεινά δεδομένα είναι ένα υποσύνολο των Μέγα δεδομένων, το οποίο αποτελεί το μεγαλύτερο μέρος του συνολικού όγκου των Μέγα δεδομένων που συλλέγονται από οργανισμούς σε ένα χρόνο. Τα σκοτεινά δεδομένα δεν αναλύονται ή επεξεργάζονται από τις εταιρείες συνήθως για διαφορετικούς λόγους.

Με τη βιομηχανία όμως να συνειδητοποιεί τελικά τη σημασία της εκπληκτικής ποσότητας πρωτογενών δεδομένων στις ανεξερεύνητες περιοχές του Βαθύ Ιστού (Deep Web), γίνονται τεράστια στοχοποιημένα βήματα για την συγκέντρωση και αποκρυπτογράφηση των διαθέσιμων μη δομημένων δεδομένων. Η αξιοποίηση αυτών των δεδομένων με την χρήση της AI θα βοηθήσει τους οργανισμούς να αποκτήσουν νέες γνώσεις και γνώσεις που θα αποφέρουν ένα μεγαλύτερο ανταγωνιστικό πλεονέκτημα. Η IDC προβάλλει ότι οι οργανισμοί που αναλύουν όλα τα σχετικά δεδομένα και παρέχουν ενημερωμένες πληροφορίες θα επιτύχουν επιπλέον κέρδη 430 δισεκατομμυρίων δολαρίων σε σχέση με τους λιγότερο προσανατολισμένους οργανισμούς προς την κατεύθυνση αυτή έως το 2020. Στο πλαίσιο των επιχειρηματικών δεδομένων, ο όρος "Σκοτεινό (Dark)" περιγράφει κάτι που είναι κρυμμένο ή αδιαπέραστο. Η ανάλυση του σκοτεινού ιστού επικεντρώνεται κυρίως σε ακατέργαστα δεδομένα που βασίζονται σε κείμενα που δεν έχουν αναλυθεί - με

έμφαση στα μη δομημένα δεδομένα, τα οποία μπορεί να περιλαμβάνουν πράγματα όπως μηνύματα κειμένου, έγγραφα, email, αρχεία βίντεο και ήχου και φωτογραφίες. Σε ορισμένες περιπτώσεις, οι εξερευνήσεις σκοτεινών αναλυτικών στοιχείων θα μπορούσαν επίσης να στοχεύουν στο Βαθύ Ιστό, ο οποίος περιλαμβάνει όλα τα διαδικτυακά που δεν ευρετηριάζονται από μηχανές αναζήτησης, συμπεριλαμβανομένου ενός μικρού υποσυνόλου ανώνυμων, απρόσιτων ιστότοπων που είναι γνωστοί ως "Σκοτεινός Ιστός".

Υπολογίζεται ότι το 90 τοις εκατό όλων των δεδομένων που υπάρχουν σήμερα δημιουργήθηκαν τα τελευταία πέντε χρόνια.

Είναι απολύτως σαφές ότι η υιοθέτηση των τεχνολογιών των Μέγα Δεδομένων επιτρέπει την αύξηση της οικονομικής ανάπτυξης. Ωστόσο, πρέπει να ληφθούν υπόψη και οι κοινωνικές πτυχές των Μέγα Δεδομένων, που είναι η ασφάλεια, η προστασία της ιδιωτικής ζωής και τα πνευματικά δικαιώματα.

Στο επόμενο τεύχος:

4. ΔΙΑΔΙΚΤΥΟ ΤΩΝ ΠΡΑΓΜΑΤΩΝ

5. ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

6. ΔΙΑΣΥΝΔΕΣΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕΣΩ ΣΥΝΔΕΔΕΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Λίγα λόγια για τον αρθρογράφο

Ο κ. **Δημήτρης Φιλίππου** είναι Τεχνικός Διευθυντής στην Εταιρεία Integrated Intelligent Quality Systems (I2QS). Σε διεθνές επίπεδο αποτελεί μέλος του Communication Society του INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS (IEEE) από το 1995 και Senior Member του IEEE από το 2020. Είναι Country Chair του BICSI στην Ελλάδα και μέλος του από το 2001. Επιπλέον, αποτελεί μέλος ως Professional Engineer στο CONSTRUCTION SPECIFICATIONS INSTITUTE (CSI) από το 2018, καθώς επίσης είναι μέλος ως Consultant στο ETHERNET ALLIANCE (EA) από το 2019.

Εργάζεται στον χώρο της πληροφορικής και των τηλεπικοινωνιών από το 1995, ξεκινώντας ως μηχανικός έρευνας, σχεδίασης και ανάπτυξης τηλεπικοινωνιακών συστημάτων (Hardware Design). Στην συνέχεια ασχολήθηκε με την σχεδίαση και ανάπτυξη τηλεπικοινωνιακών δικτύων, ενώ ταυτόχρονα επικεντρώθηκε στην μελέτη και ανάπτυξη τηλεπικοινωνιακών καλωδιακών υποδομών, παρακολουθώντας ενεργά όλα αυτά τα χρόνια την ανάπτυξη και εξέλιξη των Ευρωπαϊκών και Διεθνών προτύπων.

Σε εθνικό επίπεδο, συμμετέχει ενεργά στην εθνική Τεχνική Επιτροπή ΤΕ93 του ΕΛΟΤ από το 2007, εκπροσωπώντας την παράλληλα στην ευρωπαϊκή Τεχνική Επιτροπή TC215 της CENELEC. Σε ευρωπαϊκό επίπεδο, συμμετέχει ως μέλος στο Working Group 1 (WG1) και στο Working Group 2 (WG2) της TC215 της CENELEC από το 2007. Τα Working Groups αυτά είναι αρμόδια για την ανάπτυξη μιας ολοκληρωμένης σειράς προτύπων για την σχεδίαση και την εγκατάσταση των τηλεπικοινωνιακών καλωδιακών υποδομών σε μια σειρά εγκαταστάσεων, συμπεριλαμβάνοντας μεταξύ άλλων, γραφεία, βιομηχανίες, σπίτια και κέντρα δεδομένων. Επιπλέον, συμμετέχει στις εργασίες του Working Group 3 (WG3) της CLC TC215 για την σειρά προτύπων EN 50600 από το 2007, εξετάζοντας την εφαρμογή της ενεργειακής απόδοσης στις εγκαταστάσεις και τις υποδομές ενός κέντρου δεδομένων. Σε διεθνές επίπεδο, συμμετέχει ενεργά στα Subcommittee Power over Ethernet (PoE), High Speed Networking (HSN) και Single Pair Ethernet (SPE) του EA. Έχει γράψει ένα πλήθος άρθρων σχετικά με τα διεθνή, ευρωπαϊκά και εθνικά πρότυπα που χρησιμοποιούνται στην σχεδίαση και ανάπτυξη συστημάτων γένιας καλωδίωσης και είναι ομιλητής σε αντίστοιχο πλήθος διαλέξεων και σεμιναρίων, ενημερώνοντας την ελληνική αγορά για τις εξελίξεις και την πρόοδο των προτύπων των τηλεπικοινωνιακών καλωδιακών υποδομών.

Εάν επιθυμείτε το COMMUNICATION SOLUTIONS να δημοσιεύσει περισσότερα άρθρα για **Information Technologies** επικοινωνήστε μαζί μας στο: **info@comsol.gr**