

Το νόημα των δεδομένων στην εποχή της 4^{ης} Βιομηχανικής επανάστασης



Δημήτρης Κυρίτσης

Μέντορας καριέρας • [LinkedIn](#)

Key ideas:

- 1) Ο άυλος φυσικός κόσμος γύρω μας μπορεί να παρουσιαστεί ως ένα σύνολο ακατέργαστων δεδομένων
- 2) Η «πληροφορία» ορίζεται επίσης συχνά ως «δεδομένα σε πλαίσιο»
- 3) Η σημασιολογία των δεδομένων είναι εξίσου σημαντική με τα ίδια τα ακατέργαστα δεδομένα

Ζούμε στην εποχή των δεδομένων, όπου όλα όσα μας περιβάλλουν συνδέονται με μια πηγή δεδομένων και πολλές πτυχές της ζωής μας ψηφιοποιούνται όλο και περισσότερο. Υποτίθεται ότι ο φυσικός κόσμος γύρω μας μπορεί να παρουσιαστεί ως ένα σύνολο ακατέργαστων δεδομένων. Με δεδομένο τον τεράστιο όγκο ακατέργαστων δεδομένων, οι τεχνολογίες επεξεργασίας μεγάλου όγκου δεδομένων (big data analysis) έχουν αναδειχθεί ως ένα αποτελεσματικό και απαραίτητο μέσο για τη διερεύνηση και αξιοποίηση τους.

Τα Big Data περιγράφονται ευρέως με τέσσερις κύριες διαστάσεις: όγκος, ταχύτητα, ποικιλία και ακρίβεια.

Ο **όγκος** αναφέρεται στο πρόβλημα της αντιμετώπισης πολύ μεγάλων συνόλων δεδομένων, το οποίο συνήθως απαιτεί εκτέλεση σε μια κατανεμημένη υποδομή που βασίζεται σε νέφος.

Η **ταχύτητα** αναφέρεται στην αντιμετώπιση δεδομένων ροής σε πραγματικό χρόνο, όπως ροές βίντεο, όπου μπορεί να είναι

αδύνατη η αποθήκευση όλων των δεδομένων για μεταγενέστερη επεξεργασία.

Η **ποικιλία** αναφέρεται στην αντιμετώπιση διαφορετικών τύπων πηγών, διαφορετικών μορφών δεδομένων και μεγάλου αριθμού πηγών.

Η **ακρίβεια** ισχύει για την επαληθευσσιμότητα των ακατέργαστων δεδομένων και την εμπιστευτικότητα των πληροφοριών που προκύπτουν από μια ανάλυση.

Μεγάλο μέρος της εργασίας στα μεγάλα δεδομένα έχει επικεντρωθεί στον όγκο και την ταχύτητα, αλλά τα προβλήματα της ποικιλίας και της ακρίβειας είναι εξίσου σημαντικά για την επίλυση πολλών προβλημάτων του πραγματικού κόσμου. Υπό το πρίσμα της ετερογένειας των δεδομένων που συλλέγονται από πολλές και διάφορες πηγές, μπορεί να υποστηριχθεί ότι **η έρευνα για να φτάσουμε στο νόημα των δεδομένων αποτελεί το Άγιο Δισκοπότηρο της ενοποίησης δεδομένων.**

Ας δείξουμε τη σημασία του μέσα από ένα παράδειγμα.

Σήμερα, μυριάδες δεδομένα συλλέγονται κάθε στιγμή από διάφορους τύπους αισθητήρων και η τάση αυξάνεται με τη διαθεσιμότητα μικρότερων, φθηνότερων και πιο αποτελεσματικών αισθητήρων που μπορούν να ανιχνεύουν σχεδόν τα πάντα.

Ας δούμε πιο προσεκτικά κάτι που κάθε άνθρωπος κάνει πολύ συχνά: να μετράει τη θερμοκρασία του σώματός μας με ένα απλό θερμόμετρο. Είναι απλό για όλους πώς να το κάνουν, πώς να διαβάσουν την τιμή της θερμοκρασίας του σώματος σε βαθμούς Κελσίου ή Φαρενάιτ και να ερμηνεύσουν την τιμή που εμφανίζεται στο θερμόμετρο. Για παράδειγμα, εάν δούμε μια τιμή 38,5 βαθμών Κελσίου, συμπεραίνουμε ότι ο πυρετός μας είναι κάπως υψηλός και ίσως χρειαστεί να καλέσουμε τον γιατρό μας ή να πάρουμε κάποιο φάρμακο. Τώρα ας σκεφτούμε αυτό: ποια είναι η πηγή των δεδομένων που διαβάζουμε στο θερμόμετρο; Το ίδιο το θερμόμετρο όπως πολλοί άνθρωποι απαντούν σε αυτήν την ερώτηση; Όμως το ίδιο το θερμόμετρο δεν είναι ικανό να δημιουργήσει θερμότητα ή να αυξήσει τη θερμοκρασία του σώματός μας. Είναι μάλλον ένα φαινόμενο στο σώμα μας που δημιουργεί θερμότητα και αυξάνει τη θερμοκρασία του η οποία με τη σειρά της συλλαμβάνεται από το θερμόμετρο αν το εφαρμόσουμε σωστά και στο σωστό σημείο του σώματός μας.

Ο παραπάνω συλλογισμός μπορεί να γίνει με οποιονδήποτε τύπο αισθητήρα που χρησιμοποιείται για να συλλάβει την τιμή για κάτι που θέλουμε να μετρήσουμε: θερμοκρασία, δόνηση, επιτάχυνση, κίνηση, κ.λπ. κ.λπ. Η πραγματική πηγή των δεδομένων που καταγράφονται δεν είναι ο ίδιος ο αισθητήρας αλλά ένα συγκεκριμένο φαινόμενο του αντικειμένου στο οποίο είναι ενσωματωμένος ο αισθητήρας. Η λειτουργία του αισθητήρα είναι να συλλαμβάνει την τιμή μιας παραμέτρου που επιτρέπει την αξιολόγηση της συμπεριφοράς του

παρατηρούμενου φαινομένου.

Επιπλέον των παραπάνω, πρέπει να σημειωθεί ότι τα δεδομένα που συλλέγονται ερμηνεύονται πάντα εντός των ορίων ενός καλά καθορισμένου «πλαισίου» (context).

Το προαναφερθέν παράδειγμα απεικονίζει τη σημασία της αναγνώρισης και σύλληψης του νοήματος των δεδομένων. Ως λύση σε αυτήν την πρόκληση, ένα σημασιολογικό μοντέλο δεδομένων προσπαθεί να ορίσει ρητά τι είναι έννοιες όπως, σε ένα άλλο ευνόητο παράδειγμα, «πελάτης», «προϊόν», «πιστωτικό όριο», «καθαρές πωλήσεις» και ούτω καθεξής. Αυτό που πρέπει να αντιμετωπίσει ο σημασιολογικός μελετητής είναι το «πλαίσιο» (context) της έννοιας στην οποία αναφέρεται κάποιο δεδομένο, και πώς σχετίζεται με άλλα δεδομένα σε σχέση με την πηγή κάθε δεδομένου. Για παράδειγμα, ένας πελάτης είναι φυσικό πρόσωπο ή εταιρεία; Πρέπει ένας πελάτης να έχει πράγματι αγοράσει ένα προϊόν ή μπορεί ένας πελάτης να είναι επίσης κάποιος που βρίσκεται στην αγορά και ψάχνει για ένα (το) προϊόν; Αυτό που σε ορισμένα πλαίσια μπορεί να ονομαστεί «δυνατότητα» μπορεί να ονομαστεί «πελάτης» σε άλλα. Είναι ο χονδρέμπορος πελάτης ή μόνο ο τελικός καταναλωτής είναι ο πελάτης; Ο πελάτης του χονδρέμπορου ονομάζεται επίσης πελάτης; Αυτή ακριβώς τη σύγχυση προσπαθεί να αποκαλύψει και να επιλύσει το σημασιολογικό μοντέλο. Με μια γενική έννοια, η σημασιολογία είναι η μελέτη των νοημάτων του μηνύματος πίσω από λέξεις, αριθμούς ή και άλλους τύπους δεδομένων. Ο σημασιολογικός μελετητής πρέπει να διερευνήσει και να συλλάβει την απόχρωση κάθε σημασίας και πρέπει να συνεργαστεί με τους τελικούς χρήστες ή και άλλους που εμπλέκονται σε σχετικές διαδικασίες για να αναπτύξει μια σύμβαση ονομασίας ή σύνταξη που να παρέχει σαφήνεια. Όλες οι πιθανές σημασίες πρέπει αντιπροσωπεύονται στο σημασιολογικό μοντέλο.

Δεν αποτελεί έκπληξη, όπως σημειώθηκε προηγουμένως, ότι η «πληροφορία» ορίζεται επίσης συχνά ως «δεδομένα σε πλαίσιο (data in context)». Για παράδειγμα, ένας αριθμός ως δεδομένο είναι κάτι το πολύ αφηρημένο και μόνο όταν προσθέτουμε νοηματικό περιεχόμενο και μετατρέπουμε τον αριθμό σε ποσότητα με συγκεκριμένο νόημα, τα δεδομένα μπορούν να μετατραπούν σε πολύτιμες πληροφορίες. Ωστόσο, είναι αρκετά σύνηθες ότι παρόλο που έχουμε άφθονα δεδομένα, το πληροφοριακό τους περιεχόμενο είναι ασαφές και δυστυχώς όχι τόσο κατατοπιστικό. Ο κύριος λόγος είναι ότι παρόλο ότι δίνουμε μεγάλη προσοχή στη διαδικασία συλλογής των δεδομένων, δίνουμε σημαντικά λιγότερη προσοχή στην τυποποίηση της σημασίας τους, καθιστώντας σαφές τι περιγράφουν αυτά τα δεδομένα. Για παράδειγμα, το «πιστωτικό όριο» αναφέρεται στο μέγιστο ποσό πίστωσης που θα επεκτείνει ένα χρηματοπιστωτικό ίδρυμα σε έναν οφειλέτη για ένα συγκεκριμένο πιστωτικό όριο. Ωστόσο, αυτός ο ορισμός μας λέει λίγα σχετικά με το νόμισμα στο οποίο μετράται το πιστωτικό όριο ή τον τύπο της αξίας του. Έτσι, θα μπορούσαμε να πούμε ότι οι συναλλασσόμενοι «χάνονται στη μετάφραση».

Στην πραγματικότητα, γίνεται όλο και πιο εμφανές ότι η σημασιολογία των δεδομένων είναι εξίσου σημαντική με τα ίδια τα ακατέργαστα δεδομένα. Προς αυτή την κατεύθυνση, οι σημασιολογικές τεχνολογίες προσφέρουν ένα πλήρες σύνολο εργαλείων για την ταυτόχρονη παροχή των ακατέργαστων δεδομένων και της σημασίας τους, προσφέροντας έναν μοναδικό τρόπο για τη δημιουργία ενός πλαισίου για την έκθεση σε πηγές δεδομένων. Ταυτόχρονα, παρέχουν τα εργαλεία που απαιτούνται για τον επίσημο ορισμό των όρων και των σχέσεων που χαρακτηρίζουν έναν δεδομένο τομέα, παρέχοντας έτσι ένα τυποποιημένο και αναγνώσιμο από ηλεκτρονικό υπολογιστή αναπαραστατικό μοντέλο (οντολογία)

που μπορεί να χρησιμοποιηθεί σε μεταγενέστερο βήμα για να φτάσουμε στο νόημα των δεδομένων.