Παρουσίαση Big Data

Λάμπρου Ανδρέας
Επιβλέπων καθηγητής: Δρ. Μηνάς Δασυγένης
Πανεπιστήμιο Δυτικής Μακεδονίας
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εργαστήριο Ψηφιακών Συστημάτων και Αρχιτεκτονικής Υπολογιστών,
http://arch.icte.uowm.gr/ Κοζάνη 2019

Πανεπιστήμιο Δυτικής Μακεδονίας

Περιεχόμενα παρουσίασης

- Τι είναι τα Big data
- Τα 3 v των Big data
- Ιστορία του Big data
- Καθημερινές εφαρμογές του Big data
- Γιατί είναι σημαντικά
- Ποιος τα χρησιμοποιεί
- Χτίζοντας την αρχιτεκτονική των Big data

Πανεπιστήμιο Δυτικής Μακεδονίας

Περιεχόμενα παρουσίασης

- Δομημένα δεδομένα
- Μη δομημένα δεδομένα
- Ημι δομημένα δεδομένα
- Συστήματα διαχείρισης περιεχομένου
- Διαχείριση διαφορετικών τύπων δεδομένων
- Hadoop
- Mapreduce
- Hive
- Hdfs
- Nosql

Πανεπιστήμιο Δυτικής Μακεδονίας

Περιεχόμενα παρουσίασης

- Pig
- Scoop
- Pig latin
- κλασικές μεθόδους εξόρυξης
- k clustering
- decision trees
- logistic regression
- association analysis
- αλγόριθμος apriori
- zookeeper

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα big data και η μηχανική μάθηση
- Επιδόσεις επεξεργασίας
- Ποιοτική διάσταση
- Μηχανική Χαρακτηριστικών
- Διαφορά και βάση
- Big data στον τομέα της υγείας
- Big data στο τομέα της ενέργειας
- Big data και περιβαλλοντικό αντίκτυπο
- Big data στη βιομηχανία
- Big data και cloud computing

Περιεχόμενα παρουσίασης

Πανεπιστήμιο Δυτικής Μακεδονίας

- Παραδοσιακή αποθήκευση των δεδομένων
- Αρχιτεκτονική αποθήκευσης
- Προκλήσεις στην αποθήκευση των μεγάλων δεδομένων
- Νέα προσέγγιση στην αποθήκευση των δεδομένων
- Αναλυτικές βάσεις δεδομένων
- Εκσυγχρονισμός της αποθήκευσης των Big data
- Πλεονεκτήματα του εκσυγχρονισμού

Περιεχόμενα παρουσίασης

Πανεπιστήμιο Δυτικής Μακεδονίας

Τι είναι τα Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα είναι ένα πεδίο που αντιμετωπίζει τους τρόπους ανάλυσης, συστηματικής απόσπασης πληροφοριών ή άλλης αντιμετώπισης με σύνολα δεδομένων που είναι πολύ μεγάλα ή περίπλοκα για να αντιμετωπιστούν από το παραδοσιακό λογισμικό εφαρμογών επεξεργασίας δεδομένων.

- Με απλά λόγια, τα μεγάλα δεδομένα είναι μεγαλύτερα, πιο σύνθετα σύνολα δεδομένων από διαφορετικές πηγές δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τι είναι τα Big Data

- Αυτά τα σύνολα δεδομένων είναι τόσο ογκώδη που το παραδοσιακό λογισμικό επεξεργασίας δεδομένων δεν μπορεί να τα διαχειριστεί.

- Ωστόσο, αυτοί οι τεράστιοι όγκοι δεδομένων μπορούν να χρησιμοποιηθούν για την αντιμετώπιση πιο σύνθετων προβλημάτων που δεν θα μπορούσαν να αντιμετωπιστούν προηγουμένως.

Πανεπιστήμιο Δυτικής Μακεδονίας

Διαχείριση των Big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η ανάλυση των δεδομένων και η διαχείριση τους προσφέρουν πάντα οφέλη και τις μεγαλύτερες προκλήσεις για τους οργανισμούς.
- Οι οργανισμοί εδώ και καιρό είναι σε αναζήτηση για να βρουν μια ρεαλιστική προσέγγιση για τη συλλογή πληροφοριών σχετικά με τους πελάτες, τα προϊόντα και τις υπηρεσίες τους.
- Οι οργανισμοί και οι εταιρίες ειδικότερα που έχουν τμήμα έρευνας και ανάπτυξης (R & D), έχουν αρκετή υπολογιστική ισχύ για να τρέξουν εξελιγμένα μοντέλα ή να επεξεργαστούν τεράστιους όγκους δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Πράγματι, έχουμε πολλή πολυπλοκότητα και όγκο όσον αφορά τα δεδομένα.
- Ορισμένα δεδομένα είναι δομημένα και αποθηκευμένα σε μια παραδοσιακή σχεσιακή βάση δεδομένων.
- Ενώ άλλα δεδομένα, συμπεριλαμβανομένων των εγγράφων, των αρχείων εξυπηρέτησης πελατών, ακόμα και των εικόνων και των βίντεο, είναι αδόμητα.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Άλλες νέες πηγές πληροφοριών δημιουργούνται από ανθρώπους, όπως δεδομένα από κοινωνικά μέσα και δεδομένα που παράγονται από αναζητήσεις σε ιστότοπους.

- Επιπλέον, η διαθεσιμότητα και η υιοθέτηση νεώτερων, ισχυρότερων κινητών συσκευών, σε συνδυασμό με την πρόσβαση σε παγκόσμια δίκτυα, θα οδηγήσουν στη δημιουργία νέων πηγών δεδομένων.

- Παρόλο που κάθε πηγή δεδομένων μπορεί να διαχειριστεί και να αναζητηθεί ανεξάρτητα, η πρόκληση σήμερα είναι πως ο κόσμος γενικότερα θα αντιμετωπίσει τον τεράστιο όγκο της πληροφορίας.

- Όταν υπάρχουν τόσες πολλές πληροφορίες σε πολλές διαφορετικές μορφές, είναι αδύνατη η διαχείριση των δεδομένων με παραδοσιακούς τρόπους.

- Παρόλο που στην ιστορία του ανθρώπου υπήρχαν πάντα πολλά δεδομένα, η διαφορά σήμερα είναι ότι υπάρχουν πολύ περισσότερα από αυτά, και ποικίλλουν ανάλογα με τον τύπο και την επικαιρότητα.

- Οι επιστήμονες και οι μηχανικοί πρέπει να βρίσκουν επίσης περισσότερους τρόπους και να κάνουν χρήση αυτών των πληροφοριών.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα τρία Vs των Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα ορίζονται ως οποιαδήποτε πηγή δεδομένων που έχει τουλάχιστον τρία κοινά χαρακτηριστικά:

1. *Εξαιρετικά μεγάλες ποσότητες δεδομένων (Volume).*
2. *Εξαιρετικά μεγάλη ταχύτητα δεδομένων (Velocity).*
3. *Εξαιρετικά ευρεία ποικιλία δεδομένων (Variety).*

- Τα Μεγάλα δεδομένα είναι σημαντικά επειδή επιτρέπουν σε οργανισμούς να συλλέγουν, να αποθηκεύουν, να διαχειρίζονται και να χειρίζονται τεράστια ποσά δεδομένων με τη σωστή ταχύτητα, την κατάλληλη στιγμή, για να αποκτήσουν τις σωστές γνώσεις.

Τα τρία Vs των Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

- *Volume(Όγκος)*: Η ποσότητα των δεδομένων έχει σημασία. Τα μεγάλα δεδομένα, θα χρειαστεί να επεξεργαστούν μεγάλους όγκους δεδομένων.

- Αυτά μπορεί να είναι δεδομένα χωρίς να έχουν μετρήσιμη αξία, όπως ροές δεδομένων στο Twitter, δεδομένα σε μια ιστοσελίδα ή μια εφαρμογή για κινητά ή εξοπλισμό αισθητήρων.

- Για ορισμένους οργανισμούς, αυτός ο όγκος μπορεί να είναι δεκάδες terabyte δεδομένων. Για άλλους, μπορεί να είναι εκατοντάδες petabytes δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα τρία Vs των Big Data

- *Velocity(Ταχύτητα):* Η ταχύτητα είναι ο ρυθμός με τον οποίο λαμβάνονται τα δεδομένα .

- Κανονικά, η υψηλότερη ταχύτητα των δεδομένων εγγράφεται απευθείας στη μνήμη, αντί να γράφεται στο σκληρό δίσκο.

- Ορισμένες έξυπνες συσκευές όπως wearables και smartphones με δυνατότητα να είναι συνεχώς στο Internet λειτουργούν σε πραγματικό χρόνο και επιτρέπουν την ακόμα πιο γρήγορη ροή των δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα τρία Vs των Big Data

- *Variety(Ποικιλία):* Η ποικιλία αναφέρεται στους πολλούς τύπους δεδομένων που είναι διαθέσιμοι και στην ποικιλομορφία .

- Οι παραδοσιακοί τύποι δεδομένων διαρθρώθηκαν και προσαρμόστηκαν σε μια σχεσιακή βάση δεδομένων.

- Ορισμένες έξυπνες συσκευές όπως wearables και smartphones με δυνατότητα να είναι συνεχώς στο Internet λειτουργούν σε πραγματικό χρόνο και επιτρέπουν την ακόμα πιο γρήγορη ροή των δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα τρία Vs των Big Data

- Δύο επιπλέον Vs έχουν προκύψει τα τελευταία χρόνια: αξία(value) και αλήθεια (veracity).

- Τα δεδομένα έχουν αξία. Αλλά δεν έχουν καμία χρησιμότητα μέχρι να επαληθευθεί η αξία που προσδίδουν στην εγγύτητα της πληροφορίας. Επίσης είναι πολύ σημαντικό τα δεδομένα να είναι αξιόπιστα.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα 5 Vs των Big Data

Τα 5 Vs των Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι πρόσφατες τεχνολογικές εξελίξεις έχουν μειώσει εκθετικά το κόστος αποθήκευσης δεδομένων και τον υπολογισμό τους, καθιστώντας ευκολότερη και λιγότερο δαπανηρή αποθήκευση περισσότερων δεδομένων από ποτέ.

- Με τον αυξημένο όγκο μεγάλων δεδομένων οι επιχειρήσεις επωφελούνται και μπορεί να παρθούν πιο ακριβείς και επιχειρηματικές αποφάσεις.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα 5 Vs των Big Data

- Η εύρεση της πληροφορίας μέσα στην ποικιλία των μεγάλων δεδομένων δεν αφορά μόνο τον εντοπισμό της, που αποτελεί σίγουρα μια σημαντική ανακάλυψη.

- Αλλά είναι μια ολόκληρη διαδικασία ανακάλυψης που απαιτεί διορατικές αναζητήσεις, μηχανικούς της πληροφορικής ικανούς να αναγνωρίζουν πρότυπα, να κάνουν ερεύνα χρησιμοποιώντας την υπάρχουσα τεχνολογία.

- Είναι μια διαδικασία που ενώνει πολλούς τομείς της πληροφορικής και την εξόρυξη των δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τα 5 Vs των Big Data

Ιστορία του Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Αν και η έννοια των μεγάλων δεδομένων (Big Data) είναι σχετικά νέα, η προέλευση των μεγάλων δεδομένων αρχίζει από τη δεκαετία του 1960 - 1970, καθώς ο κόσμος των μεγάλων δεδομένων μόλις ξεκίνησε, με τις πρώτες βάσεις δεδομένων και την ανάπτυξη της σχεσιακής βάσης δεδομένων.

- Ωστόσο στην αρχή του 2005, οι άνθρωποι άρχισαν να συνειδητοποιούν το μέγεθος των δεδομένων.

- Με πρώτους τους χρήστες του Facebook, του YouTube και άλλων online πλατφορμών να χρησιμοποιούν τεράστιο όγκο δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Ιστορία του Big Data

- Με τη δημιουργία του Hadoop (δημιουργήθηκε ένα πλαίσιο ανοιχτού κώδικα ειδικά για την αποθήκευση και την ανάλυση μεγάλων συνόλων δεδομένων) την ίδια χρονιά.

- Το NoSQL άρχισε επίσης να κερδίζει δημοτικότητα κατά τη διάρκεια αυτής της περιόδου.

- Η ανάπτυξη πλαισίων ανοιχτού κώδικα, όπως το Hadoop (και πιο πρόσφατα η Spark), ήταν ουσιαστικής σημασίας για την ανάπτυξη μεγάλων δεδομένων επειδή καθιστούν τα δεδομένα μεγάλης χωρητικότητας πιο εύχρηστα και φθηνότερα στην αποθήκευση, όπως θα δούμε παρακάτω.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Με την εμφάνιση του Ίντερνετ των πραγμάτων (IoT), περισσότερα αντικείμενα και συσκευές συνδέονται στο διαδίκτυο, συγκεντρώνοντας δεδομένα σε πραγματικό χρόνο.

- Η εμφάνιση της μηχανικής μάθησης έχει παίξει επίσης σημαντικό ρόλο στην κατανόηση και ανάλυση της πληροφορίας.

- Το Cloud computing έχει επεκτείνει περισσότερο τις δυνατότητες των μεγάλων δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ανάπτυξη προϊόντων: Εταιρείες όπως η Netflix και η Procter & Gamble χρησιμοποιούν μεγάλα δεδομένα για να προβλέψουν τη ζήτηση και τις ανάγκες των πελατών.

- Επιπλέον, η P & G χρησιμοποιεί δεδομένα και αναλύσεις από κοινωνικά δίκτυα και μηχανές αναζήτησης για να προγραμματίσει, να δημιουργήσει και να προωθήσει στην αγορά νέα προϊόντα.

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

- Συντήρηση: Παράγοντες που μπορούν να προβλέψουν μηχανικές βλάβες μπορεί να είναι βαθιά θαμμένοι σε δομημένα δεδομένα, όπως το έτος του εξοπλισμού, ημερομηνία κατασκευής και μοντέλο μιας μηχανής.

- Καθώς και σε μη δομημένα δεδομένα που καλύπτουν εκατομμύρια καταχωρήσεις καταγραφής, δεδομένα αισθητήρων, σφάλματα μηνυμάτων και τη θερμοκρασία του εξοπλισμού.

- Με την ανάλυση των μεγάλων δεδομένων παίρνουμε ενδείξεις πιθανών ζητημάτων-σφαλμάτων πριν συμβούν προβλήματα-βλάβες, οι οργανώσεις μπορούν να αναπτύξουν πιο αποτελεσματικά τη συντήρηση και να μεγιστοποιήσουν το χρόνο λειτουργίας των εξαρτημάτων και εξοπλισμού.

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

- Εμπειρία πελατών: Με την ύπαρξη των Μεγάλων δεδομένων (Big Data) οι εταιρείες έχουν μια πιο σαφή εικόνα για την ικανοποίηση των πελατών.

- Μια σαφέστερη εικόνα της εμπειρίας των πελατών είναι πιο δυνατή τώρα από ποτέ.

- Με τα μεγάλα δεδομένα τώρα υπάρχει η δυνατότητα της συλλογής δεδομένων από κοινωνικά δίκτυα, επισκέψεις σε ιστοσελίδες (cookies), αρχεία καταγραφής κλήσεων και άλλες πηγές δεδομένων για τη βελτίωση της εμπειρίας του πελάτη.

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

- Μηχανική μάθηση: Η μηχανική μάθηση είναι ένα φλέγον θέμα τώρα. Επιπλέον με την ύπαρξη των δεδομένων, ειδικά των μεγάλων δεδομένων η μηχανική μάθηση είναι ένα από τα πιο πρωτοπόρα συστήματα όσον αφορά τα αυτοματοποιημένα συστήματα.

- Η μηχανική μάθηση (machine learning) μπορεί να διδάξει μηχανές αντί να τις προγραμματίσει.

- Η διαθεσιμότητα μεγάλων δεδομένων για την εκμάθηση μοντέλων μηχανικής μάθησης συμβάλλει στη βελτίωση τους όπως θα δούμε παρακάτω.

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

- Επιχειρησιακή απόδοση: Η λειτουργική αποτελεσματικότητα μπορεί να είναι ένας τομέας στον οποίο τα μεγάλα δεδομένα έχουν τον μεγαλύτερο αντίκτυπο.

- Με τα μεγάλα δεδομένα, μπορεί να αναλυθούν και να αξιολογηθούν η παραγωγή, τα σχόλια και οι ανάγκες των πελατών και άλλοι παράγοντες οι οποίοι πρέπει να ληφθούν υπόψη για να προβλεφθούν οι μελλοντικές απαιτήσεις.

- Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για τη βελτίωση της λήψης αποφάσεων σύμφωνα με τη ζήτηση της αγοράς.

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

Βέλτιστες Πρακτικές Μεγάλων Δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα συνδυάζονται πρακτικά με το τομέα της πληροφορικής.
- Παραδείγματα περιλαμβάνουν: την κατανόηση του παγκόσμιου ιστού και τις αναζητήσεις του μέσου χρήστη.
- Με τη συλλογή των πληροφοριών αυτών μπορεί να εξαχθούν δεδομένα με τις ηλεκτρονικές αγορές, την αλληλεπίδραση με τα μέσα κοινωνικής δικτύωσης.

- Τα αναλυτικά στοιχεία Big Data είναι πράγματι μια επανάσταση στον τομέα της Πληροφορικής.
- Η χρήση της ανάλυσης δεδομένων από τις εταιρείες ενισχύεται κάθε χρόνο καθώς οι ανάγκες των εταιρειών αυξάνονται.
- Τα οφέλη σε πραγματικό χρόνο του Big Data Analytic είναι πολλά.

Big data-Γιατί είναι σημαντικά

Πανεπιστήμιο Δυτικής Μακεδονίας

- Έχει σημειωθεί τεράστια αύξηση στον τομέα των Big Data με τα οφέλη της τεχνολογίας.

- Η χρήση των μεγάλων δεδομένων έχει οδηγήσει στην ανάπτυξη πολλών τομέων όπως:

1. Τραπεζικές συναλλαγές
2. *Υγεία*
3. *Ενέργεια*
4. *Τεχνολογία*
5. *Βιομηχανία*

Big data-Γιατί είναι σημαντικά

Πανεπιστήμιο Δυτικής Μακεδονίας

- Υπάρχουν πολλές βιομηχανίες που χρησιμοποιούν μεγάλες αναλύσεις δεδομένων.
- Ο τραπεζικός τομέας θεωρείται ως το πεδίο που κάνει τη μέγιστη χρήση των Big Data.
- Ο εκπαιδευτικός τομέας κάνει επίσης χρήση των αναλύσεων δεδομένων σε μεγάλο βαθμό.

- Υπάρχουν νέες δυνατότητες για την έρευνα και την ανάλυση χρησιμοποιώντας την ανάλυση δεδομένων.

- Τα θεσμικά στοιχεία μπορούν να χρησιμοποιούν τα δεδομένα για καινοτομίες σε συνδυασμό με τα τεχνικά εργαλεία που είναι διαθέσιμα σήμερα.

- Λόγω των τεράστιων ευκαιριών, η ανάλυση δεδομένων έχει γίνει μία ελκυστική επιλογή για σπουδές, τόσο για τους εκπαιδευτικούς όσο και για τους σπουδαστές.

- Οι γνώσεις που παρέχουν τα μεγάλα εργαλεία ανάλυσης δεδομένων βοηθούν στην καλύτερη γνώση των αναγκών των πελατών.
- Αυτό βοηθά στην ανάπτυξη νέων και καλύτερων προϊόντων.
- Βελτιωμένα προϊόντα και υπηρεσίες με νέες γνώσεις μπορούν να βοηθήσουν τις επιχείρησής και γενικά τους τομείς των big data.

Big data-Γιατί είναι σημαντικά

Πανεπιστήμιο Δυτικής Μακεδονίας

- Επίσης με τα big data δημιουργούνται νέες ευκαιρίες και θέσεις εργασίας.

- Είναι μεγάλο το ενδιαφέρον για επενδύσεις στις τεχνολογίες των Big Data. Οι επαγγελματίες που διαθέτουν τους πόρους πληρώνουν ελκυστικά πακέτα απολαβών και προσλαμβάνουν ειδικευμένους επαγγελματίες.

- Οι επαγγελματίες πληροφορικής, όπως μηχανικοί και διαχειριστές δεδομένων, μπορούν να μάθουν τα εργαλεία ανάλυσης και να έχουν μια πολλά υποσχόμενη καριέρα.

Αρχιτεκτονική των Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

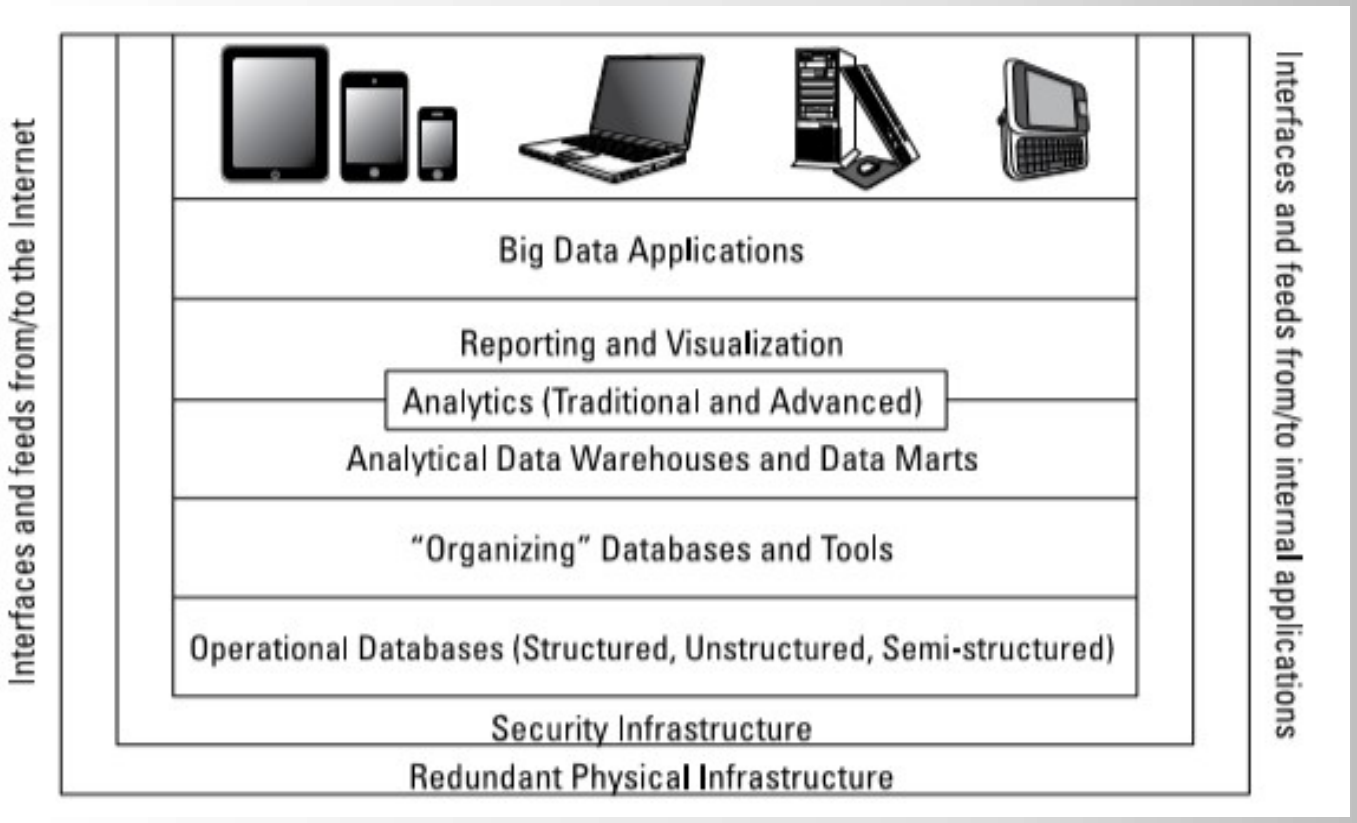
- Όπως σε κάθε σημαντική αρχιτεκτονική δεδομένων, θα πρέπει να σχεδιαστεί ένα μοντέλο το οποίο να προτείνει μια ολοκληρωμένη λύση σχετικά με το πώς όλα τα στοιχεία πρέπει να συνδέονται.

- Αν και αυτό θα χρειαστεί αρκετό χρόνο στην αρχή, έπειτα θα εξοικονομήσει πολλές ώρες ανάπτυξης κατά τη διάρκεια των μεταγενέστερων ενεργειών στην ανάπτυξη και συγκέντρωση του όγκου των δεδομένων.

- Γενικά τα μεγάλα δεδομένα πρέπει να θεωρούνται ιδιαιτέρως σημαντικά.

- Οι καλές αρχές σχεδιασμού είναι κρίσιμες όταν δημιουργείται ένα περιβάλλον για την υποστήριξη μεγάλων δεδομένων.
- Το περιβάλλον θα ασχολείται με την αποθήκευση, την ανάλυση, τις αναφορές ή τις εφαρμογές των δεδομένων αυτών.
- Το περιβάλλον πρέπει να περιλαμβάνει εκτιμήσεις για το υλικό, το λογισμικό υποδομής, τον επιχειρησιακό λογισμικό, τη διαχείριση του λογισμικού, καθώς και τις καθορισμένες διεπαφές προγραμματισμού εφαρμογών (API) και ακόμη τα εργαλεία ανάπτυξης λογισμικού.

- Στο παρακάτω σχήμα παρουσιάζεται ένα μοντέλο αρχιτεκτονικής Big data.



Πηγή από το βιβλίο Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman - Big Data For Dummies- Wiley (2013)

Αρχιτεκτονική των Big Data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Στο χαμηλότερο επίπεδο της στοίβας είναι η φυσική υποδομή, το υλικό, το δίκτυο και ο εξοπλισμός που απαιτείται.
- Οι μεγάλες υλοποιήσεις δεδομένων έχουν πολύ ειδικές απαιτήσεις σε όλα τα στοιχεία της αρχιτεκτονικής , έτσι κάθε υλοποίηση απαιτεί και διαφορετική προσέγγιση.
- Δεδομένου ότι τα μεγάλα δεδομένα αφορούν μόνο την υψηλή ταχύτητα, το υψηλό όγκο και την μεγάλη ποικιλία δεδομένων, η υλική υποδομή κυριολεκτικά βοηθάει στην υλοποίηση τους.

- Οι περισσότερες μεγάλες υλοποιήσεις δεδομένων πρέπει να έχουν πολύ καλές βάσεις εξοπλισμού, έτσι ώστε τα δίκτυα, οι διακομιστές και η φυσική αποθήκευση πρέπει να άμεσα διαθέσιμα.

- Πρέπει το κάθε τμήμα της αρχιτεκτονικής να είναι αλληλένδετο.

- Μια υποδομή ή ένα σύστημα (που οφείλει να είναι ανθεκτικό και έμπιστο) πρέπει να είναι έτοιμο να επεξεργαστεί τον όγκο της πληροφορίας και να τον αποθηκεύσει.

- Στο επόμενο στρώμα της στοίβας, έχουμε την ασφάλεια.
- Οι απαιτήσεις ασφάλειας και προστασίας προσωπικών δεδομένων για τα μεγάλα δεδομένα είναι αυξημένες λόγου του όγκου της πληροφορίας.
- Στο επόμενο στρώμα της στοίβας, έχουμε τις λειτουργικές βάσεις δεδομένων.

- Στον πυρήνα οποιουδήποτε μεγάλου περιβάλλοντος δεδομένων είναι οι μηχανές βάσεων δεδομένων που περιέχουν συλλογές στοιχείων δεδομένων οι οποίες σχετίζονται με τον τομέα εφαρμογής των Big Data.
- Αυτές οι βάσεις πρέπει να είναι άμεσα διαθέσιμες, κλιμακούμενες και σταθερές.
- Ότι πληροφορία αποθηκεύεται εκεί θα πρέπει να είναι άμεσα προσπελάσιμη από τον χρήστη.

- Στο τρίτο στρώμα υπάρχει η οργάνωση δεδομένων υπηρεσιών και τα εργαλείων συλλογής.
- Επειδή τα μεγάλα δεδομένα είναι τεράστια, οι τεχνικές έχουν εξελιχθεί ώστε να επεξεργάζονται αποτελεσματικά και χωρίς προβλήματα τα δεδομένα.
- Υπάρχουν πολλές οργανωτικές υπηρεσίες δεδομένων όπως η μηχανή Map Reduce, η οποία είναι ειδικά σχεδιασμένη για τη βελτιστοποίηση της οργάνωσης μεγάλων ροών δεδομένων.

- Στο τέταρτο στάδιο υπάρχει η αποθήκευση των δεδομένων.
- Η αποθήκευση των δεδομένων, αποτελείται από πρωταρχικές τεχνικές που χρησιμοποιούν οι οργανισμοί για τη βελτιστοποίηση των δεδομένων.
- Τυπικά, η αποθήκευση των δεδομένων, περιέχει κανονικοποιημένα δεδομένα συγκεντρωμένα από διάφορες πηγές τα οποία χρησιμοποιούνται για να διευκολύνουν την ανάλυση των δεδομένων.
- Η αποθήκευση των δεδομένων απλοποιεί τη δημιουργία αναφορών και την απεικόνιση διαφόρων στοιχείων δεδομένων.

- Στο πέμπτο στάδιο έχουμε τα εργαλεία και τις τεχνικές ανάλυσης τα οποία θα βοηθήσουν πολύ στην κατανόηση μεγάλων δεδομένων.

- Οι αλγόριθμοι που αποτελούν μέρος αυτών των εργαλείων πρέπει να είναι σε θέση να λειτουργούν και να επεξεργάζονται μεγάλες ποσότητες διαφορετικών δεδομένων.

- Λόγω αυτών των αλγορίθμων, έχει επίσης δημιουργηθεί μια νέα κατηγορία εργαλείων που θα βοηθήσουν στην κατανόηση των μεγάλων δεδομένων.

- Στο τελικό στάδιο της στοίβας εντοπίζονται οι προσαρμοσμένες εφαρμογές που προσφέρουν μια εναλλακτική μέθοδο ανταλλαγής και εξέτασης των μεγάλων πηγών δεδομένων.

- Παρόλο που όλα τα στρώματα της αρχιτεκτονικής στοίβας είναι σημαντικά από μόνα τους, αυτό το στρώμα είναι το μέρος όπου η καινοτομία και η δημιουργικότητα είναι εμφανής.

- Τα μεγάλα δεδομένα κινούνται γρήγορα και αλλάζουν εξίσου γρήγορα, έτσι οι ομάδες ανάπτυξης λογισμικού πρέπει να είναι σε θέση να δημιουργήσουν γρήγορα εφαρμογές για την αντιμετώπιση των συνεχών αλλαγών.

Τύποι big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Δομημένα Δεδομένα: Ο όρος δομημένα δεδομένα αναφέρεται σε δεδομένα που έχουν καθορισμένο μήκος και μορφή.

- Παραδείγματα δομημένων δεδομένων περιλαμβάνουν αριθμούς, ημερομηνίες και ομάδες λέξεων και αριθμών που ονομάζονται strings.

- Οι πηγές των δομημένων δεδομένων είναι συνήθως 2:

1. *Δημιουργία από υπολογιστή ή μηχανή: Τα δεδομένα που παράγονται από μηχανή γενικά αναφέρονται σε δεδομένα που δημιουργούνται από μηχάνημα χωρίς ανθρώπινη παρέμβαση.*
2. *Ανθρώπινα: Αυτά είναι δεδομένα που παρέχουν οι άνθρωποι, σε αλληλεπίδραση με τους υπολογιστές.*

Πανεπιστήμιο Δυτικής Μακεδονίας

- Μη δομημένα δεδομένα: Τα μη δομημένα δεδομένα είναι δεδομένα που δεν ακολουθούν μια καθορισμένη μορφή.
- Τα μη δομημένα δεδομένα είναι πραγματικά τα περισσότερα από τα δεδομένα που υπάρχουν.
- Μέχρι πρόσφατα, όμως, η τεχνολογία δεν μπορούσε να κάνει πολλά μαζί με τα δομημένα αυτά, εκτός από την αποθήκευση ή την ανάλυσή τους.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τύποι big data

- Τα μη δομημένα δεδομένα είναι παντού.
- Στην πραγματικότητα, τα περισσότερα άτομα και οργανισμοί περιτριγυρίζονται γύρω από μη δομημένα δεδομένα.
- Όπως συμβαίνει με τα δομημένα δεδομένα, τα μη δομημένα δεδομένα είτε είναι μηχανικά είτε παράγονται από ανθρώπους.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα ημι-δομημένα δεδομένα είναι ένα είδος δεδομένων που εμπίπτει μεταξύ δομημένων και μη δομημένων δεδομένων.
- Τα ημι-δομημένα δεδομένα δεν συμμορφώνονται υποχρεωτικά με ένα καθορισμένο σχήμα, δηλαδή δομή, αλλά μπορεί να έχουν την δική τους αξία.
- Παραδείγματα ημιδομημένων δεδομένων περιλαμβάνουν τα EDI, SWIFT και XML.
- Μπορεί να θεωρηθούν ως είδος βοήθειας για την επεξεργασία , πολυσύνθετων και πολύπλοκων δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τύποι big data

Τύποι big data

	Batch	Streaming	Complex
structured	hadoop	Key/value	Rdbms
unstructed	document	graph	columnar
both	hybrid	hybrid	hybrid

Πανεπιστήμιο Δυτικής Μακεδονίας

- *Μεταδεδομένα (Metadata):* Ένα κρίσιμο στοιχείο για την ενσωμάτωση όλων αυτών των δεδομένων είναι τα μεταδεδομένα.

- Τα μεταδεδομένα είναι οι ορισμοί, αντιστοιχίσεις και άλλα χαρακτηριστικά που χρησιμοποιούνται για να περιγράψουν τον τρόπο εύρεσης, πρόσβασης και χρήσης των στοιχείων δεδομένων.

- Ένα παράδειγμα μεταδεδομένων είναι δεδομένα σχετικά με έναν αριθμό λογαριασμού.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Αυτό μπορεί να περιλαμβάνει τον αριθμό, την περιγραφή, τον τύπο δεδομένων, το όνομα, τη διεύθυνση, τον αριθμό τηλεφώνου και το επίπεδο προστασίας προσωπικών δεδομένων.

- Τα μεταδεδομένα μπορούν να χρησιμοποιηθούν για να βοηθήσουν να οργανώσουν τα δεδομένα και την οργάνωση των νέων και μεταβαλλόμενων πηγών δεδομένων.

- Παρόλο που η ιδέα των μεταδεδομένων δεν είναι νέα, αλλάζει και εξελίσσεται στο πλαίσιο των μεγάλων δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Τύποι big data

Hadoop

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι καινοτόμοι των μηχανών αναζήτησης όπως το Yahoo! και η Google χρειάστηκαν να βρουν έναν τρόπο να κατανοήσουν τα τεράστια ποσά των δεδομένων που συλλέγουν από τους χρήστες τους.

- Οι εταιρείες αυτές έπρεπε να κατανοήσουν και τις πληροφορίες που συλλέγουν και πώς θα μπορούσαν να αποκομίσουν κέρδος από αυτά τα δεδομένα για να στηρίξουν το επιχειρηματικό μοντέλο τους.

- Το Hadoop αναπτύχθηκε επειδή αντιπροσωπεύει τον πιο ρεαλιστικό τρόπο που επιτρέπει στις εταιρείες να διαχειρίζονται τεράστιους όγκους δεδομένων εύκολα, γρήγορα και αποτελεσματικά.

- Το Hadoop επέτρεψε τα μεγάλα προβλήματα να αναλυθούν σε μικρότερα στοιχεία, ώστε η ανάλυση να μπορεί να γίνει γρήγορα, οικονομικά και πιο αποδοτικά.

- Με τη διάσπαση του προβλήματος σε μικρότερα κομμάτια είναι ευκολότερη η παράλληλη επεξεργασία, η επεξεργασία των πληροφοριών ο διαμερισμός σε μικρότερα κομμάτια.

- Το Hadoop δημιουργήθηκε αρχικά από ένα μηχανικό της Yahoo!, ο οποίος ονομάζεται Doug Cutting και είναι πλέον ένα πρόγραμμα ανοιχτού κώδικα που διαχειρίζεται το Apache Software Foundation.

- Διατίθεται υπό την Άδεια Apache v2.0. Το Hadoop είναι ένα θεμελιώδες δομικό στοιχείο που βοηθάει στην καταγραφή και επεξεργασία μεγάλων δεδομένων.

- Το Hadoop έχει σχεδιαστεί για να παραλληλίζει την επεξεργασία δεδομένων μεταξύ υπολογιστικών κόμβων με υπολογιστές ταχύτητας και να μειώνει την καθυστέρηση στην επεξεργασία τους.

- Το Hadoop έχει δύο βασικά συστατικά:

1. *Hadoop Distributed File System: Ένα αξιόπιστο, υψηλού εύρους ζώνης, χαμηλού κόστους, cluster αποθήκευσης δεδομένων που διευκολύνει τη διαχείριση των δεδομένων σε όλα τα μηχανήματα.*
2. *MapReduce: Μια παράλληλη / κατανεμημένη εφαρμογή επεξεργασίας δεδομένων υψηλής απόδοσης του αλγορίθμου MapReduce.*

- Το Hadoop έχει σχεδιαστεί για να επεξεργάζεται τεράστιες ποσότητες δομημένων και αδόμητων δεδομένων (terabytes σε petabytes).

- Το Hadoop είναι σε θέση να ανιχνεύει αλλαγές (συμπεριλαμβανομένων των βλαβών) και να προσαρμόζεται στις αλλαγές αυτές και να συνεχίζει να λειτουργεί χωρίς διακοπή, έχοντας συνεχόμενη ροή δεδομένων.

- Το σύστημα κατανομής αρχείων Hadoop (Hadoop Distributed File System) είναι μια ευέλικτη, ανθεκτική, προσέγγιση για τη διαχείριση αρχείων σε ένα μεγάλο περιβάλλον δεδομένων.

- Πρόκειται για μια υπηρεσία δεδομένων που προσφέρει ένα μοναδικό σύνολο δυνατοτήτων που απαιτούνται όταν οι όγκοι δεδομένων και οι ταχύτητες είναι υψηλές.

- Επειδή τα δεδομένα γράφονται μία φορά και στη συνέχεια διαβάζονται πολλές φορές στη συνέχεια, αντί για τις συνεχείς αναγνώσεις άλλων συστημάτων αρχείων, το HDFS (Hadoop Distributed File System) παρουσιάζεται ως μια εξαιρετική επιλογή για την υποστήριξη της μεγάλης ανάλυσης δεδομένων.

- Το Hadoop έχει δυο βασικά στοιχεία:

1. *Namenode(ονόματα κόμβων)*
2. *Datanode(Κόμβοι δεδομένων)*

- Οι κόμβοι δεδομένων δεν είναι πολύ “έξυπνοι”, αλλά η πλατφόρμα NameNode είναι.
- Οι κόμβοι δεδομένων ζητούν συνεχώς από την πλατφόρμα NameNode να τους εξυπηρετήσει.
- Αυτή η συνεχής επικοινωνία στο NameNode πρακτικά δείχνει ποιοι κόμβοι δεδομένων είναι ελεύθεροι και ποιοι είναι απασχολημένοι.

- Το NameNode είναι το κεντρικό στοιχείο ενός συστήματος αρχείων HDFS.
- Διατηρεί την δομή καταλόγου όλων των αρχείων στο σύστημα αρχείων και παρακολουθεί όπου διατηρούνται και αποθηκεύονται τα δεδομένα αρχείων σε ολόκληρο το σύστημα.
- Δεν αποθηκεύει τα δεδομένα αυτών των αρχείων.

- Ένα DataNode αποθηκεύει δεδομένα στο HadoopFileSystem.
- Ένα λειτουργικό σύστημα αρχείων έχει περισσότερους από έναν DataNode, με τα δεδομένα να πολλαπλασιάζονται σε αυτά.
- Ένα DataNode επικοινωνεί με το NameNode. Στη συνέχεια, ανταποκρίνεται σε αιτήματα από το NameNode για λειτουργίες συστήματος αρχείων.

- Οι κόμβοι δεδομένων χρησιμοποιούν τοπικούς δίσκους στο (server) διακομιστή για αποθήκευση.
- Όλα τα δεδομένα αποθηκεύονται τοπικά, κυρίως για λόγους απόδοσης.
- Τα δεδομένα αναπαράγονται σε διάφορους κόμβους δεδομένων, οπότε η αποτυχία ενός διακομιστή (server) ενδέχεται να μην καταστρέφει απαραιτήτως ένα αρχείο.

Hadoop Distributed File System (HDFS)

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα φέρνουν τις μεγάλες προκλήσεις του όγκου, της ταχύτητας και της ποικιλίας στις βάσεις δεδομένων.

- Το HDFS αντιμετωπίζει αυτές τις προκλήσεις με το ‘’σπάσιμο’’των αρχείων αυτών σε μια σχετική συλλογή μικρότερων ομάδων που αποτελούνται από δεδομένα.

- Αυτές οι ομάδες διανέμονται μεταξύ των κόμβων δεδομένων στο σύμπλεγμα του HDFS και διαχειρίζονται από τα NameNode.

Hadoop Distributed File System (HDFS)

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το HDFS υποστηρίζει τη δυνατότητα δημιουργίας αγωγών δεδομένων.
- Ο κόμβος δεδομένων αναλαμβάνει και προωθεί τα δεδομένα στο επόμενο κόμβο, στον αγωγό. Αυτό συνεχίζεται μέχρι όλα τα δεδομένα και όλα τα αντίγραφα των δεδομένων, να είναι γραμμένα στο δίσκο.
- Ένας αγωγός δεδομένων είναι μια σύνδεση μεταξύ πολλαπλών κόμβων δεδομένων (datanodes).

- Με όλον αυτόν τον όγκο των αρχείων και των αγωγών δεδομένων στους διακομιστές (servers), είναι ιδιαίτερά σημαντικό τα πράγματα να διατηρούνται σε ισορροπία.
- Για παράδειγμα υπάρχει περίπτωση, σε έναν κόμβο δεδομένων να γίνει συμφόρηση ενώ ένας άλλος μπορεί να είναι σχεδόν άδειος.
- Το HDFS έχει μια υπηρεσία που έχει σχεδιαστεί για να αντιμετωπίσει αυτές τις αντιξοότητες.

- Ο στόχος είναι η εξισορρόπηση των κόμβων δεδομένων με βάση το πόσο πλήρες είναι κάθε σύνολο τοπικών δίσκων.
- Η εξισορρόπηση εκτελείται ενώ το σύμπλεγμα (cluster) είναι ενεργό και για να αποφευχθεί η συμφόρηση της κυκλοφορίας του δικτύου.
- Το HDFS χρειάζεται να διαχειριστεί τα αρχεία και φροντίσει να είναι ισορροπημένο το σύμπλεγμα.

Mapreduce

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα έχουν κυριαρχήσει στον τομέα της πληροφορικής, ωστόσο έχουν δημιουργηθεί και μεγάλα υπολογιστικά προβλήματα.

- Κάθε φορά που ένα νεότερο, ταχύτερο και πιο ισχυρό σύστημα πληροφορικής εφευρισκόταν οι μηχανικοί έβρισκαν προβλήματα που ήταν πολύ μεγάλα για το εκάστοτε σύστημα να διαχειριστεί.

- Μαζί με τα δίκτυα υπολογιστών και η βιομηχανία στράφηκε στο συνδυασμό των υπολογιστικών και αποθηκευτικών δυνατοτήτων των συστημάτων στο δίκτυο προς την επίλυση όλο και μεγαλύτερων προβλημάτων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η κατανομή των υπολογιστών και των δεδομένων είναι έντονη, οι εφαρμογές βρίσκονται στο επίκεντρο μιας λύσης για τις μεγάλες προκλήσεις των δεδομένων.
- Για να επιτευχθεί η αξιόπιστη κατανομή σε κλίμακα, νέες τεχνολογικές προσεγγίσεις και καινοτομίες απαιτούνται.
- Το MapReduce είναι μία από αυτές τις νέες προσεγγίσεις.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το MapReduce είναι ένα λογισμικό το οποίο επιτρέπει στους προγραμματιστές να γράφουν προγράμματα που μπορούν να επεξεργαστούν μεγάλες ποσότητες μη δομημένων δεδομένων παράλληλα σε μια κατανεμημένη ομάδα επεξεργαστών.

- Στις αρχές της δεκαετίας του 2000, οι μηχανικοί της Google ερεύνησαν τις μελλοντικές απαιτήσεις και διαπίστωσαν ότι οι τρέχουσες λύσεις τους για εφαρμογές όπως ο παγκόσμιος ιστός ήταν επαρκή για τα περισσότερα υπάρχοντα προβλήματα των μεγάλων δεδομένων.

- Ωστόσο ήταν ανεπαρκείς για την πολυπλοκότητα που αναμενόταν στο εγγύς μέλλον.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι μηχανικοί διαπίστωσαν ότι αν η εργασία μπορούσε να διανεμηθεί σε διάφορους υπολογιστές και στη συνέχεια να συνδεθεί στο δίκτυο με τη μορφή ενός «συμπλέγματος», το γνωστό cluster, θα μπορούσαν να λύσουν το πρόβλημα της πολυπλοκότητας.

- Η διανομή μόνο δεν ήταν επαρκής απάντηση.

- Η διανομή του φόρτου της εργασίας πρέπει να εκτελεστεί παράλληλα για τους ακόλουθους τρεις λόγους:

1. *Η επεξεργασία να μπορεί να επεκταθεί και να συρρικνωθεί αυτόματα.*
2. *Η επεξεργασία να μπορεί να γίνει ανεξάρτητα από τις αστοχίες στο δίκτυο ή τα μεμονωμένα συστήματα.*
3. *Οι προγραμματιστές που αξιοποιούν αυτή την προσέγγιση πρέπει να είναι σε θέση να δημιουργήσουν υπηρεσίες που είναι εύκολο να αξιοποιηθεί από άλλους προγραμματιστές.*

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το MapReduce σχεδιάστηκε ως γενικό μοντέλο προγραμματισμού.
- Μερικές από τις αρχικές εφαρμογές παρείχαν όλες τις βασικές απαιτήσεις της παράλληλης εκτέλεσης, βλάβης, εξισορρόπησης φορτίου και χειρισμού δεδομένων.
- Οι μηχανικοί το ονόμασαν MapReduce επειδή συνδυάζει δύο δυνατότητες από τις υπάρχουσες λειτουργικές γλώσσες υπολογιστών: χάρτης (Map) και μείωση (Reduce).

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι μηχανικοί της Google σχεδίασαν το MapReduce για να λύσουν ένα συγκεκριμένο πρακτικό πρόβλημα.
- Ως εκ τούτου, σχεδιάστηκε ως ένα μοντέλο προγραμματισμού.
- Χρησιμοποιήθηκε για να αποδειχθεί η πρακτικότητα και η αποτελεσματικότητα του στη διαχείριση των μεγάλων δεδομένων και να συμβάλει ώστε το μοντέλο αυτό να υιοθετηθεί ευρέως από την βιομηχανία των ηλεκτρονικών υπολογιστών.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Με τα χρόνια, έχουν δημιουργηθεί και άλλες εφαρμογές του MapReduce και είναι διαθέσιμα τόσο ως προϊόντα ανοικτού κώδικα όσο και ως εμπορικά προϊόντα.

- Η λειτουργία του map υπήρξε μέρος πολλών λειτουργικών γλωσσών προγραμματισμού για χρόνια, κερδίζοντας πρώτα τη δημοτικότητα με μια γλώσσα τεχνητής νοημοσύνης που ονομάζεται LISP.

- Οι προγραμματιστές λογισμικού κατανοούν την αξία της , έτσι ο χάρτης ανανεώθηκε ως βασική τεχνολογία για την επεξεργασία των στοιχείων των μεγάλων δεδομένων.

- Όπως και η λειτουργία του χάρτη (map), η μείωση (reduce) έχει αποτελέσει χαρακτηριστικό του λειτουργικού προγραμματισμού για πολλά χρόνια.

- Η λειτουργία μείωσης (reduce) λαμβάνει την έξοδο ενός αποτελέσματος και τη διαμορφώνει με οποιονδήποτε τρόπο επιθυμεί ο προγραμματιστής.

- Η λειτουργία του reduce επεξεργάζεται κάθε στοιχείο της λίστας ξεχωριστά και επιστρέφει κάθε στοιχείο της λίστας, διαμερισμένο σε μικρότερα κομμάτια.

Πανεπιστήμιο Δυτικής Μακεδονίας

Hive

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η Hive ή αλλιώς κυψέλη είναι ένα είδος αποθήκευσης δεδομένων, που βασίζεται σε στοιχεία του πυρήνα του Hadoop (HDFS και MapReduce).
- Παρέχεται στους χρήστες που γνωρίζουν την SQL με μια απλή εφαρμογή SQL-lite που ονομάζεται και HiveQL.
- Με την κυψέλη, δανείζεται η δυνατότητα της SQL, όπως η πρόσβαση σε δομημένα δεδομένα και η εξελιγμένη ανάλυση μεγάλων δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η Hive χρησιμοποιεί τρεις μηχανισμούς για την οργάνωση δεδομένων:

1. *Πίνακες: Οι πίνακες της Hive αποτελούνται από σειρές και στήλες. Επειδή η κυψέλη (Hive) είναι δομημένη πάνω στο Hadoop και το HDFS, υπάρχουν πίνακες που έχουν αντιστοιχιστεί σε καταλόγους στο σύστημα αρχείων. Επιπλέον, υποστηρίζει πίνακες αποθηκευμένους σε άλλα συστήματα εγγενών αρχείων.*

1. *Τμήματα: Ο πίνακας της Hive μπορεί να υποστηρίξει ένα ή περισσότερα διαμερίσματα. Αυτά τα διαμερίσματα μοιράζονται σε υποκατάλογους στο σύστημα αρχείων και αντιπροσωπεύουν τη διανομή των δεδομένων σε όλο τον πίνακα.*

1. *Κάδοι: Με τη σειρά τους, τα δεδομένα μπορούν να χωριστούν σε κάδους. Οι κάδοι αποθηκεύονται ως αρχεία στον κατάλογο στο σύστημα αρχείων.*

Pig και Pig Latin

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η ισχύς και η ευελιξία του Hadoop είναι γνωστή από το λογισμικό κυρίως επειδή το οικοσύστημα του Hadoop χτίστηκε από προγραμματιστές, για προγραμματιστές.
- Ωστόσο, δεν είναι όλοι προγραμματιστές λογισμικού.
- Το Pig σχεδιάστηκε για να κάνει το Hadoop πιο προσιτό και εύχρηστο.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το Pig είναι μια πλατφόρμα που βασίζεται σε script, υποστηρίζοντας το Pig latin, μια γλώσσα που χρησιμοποιείται για την έκφραση ροών δεδομένων.

- Η γλώσσα Pig Latin υποστηρίζει τη φόρτωση και την επεξεργασία των δεδομένων εισόδου και παράγει την επιθυμητή έξοδο της ροής των δεδομένων.
- Το περιβάλλον εκτέλεσης του Pig έχει δύο λειτουργίες:

1. *Τοπική λειτουργία: Όλα τα σενάρια εκτελούνται σε ένα μόνο μηχάνημα. Hadoop MapReduce και HDFS δεν απαιτούνται.*
2. *Hadoop: Με τη λειτουργία του MapReduce, όλα τα σενάρια τρέχουν σε ένα δεδομένο Hadoop cluster*.

Πανεπιστήμιο Δυτικής Μακεδονίας

Pig και Pig Latin

- Η γλώσσα Pig Latin παρέχει έναν αφηρημένο τρόπο για να παρθούν οι απαντήσεις από τα μεγάλα δεδομένα εστιάζοντας στα δεδομένα και όχι στη δομή ενός προσαρμοσμένου προγράμματος λογισμικού.

- Το Pig απλουστεύει κατά πολύ την επεξεργασία των δεδομένων.

- Για παράδειγμα, μπορεί να εκτελεστεί ένα σενάριο Pig σε μια μικρή προσομοίωση της μεγάλης βάσης των δεδομένων για να εξακριβωθούν ότι έχουμε τα επιθυμητά αποτελέσματα προτού δεσμεύσουμε πρακτικά όλα τα δεδομένα.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα προγράμματα του περιβάλλοντος Pig μπορούν να εκτελούνται με τρεις διαφορετικούς τρόπους:

1. *Σενάριο*: *Ουσιαστικά ένα αρχείο που περιέχει εντολές της Pig Latin, ακολουθούμενες από την κατάληξη .pig. Οι εντολές μεταφράζονται από το ίδιο το Pig και εκτελούνται με διαδοχική σειρά.*
2. *Grunt: Το Grunt είναι διερμηνέας εντολών. Εδώ υπάρχει η δυνατότητα να πληκτρολογηθεί Pig Latin στη γραμμή εντολών του Grunt και το Grunt θα εκτελέσει την εντολή.*
3. *Ενσωματωμένα: Τα προγράμματα Pig μπορούν να εκτελεστούν ως μέρος ενός προγράμματος Java.*

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το Pig latin έχει μια πολύ πλούσια σύνταξη. Υποστηρίζει τις ακόλουθες λειτουργίες:

1. *Φόρτωση και αποθήκευση δεδομένων*
2. *Δεδομένα ροής*
3. *Φιλτράρισμα δεδομένων*
4. *Ομαδοποίηση και σύνδεση δεδομένων*
5. *Ταξινόμηση δεδομένων*
6. *Συνδυασμός και διαίρεση δεδομένων*

Πανεπιστήμιο Δυτικής Μακεδονίας

Scoop

Πανεπιστήμιο Δυτικής Μακεδονίας

- Πολλές επιχειρήσεις αποθηκεύουν τις πληροφορίες των δεδομένων τους σε μεγάλες βάσεις δεδομένων, έτσι χρειάζονται έναν τρόπο να μεταφέρονται τα δεδομένα από και προς αυτές τις βάσεις στο Hadoop.

- Είναι σημαντικό να μεταφέρονται τα δεδομένα σε πραγματικό χρόνο και είναι επίσης σημαντικό να γίνεται αποφυγή συμφόρησης της κίνησης των δεδομένων.

- Το Sqoop (SQL-to-Hadoop) είναι ένα εργαλείο που προσφέρει τη δυνατότητα εξαγωγής δεδομένων, τη μετατροπή των δεδομένων σε μια μορφή που μπορεί να χρησιμοποιήσει το Hadoop και στη συνέχεια, τη φόρτωση των δεδομένων σε HDFS.

Πανεπιστήμιο Δυτικής Μακεδονίας

Scoop

- Όπως το Pig, το Sqoop είναι ουσιαστικά μια γραμμή εντολών. Γίνεται η πληκτρολόγηση των εντολών στο Sqoop και εκτελούνται μία φορά τη φορά.

- Τέσσερα βασικά χαρακτηριστικά βρίσκονται στο Sqoop:

1. *Μαζική εισαγωγή: Το Sqoop μπορεί να εισάγει μεμονωμένους πίνακες ή ολόκληρες βάσεις δεδομένων σε HDFS. Τα δεδομένα αποθηκεύονται στους εγγενείς καταλόγους και αρχεία στο σύστημα αρχείων HDFS.*
2. *Άμεση είσοδος: Το Sqoop μπορεί να εισάγει και να αντιστοιχίζει βάσεις δεδομένων SQL (σχεσιακών) απευθείας σε Hive και Hbase.*

Πανεπιστήμιο Δυτικής Μακεδονίας

3. *Αλληλεπίδραση δεδομένων: το Sqoop μπορεί να δημιουργήσει κλάσεις Java, ώστε να μπορεί να γίνει η αλληλεπίδραση με τα δεδομένα προγραμματιστικά.*

3. *Εξαγωγή δεδομένων: Το Sqoop μπορεί να εξάγει δεδομένα απευθείας από το HDFS σε μια σχεσιακή βάση δεδομένων χρησιμοποιώντας έναν πίνακα στόχων με βάση τις ιδιαιτερότητες της βάσης δεδομένων.*

- *Το Sqoop λειτουργεί εξετάζοντας τη βάση δεδομένων που θέλουμε να εισάγουμε και επιλέγει μια κατάλληλη λειτουργία εισαγωγής για τα δεδομένα.*

Πανεπιστήμιο Δυτικής Μακεδονίας

Scoop

- Αφού αναγνωρίσει την είσοδο, διαβάζει τα μεταδεδομένα στον πίνακα και δημιουργεί μια κλάση.

- Το Sqoop μπορεί να προγραμματιστεί να είναι πολύ επιλεκτικό, ώστε να έχει μόνο τις στήλες που ψάχνουμε πριν από την είσοδο των δεδομένων παρά να ψάχνει τα δεδομένα συνολικά.

- Αυτό μπορεί να εξοικονομήσει χρόνο. Η εισαγωγή από την εξωτερική βάση δεδομένων σε HDFS εκτελείται από την MapReduce.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το Sqoop είναι ένα αποτελεσματικό εργαλείο για τους μη προγραμματιστές.
- Το άλλο σημαντικό στοιχείο που πρέπει να σημειωθεί είναι η συσχέτιση από τις τεχνολογίες του HDFS και του MapReduce.
- Το Scoop δεν μπορεί να υπάρξει μόνο του χωρίς το Hadoop.

Πανεπιστήμιο Δυτικής Μακεδονίας

Scoop

Zookeeper

Πανεπιστήμιο Δυτικής Μακεδονίας

- Από τα μεγαλύτερα πλεονεκτήματα του Hadoop για την αντιμετώπιση μεγάλων προκλήσεων δεδομένων, είναι η ικανότητά του να διαιρεί τα δεδομένα και να τα επεξεργάζεται σε διαφορετικά υπολογιστικά συστήματα.

- Μετά το διαμερισμό του προβλήματος, χρησιμοποιούνται τεχνικές κατανεμημένης και παράλληλης επεξεργασίας σε ολόκληρο το σύμπλεγμα (cluster) του Hadoop.

- Ωστόσο για ορισμένα προβλήματα μεγάλων δεδομένων, τα διαδραστικά εργαλεία δεν είναι σε θέση να παράσχουν τις γνώσεις ή την λύση που απαιτούνται για το πρόβλημα.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Σε αυτές τις περιπτώσεις, απαιτείται η δημιουργία εφαρμογών για την επίλυση αυτών των μεγάλων προβλημάτων δεδομένων.
- Το Zookeeper είναι ο τρόπος συντονισμού του Hadoop για όλα τα στοιχεία αυτών των κατανεμημένων εφαρμογών.
- Το Zookeeper ως τεχνολογία είναι πραγματικά απλό, αλλά τα χαρακτηριστικά του είναι ισχυρά.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ορισμένες από τις δυνατότητες του Zookeeper είναι οι εξής:

1. *Συγχρονισμός διαδικασίας*: *Το Zookeeper συντονίζει την εκκίνηση και τη διακοπή πολλαπλών κόμβων στο σύμπλεγμα. Αυτό εξασφαλίζει ότι όλη η επεξεργασία πραγματοποιείται με την επιθυμητή σειρά.*
2. *Διαχείριση παραμέτρων: Το Zookeeper μπορεί να χρησιμοποιηθεί για την αποστολή χαρακτηριστικών διαμόρφωσης σε οποιονδήποτε ή σε όλους τους κόμβους του συμπλέγματος.*
3. *Αξιόπιστα μηνύματα: Το Zookeeper προσφέρει την επικοινωνία μεταξύ των κόμβων του συμπλέγματος που αφορά στην κατανεμημένη εφαρμογή.*

Πανεπιστήμιο Δυτικής Μακεδονίας

Nosql-MongoDB

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το MongoDB είναι βασισμένο σε μη σχεσιακές βάσεις δεδομένων (NoSQL).
- Είναι γραμμένο στην C++ από την 10gen Corporation ως έργο ανοιχτού κώδικα.
- Η αρχιτεκτονική του MongoDB έχει τρία κύρια στοιχεία, όπως θα δούμε παρακάτω.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η διαδικασία Mongodb χειρίζεται τα δεδομένα, τη μορφή των δεδομένων και την αποθήκευση των δεδομένων.

- Το Mongo έχει το ρόλο του δρομολογητή για το σύμπλεγμα.

- Το Mongo πρακτικά, είναι ένα κέλυφος JavaScript που βοηθά τον διαχειριστή να εκτελέσει δοκιμές αναζήτησης μέσα στη βάση των δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Big data mining

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ο όρος "εξόρυξη δεδομένων" προέκυψε από την κοινότητα του μάρκετινγκ, ανάμεσα στα τέλη της δεκαετίας του 1970 και τις αρχές της δεκαετίας του 1980.
- Οι μελετητές της στατιστικής δεν κατάλαβαν τον ενθουσιασμό που προκάλεσε αυτή η νέα τεχνική, αφού η ανακάλυψη των μοτίβων και των δομών στα δεδομένα δεν ήταν νέα για αυτούς.
- Ήξεραν για την εξόρυξη δεδομένων για μεγάλο χρονικό διάστημα, αν και με διαφορετικά ονόματα, όπως fishing και snooping.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Επειδή οποιαδήποτε διαδικασία ανακάλυψης εκμεταλλεύεται πλήρως τα δεδομένα, δημιουργήθηκαν πολλά ερωτήματα .

- Οι επιστήμονες της στατιστικής δεν θεώρησαν την εξόρυξη δεδομένων ως κάτι το θετικό.

- Οπότε στην αρχή τουλάχιστον την αντιμετώπιζαν πολύ επιφυλακτικά.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η εξόρυξη δεδομένων ασχολείται με την απόκτηση γνώσεων και την εύρεση προτύπων στο σύνολο των δεδομένων.
- Το μοντέλο, η διαδικασία εξόρυξης δεδομένων και οι αλγόριθμοι που εφαρμόζονται στα δεδομένα να έχουν ως χαρακτηριστικά τα μοτίβα και την ομαδοποίηση των δεδομένων.
- Τα μοντέλα εξόρυξης δεδομένων είναι τα περισσότερα κοινά όσων αφορά την ταξινόμηση, τους κανόνες σύνδεσης και την ομαδοποίηση τους.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η εξόρυξη δεδομένων έχει ένα ευρύ φάσμα εφαρμογών στην επιστήμη, τη μηχανική και γενικά στον τομέα της πληροφορικής.
- Για παράδειγμα, στην πρόγνωση του καιρού, το μοντέλο ταξινόμησης των μετεωρολογικών προβλέψεων μπορεί να χρησιμοποιηθεί για την πρόβλεψη του καιρού της επόμενης ημέρας.
- Οι αλγόριθμοι συμπλέγματος (clustering algorithms) έχουν επίσης ευρύ φάσμα εφαρμογών, όπως την ανάλυση των δεδομένων και οι τεχνικές ανάκτησης τους.
- Μπορεί επίσης να χρησιμοποιηθούν για την πρόβλεψη των καταναλωτικών συνηθειών.

Πανεπιστήμιο Δυτικής Μακεδονίας

Αλγόριθμος Apriori

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ο αλγόριθμος A-priori είναι ένας από τους κοινούς αλγόριθμους εξόρυξης δεδομένων που χρησιμοποιείται για την εύρεση συχνών συνόλων στοιχείων σε βάσεις δεδομένων.
- Ο A-priori λειτουργεί με την εύρεση συχνών στοιχείων από τη βάση δεδομένων.
- Στη συνέχεια, ο αλγόριθμος προσπαθεί να βρει τις σχέσεις ή τις συσχετίσεις μεταξύ αντικειμένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Αλγόριθμος Apriori

- Ο αλγόριθμος A-priori αναπτύχθηκε στην IBM από την Agrawal για την εύρεση συχνών συνόλων στοιχείων σε βάσεις δεδομένων . Ο ψευδοκώδικας του αλγορίθμου είναι παρακάτω:

- ***Ck**: candidate itemset of size K.*
- ***Lk**: frequent itemset of size K*
- ***L1**: {frequent itemset} // 1-itemset by scanning D.*
- *For (K=1, Lk != , K++) do begin*
- ***Ck+1** = candidates generated from Lk*
- *For each transaction t in D do // scan the database for support count. Increment the count of all candidates in Ck+1 that are in t*
- ***Lk+1** = candidates in Ck+1 with minimum support.*
- ***End Return** ; Join step: Ck is generated by joining Lk-1 with itself.*
- *Prune step: any (k-1) item-set that is not frequent, its subs so is not frequent.*

Πανεπιστήμιο Δυτικής Μακεδονίας

Αλγόριθμος Apriori

- Επεξήγηση του αλγορίθμου:

1. *Αρχικά, γίνεται σάρωση ολόκληρης της βάσης δεδομένων για τον προσδιορισμό των αριθμών στοιχείων στη βάση δεδομένων D (1-itemet ή L1).*

1. *Στη συνέχεια, ο αλγόριθμος θα ενταχθεί (L1) για να παράγει το επόμενο συχνό σύνολο στοιχείων (k-itemset).*

1. *Δύο αντικείμενα μπορούν να ενωθούν τα (k-1) .*

1. *Τέλος, ο αλγόριθμος επαναλαμβάνεται μέχρις ότου το Lk να είναι κενό.*

Πανεπιστήμιο Δυτικής Μακεδονίας

Αλγόριθμος Apriori

Προεπεξεργασία

Πανεπιστήμιο Δυτικής Μακεδονίας

- Προ επεξεργασία: Τα δεδομένα έχουν διαφορετικές δομές όπως φωτογραφίες, δεδομένα αισθητήρων και κείμενα.
- Επίσης, τα δεδομένα προέρχονται από διαφορετικές πηγές, οι οποίες πρέπει να ενσωματωθούν από πολλαπλές πηγές.
- Καθαρισμός δεδομένων: Τα πραγματικά δεδομένα είναι ελλιπή και απαιτούν τη συμπλήρωση ελλιπών δεδομένων, εξομάλυνση των αποκλίσεων και αποκατάσταση πληροφοριών.

Πανεπιστήμιο Δυτικής Μακεδονίας

Προεπεξεργασία

- Μετασχηματισμός δεδομένων: Τα δεδομένα πρέπει να εξομαλυνθούν ή να μετατραπούν σε μία μορφή.
- Ενσωμάτωση δεδομένων: Μερικές φορές απαιτείται η ενσωμάτωση δεδομένων από πολλαπλές πηγές.
- Κατανόηση διαφορετικών τύπων δεδομένων: Τα δεδομένα έχουν διαφορετικές μορφές και δομές, όπως όπως αριθμητικά δεδομένα, ονομαστικά δεδομένα, δυαδικά δεδομένα κ.ο.κ.

Πανεπιστήμιο Δυτικής Μακεδονίας

Προεπεξεργασία

Κλασικές μέθοδοι

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι μέθοδοι εξόρυξης δεδομένων έχουν συμβάλει σε μεγάλο βαθμό στη στατιστική, μηχανική μάθηση, τεχνητή νοημοσύνη και στα συστήματα βάσεων δεδομένων.

- Ωστόσο με τη χρήση των στατιστικών στοιχείων δεν γίνεται η εξόρυξη δεδομένων.

- Οι στατιστικές μέθοδοι χρησιμοποιήθηκαν πολύ πριν από την εξόρυξη δεδομένων που σχεδιάστηκε για να εφαρμοστεί στις επιχειρηματικές εφαρμογές.

Κλασικές μέθοδοι

Πανεπιστήμιο Δυτικής Μακεδονίας

- Δύο από τις πιο ευρέως χρησιμοποιούμενες στατιστικές μεθόδους είναι η πολλαπλή γραμμική παλινδρόμηση και η λογιστική παλινδρόμηση.
- Η πολλαπλή γραμμική παλινδρόμηση και λογική παλινδρόμηση χρησιμοποιούνται συνήθως στην εξόρυξη δεδομένων.
- Μια έντονη διαφορά μεταξύ των εφαρμογών τους και των εφαρμογών εξόρυξης δεδομένων είναι ο τρόπος με τον οποίο καθορίζεται η καταλληλότητα του εκάστοτε μοντέλου.

- Μια τυπική εφαρμογή εξόρυξης δεδομένων είναι η πρόβλεψη ενός αποτελέσματος, με βάση τις επεξηγηματικές μεταβλητές, τις εισροές ή τα χαρακτηριστικά γνωρίσματα της ορολογίας εξόρυξης δεδομένων.

- Λόγω της έμφασης στην πρόβλεψη, οι κατανομές των στόχων ή των σφαλμάτων είναι πολύ λιγότερο σημαντικές.

- Συχνά χρησιμοποιούνται και ιστορικά δεδομένα στην ανάπτυξη του μοντέλου εξόρυξης δεδομένων.

- Οι τεχνικές προγνωστικής προσομοίωσης (supervised learning) επιτρέπουν στον αναλυτή να προσδιορίσει εάν ένα σύνολο μεταβλητών εισόδου είναι χρήσιμο στην πρόβλεψη κάποιου αποτελέσματος.

- Οι περιγραφικές τεχνικές επιτρέπουν το προσδιορισμό των υποκείμενων μοντέλων σε ένα σύνολο δεδομένων.

Κλασικές μέθοδοι

Πανεπιστήμιο Δυτικής Μακεδονίας

K-clustering

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το k-Means clustering είναι ένας περιγραφικός αλγόριθμος που ταιριάζει στη φιλοσοφία των μεγάλων δεδομένων.

- Η ανάλυση του συμπλέγματος έχει ευρεία εφαρμογή, συμπεριλαμβανομένης της κατάτμησης των στοιχείων, της αναγνώρισης προτύπων, των βιολογικών μελετών και της ταξινόμησης.

- Το k-clustering προσπαθεί να ομαδοποιήσει την εύρεση διαμερισμάτων k στα δεδομένα.

Πανεπιστήμιο Δυτικής Μακεδονίας

K-clustering

- Τα βασικά βήματα για το k-clustering είναι:

1. *Επιλέγουμε k στοιχεία αυθαίρετα .*

1. *Προσθέτουμε κάθε στοιχείο στο σύμπλεγμα.*

1. *Μόλις τοποθετηθούν όλα τα k στοιχεία , υπολογίζουμε εκ νέου τις θέσεις των k στοιχείων.*

Πανεπιστήμιο Δυτικής Μακεδονίας

- Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι τα στοιχεία να μην αλλάζουν πλέον θέση.

- Αυτή η επανάληψη συμβάλλει στην ελαχιστοποίηση της μεταβλητότητας και στη μεγιστοποίηση της μεταβλητότητας μεταξύ των συμπλεγμάτων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Association analysis

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η ανάλυση της σύνδεσης προσδιορίζει τις ομάδες προϊόντων ή υπηρεσιών που τείνουν να αγοράζονται ταυτόχρονα ή να αγοράζονται σε διαφορετικές χρονικές στιγμές από τον ίδιο πελάτη.

- Η ανάλυση της σύνδεσης εμπίπτει στην περιγραφική φάση μοντελοποίησης της εξόρυξης δεδομένων.

- Η ανάλυση της σύνδεσης βοηθά στην απάντηση πολλών ερωτήσεων στον τομέα των μεγάλων δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Association analysis

Multiple Linear Regression

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η πολλαπλή γραμμική παλινδρόμηση έχει πολλά πλεονεκτήματα λόγω της εφαρμογής στα μεγάλα δεδομένα.

- Τα μοντέλα παλινδρόμησης παράγουν επίσης άριστες εκτιμήσεις σε σχεδόν όλα τα προβλήματα.

- Πολλά φαινόμενα δεν μπορούν να περιγραφούν από γραμμικές σχέσεις μεταξύ των μεταβλητών στόχων και των μεταβλητών εισόδου.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Μπορεί να χρησιμοποιηθεί η πολυωνυμική παλινδρόμηση στο μοντέλο, που έχει χρησιμοποιηθεί σε κάθε περίπτωση ξεχωριστά για να προσεγγισθούν πιο σύνθετες μη γραμμικές σχέσεις.

- Οι επιστήμονες δεδομένων σχεδόν πάντα χρησιμοποιούν βηματική επιλογή παλινδρόμησης.

- Ούτως η άλλως ο βασικός στόχος της προγνωστικής μοντελοποίησης είναι να οικοδομηθεί ένα μοντέλο που μπορεί να χρησιμοποιηθεί γενικά στα μεγάλα δεδομένα.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι τρεις βαθμιδωτές μέθοδοι παλινδρόμησης είναι οι εξής:

1. *Η επιλογή προώθησης εισάγει τα δεδομένα κάθε φορά έως ότου δεν μπορούν να εισαχθούν περισσότερες μεταβλητές.*

1. *Η εξάλειψη προς τα πίσω αφαιρεί τις εισόδους μία κάθε φορά μέχρι να μην υπάρχει πλέον η ανάγκη για αφαίρεση των δεδομένων.*

1. *Η βηματική επιλογή είναι ένας συνδυασμός επιλογής προς τα εμπρός και εξάλειψης προς τα πίσω.*

Πανεπιστήμιο Δυτικής Μακεδονίας

- Όλες οι πιθανές ρουτίνες συνδυασμού μοντέλων παλινδρόμησης υποστηρίζονται επίσης στα εργαλεία εξόρυξης δεδομένων.

- Αυτές οι μέθοδοι θα πρέπει να είναι πιο εφικτές στον υπολογισμό των μεγάλων δεδομένων, καθώς οι αναλυτικές υπολογιστικές συσκευές υψηλής απόδοσης συνεχίζουν να γίνονται πιο ισχυρές.

- Επίσης οι εκτιμητές συρρίκνωσης(Shrinkage estimators) όπως ο απόλυτος τελεστής συρρίκνωσης και επιλογής (LASSO), προτιμώνται έναντι των πραγματικών μεθόδων βηματικής επιλογής.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Χρησιμοποιούν πληροφορίες από το πλήρες μοντέλο για να παρέχουν μια υβριδική εκτίμηση των παραμέτρων παλινδρόμησης συρρικνώνοντας τις εκτιμήσεις του πλήρους μοντέλου.

- Ο πολυπλεξία (Multicollinearity) συμβαίνει όταν μια είσοδος των δεδομένων είναι σχετικά υψηλά συσχετισμένη με τουλάχιστον άλλη είσοδο.

- Δεν είναι ασυνήθιστο στην εξόρυξη δεδομένων και δεν αποτελεί πρόβλημα.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η πολυπλεξία τείνει να διογκώσει τα τυπικά σφάλματα των εκτιμήσεων των αποτελεσμάτων των βάσεων δεδομένων.

- Σε άλλες περιπτώσεις, τα αποτελέσματα μπορούν να διπλασιαστούν ή να μειωθούν κατά το ήμισυ.

- Η πολλαπλή γραμμική παλινδρόμηση χρησιμοποιείται κυρίως για συνεχείς στόχους-αποτελέσματα. Μια από τις καλύτερες πηγές για μοντελοποίηση παλινδρόμησης γίνεται από τον Rawlings.

Πανεπιστήμιο Δυτικής Μακεδονίας

Logistic Regression

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η λογική παλινδρόμηση είναι μια μορφή ανάλυσης παλινδρόμησης στην οποία η μεταβλητή στόχος (μεταβλητή απόκρισης) είναι κατηγορηματική.
- Είναι ο αλγόριθμος που στην εξόρυξη δεδομένων χρησιμοποιείται ευρύτερα για την πρόβλεψη της πιθανότητας.
- Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί για την εκτίμηση της απάτης, της κακής πιστωτικής κατάστασης, της τάσης αγοράς και σε πολλές άλλες εφαρμογές.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η πολυτονική λογική παλινδρόμηση υποστηρίζει περισσότερα από δύο διαφορετικά αποτελέσματα αναζήτησης μέσα στη βάση των μεγάλων δεδομένων.

- Για τη λογική παλινδρόμηση, η αναμενόμενη πιθανότητα εμφάνισης του αποτελέσματος της αναζήτησης μετασχηματίζεται από μια συνάρτηση ζεύξη.

- Η λογιστική παλινδρόμηση απαιτεί επίσης πλήρη την ανάλυση των περιπτώσεων. Διαφορετικά, οποιαδήποτε παρατήρηση απορρίπτεται από την ανάλυση.

Πανεπιστήμιο Δυτικής Μακεδονίας

Decision Trees
(Δένδρα αποφάσεων)

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ένα δέντρο απόφασης είναι ένας άλλος τύπος αναλυτικής προσέγγισης που αναπτύσσεται στις κοινότητες των στατιστικών και της τεχνητής νοημοσύνης.
- Το δέντρο αντιπροσωπεύει μια κατάτμηση των δεδομένων που δημιουργείται εφαρμόζοντας μια σειρά απλών κανόνων.
- Κάθε κανόνας αντιπροσωπεύει τη τιμή μιας εισόδου.

Decision Trees(Δένδρα αποφάσεων)

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ένας κανόνας εφαρμόζεται μετά το άλλο, με αποτέλεσμα μια ιεραρχία τμημάτων μέσα σε τμήματα.
- Η ιεραρχία ονομάζεται δέντρο και κάθε τμήμα ονομάζεται κόμβος.
- Το αρχικό τμήμα περιέχει ολόκληρο το σύνολο δεδομένων και ονομάζεται (ρίζα) root και βρίσκεται στο κόμβο του δέντρου.

- Ένας κόμβος με όλους τους διαδόχους του σχηματίζει έναν κλάδο του κόμβου που τον δημιούργησε.
- Οι τελικοί κόμβοι ονομάζονται φύλλα.
- Για κάθε φύλλο, λαμβάνεται απόφαση και εφαρμόζεται όλοι οι κανόνες του δένδρου στο εκάστοτε φύλλο.

Μηχανική Μάθηση

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η επανάσταση των Big Data υπόσχεται να μεταμορφώσει τον τρόπο με τον οποίο ζούμε, εργαζόμαστε και σκεπτόμαστε, επιτρέποντας την ανακάλυψη των γνώσεων και βελτιώνοντας τη λήψη αποφάσεων.

- Η πραγματοποίηση αυτής της αλλαγής βασίζεται στην ικανότητα εξαγωγής των δεδομένων από τέτοια τεράστια δεδομένα μέσω της ανάλυσης δεδομένων.

- Η μηχανική μάθηση στον πυρήνα της, λόγω της ικανότητάς της να μαθαίνει από τα δεδομένα και να παρέχει γνώμες που βασίζονται σε δεδομένα, αποφάσεις και προβλέψεις, είναι ιδιαιτέρως σημαντική.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το Παγκόσμιο Ινστιτούτο McKinsey δήλωσε ότι η Μηχανική Μάθηση θα είναι ένας από τους βασικούς παράγοντες της επανάστασης των Big Data.
- Ο λόγος γι 'αυτό είναι η ικανότητά της Μηχανικής Μάθησης να μαθαίνει από τα δεδομένα και να παρέχει πληροφορίες, αποφάσεις και προβλέψεις.
- Ωστόσο, δεν απαιτεί τη χρήση στατιστικών αποδείξεων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ανάλογα με τη φύση των διαθέσιμων δεδομένων, οι δύο κύριες κατηγορίες μαθησιακών εργασιών είναι:

1. η εποπτευόμενη μάθηση όταν οι εισροές και οι επιθυμητές εξόδους είναι γνωστές και οι αναλυτές προσπαθούν να χαρτογραφούν τις εισροές στις εξόδους
2. και τη μη επιτηρούμενη μάθηση όταν επιθυμούν να μην γνωρίζουν και το σύστημα να ανιχνεύει τη δομή μέσα στα δεδομένα.

- Η ταξινόμηση και η παλινδρόμηση είναι παραδείγματα εποπτευόμενης μάθησης: στην ταξινόμηση οι έξοδοι παίρνουν διακριτές τιμές (ετικέτες κλάσης) ενώ κατά την παλινδρόμηση οι έξοδοι είναι συνεχείς.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα περιγράφονται συχνά από τις διαστάσεις τους, τα γνωστά Vs όπως αναφέραμε σε προηγούμενη ενότητα.

- Οι παλαιότεροι ορισμοί των Big Data επικεντρώθηκαν σε τρία Vs (όγκος, ταχύτητα και ποικιλία). Ωστόσο, ο πιο κοινός αποδεκτός ορισμός βασίζεται τώρα στα ακόλουθα τέσσερα Vs : όγκος, ταχύτητα, ποικιλία και αλήθεια.

- Αυτό δίνει τη δυνατότητα να συνδέονται άμεσα οι προκλήσεις με τα χαρακτηριστικά των Μεγάλων Δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

Μηχανική Μάθηση- Επιδόσεις Επεξεργασίας

Πανεπιστήμιο Δυτικής Μακεδονίας

- Μία από τις κύριες προκλήσεις που αντιμετωπίζουν οι υπολογισμοί με το Big Data προέρχεται από την απλή αρχή ότι ο όγκος προσθέτει υπολογιστική πολυπλοκότητα.

- Κατά συνέπεια, καθώς η κλίμακα γίνεται μεγάλη, ακόμη και οι ασήμαντες πράξεις μπορεί να ιδιαίτερα πολύπλοκες.

- Είναι λοιπόν επιτακτική ανάγκη να υπάρχουν αλγόριθμοι για να αντιμετωπίσουν το κάθε πρόβλημα που θα παρουσιαστεί.

- Επιπλέον, καθώς αυξάνεται το μέγεθος των δεδομένων, η απόδοση των αλγορίθμων εξαρτάται από την αρχιτεκτονική που χρησιμοποιείται για την αποθήκευση των δεδομένων.

- Ως εκ τούτου, όχι μόνο δεν είναι δυνατή η επίτευξη της αποτελεσματικότητας, αλλά και η ανάγκη επανεξέτασης της τυπικής αρχιτεκτονικής που χρησιμοποιείται για την υλοποίηση και την ανάπτυξη αλγορίθμων.

Μηχανική Μάθηση-Πολλαπλασιαστικότητα

Πανεπιστήμιο Δυτικής Μακεδονίας

- Πολλοί αλγόριθμοι μηχανικής μάθησης βασίζονται στην υπόθεση ότι τα δεδομένα που επεξεργάζονται μπορούν να κρατηθούν εξ ολοκλήρου στη μνήμη ή σε ένα μόνο αρχείο στο δίσκο αποθήκευσης.

- Πολλοί τύποι αλγορίθμων σχεδιάζονται για στρατηγικές και δημιουργούν δυσκολίες που εξαρτώνται από την εγκυρότητα των τύπων των δεδομένων.

- Ωστόσο, όταν το μέγεθος των δεδομένων οδηγεί στην αποτυχία της εγκυρότητας, επηρεάζονται έτσι ολόκληροι τύποι αλγορίθμων. Αυτή η πρόκληση αναφέρεται ως πολλαπλασιατικότητα των αλγορίθμων.

- Μία από τις προσεγγίσεις που προτάθηκαν ως λύση για αυτό το πρόβλημα είναι το MapReduce, ένα κλιμακωτό παράδειγμα προγραμματισμού για την επεξεργασία μεγάλων συνόλων δεδομένων μέσω παράλληλης εκτέλεσης σε μεγάλο αριθμό κόμβων.

- Ορισμένοι αλγόριθμοι μηχανικής μάθησης είναι εγγενώς παράλληλοι και μπορούν να προσαρμοστούν στο MapReduce, ενώ οι υπόλοιποι αλγόριθμοι εκμεταλλεύονται το μεγάλο αριθμό κόμβων υπολογιστών.

- Οι τρεις κατηγορίες αλγορίθμων που αντιμετωπίζουν το πρόβλημα της πολλαπλασιαστικότητας όταν επιχειρούν να χρησιμοποιήσουν το παράδειγμα MapReduce περιλαμβάνουν επαναληπτικό γράφημα, κλίση κλίσης και αλγόριθμους μεγιστοποίησης.

- Ο επαναληπτικός τους χαρακτήρας μαζί με την εξάρτησή τους από τα δεδομένα μνήμης δημιουργούν δυσκολίες με τη λειτουργία του MapReduce.

- Αυτό οδηγεί σε δυσκολίες, όπως στην προσαρμογή αυτών των αλγορίθμων στο MapReduce ή σε ένα άλλο κατανεμημένο παράδειγμα υπολογισμού.

- Συνεπώς, αν και ορισμένοι αλγόριθμοι όπως το k-cluster που αναφέραμε σε προηγούμενη ενότητα μπορούν να προσαρμοστούν για να αντιμετωπίσουν την αύξηση των δεδομένων μέσω παραλληλισμού και κατανεμημένων υπολογισμού, άλλοι εξακολουθούν να είναι οροθετημένοι.

Μηχανική Μάθηση- Ανισορροπία Κλάσεων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Καθώς τα σύνολα δεδομένων μεγαλώνουν, η υπόθεση ότι τα δεδομένα είναι ομοιόμορφα κατανεμημένα σε όλες τις τάξεις είναι συχνά διφορούμενη.

- Αυτό οδηγεί σε μια πρόκληση που αναφέρεται ως ταξική ανισορροπία: η απόδοση ενός αλγορίθμου μηχανικής μάθησης μπορεί να επηρεαστεί αρνητικά όταν τα σύνολα δεδομένων περιέχουν δεδομένα από διαφορετικές τάξεις.

- Αυτό το πρόβλημα είναι ιδιαίτερα εμφανές όταν ορισμένες κατηγορίες αντιπροσωπεύονται από ένα μεγάλο αριθμό δεδομένων και μερικές από πολύ λίγες.

- Η ανισορροπία των κλάσεων δεν είναι αποκλειστική για τα μεγάλα δεδομένα και έχει αποτελέσει αντικείμενο έρευνας για περισσότερο από μια δεκαετία.

- Η πολυπλοκότητα των μεγάλων δεδομένων αναμένεται να είναι υψηλή, πράγμα που θα μπορούσε να έχει σοβαρές επιπτώσεις από την ταξική ανισορροπία.

- Αυτή η πρόκληση είναι πιο κοινή, σοβαρή και πολύπλοκη στο πλαίσιο των μεγάλων δεδομένων, διότι η έκταση της ανισορροπίας επεκτείνεται συνεχώς λόγω του αυξημένου μεγέθους των δεδομένων.

- Η πιθανότητα να προκύψει ανισορροπία κλάσης είναι υψηλή.
- Επιπλέον, λόγω των σύνθετων προβλημάτων που περιλαμβάνονται σε αυτά τα δεδομένα, οι πιθανές επιπτώσεις της ανισορροπίας των κλάσεων στη μηχανική μάθηση είναι σοβαρές.
- Επίσης τα δέντρα αποφάσεων, και οι αλγόριθμοι αναζήτησης είναι πολύ ευαίσθητα στην ανισορροπία της τάξης.

Μηχανική Μάθηση- Αντιμετώπιση των Διαστάσεων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ένα άλλο ζήτημα που σχετίζεται με τον όγκο των μεγάλων δεδομένων είναι το πρόβλημα των διαστάσεων, η οποία αναφέρεται στις δυσκολίες που συναντάμε όταν εργαζόμαστε σε χώρο μεγάλων διαστάσεων.

- Ειδικά, η διαστασιολογία περιγράφει τον αριθμό των χαρακτηριστικών που υπάρχουν στο σύνολο δεδομένων.

- Ως εκ τούτου, καθώς ο αριθμός των χαρακτηριστικών αυξάνεται, η απόδοση και η ακρίβεια των αλγορίθμων μηχανικής μάθησης μειώνεται.

- Αυτό μπορεί να εξηγηθεί από την ανάλυση της λογικής που βασίζονται πολλοί αλγόριθμοι μηχανικής μάθησης.

- Δυστυχώς, όσο μεγαλύτερη είναι η ποσότητα των διαθέσιμων δεδομένων για την περιγραφή ενός φαινομένου, τόσο μεγαλύτερη γίνεται η δυσκολία της επεξεργασίας, επειδή υπάρχουν περισσότερα χαρακτηριστικά.

- Συνεπώς, καθώς ο όγκος των μεγάλων δεδομένων αυξάνεται, το ίδιο συμβαίνει και με τις διαστάσεις τους.

- Επιπλέον, η αντιμετώπιση των διαστάσεων επηρεάζει την απόδοση της επεξεργασίας: η χρονική και χωρική πολυπλοκότητα των αλγορίθμων της μηχανικής μάθησης σχετίζεται στενά με τη διάσταση και την πολυπλοκότητα των δεδομένων.

- Η χρονική πολυπλοκότητα πολλών αλγορίθμων της μηχανικής μάθησης είναι ανάλογη ως προς τον αριθμό των διαστάσεων.

Μηχανική Μάθηση-Μηχανική Χαρακτηριστικών

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι μεγάλες διαστάσεις σχετίζονται στενά με μια άλλη πρόκληση: τη μηχανική των χαρακτηριστικών.
- Αυτή είναι η διαδικασία δημιουργίας χαρακτηριστικών, που συνήθως χρησιμοποιούν υπάρχουσα γνώση, για να καταστεί η μηχανική μάθηση καλύτερη.
- Πράγματι, η επιλογή των πλέον κατάλληλων χαρακτηριστικών είναι ένα από τα πλέον χρονοβόρα καθήκοντα της προ επεξεργασίας στη μηχανική μάθηση.

- Καθώς το σύνολο δεδομένων αναπτύσσεται τόσο κατακόρυφα όσο και οριζόντια, γίνεται πιο δύσκολο να δημιουργηθούν νέα, ιδιαίτερα συναφή χαρακτηριστικά.

- Κατά συνέπεια, με τρόπο παρόμοιο με το μέγεθος, καθώς το μέγεθος του συνόλου δεδομένων αυξάνεται, έτσι και η δυσκολία συνδέεται με τη μηχανική των χαρακτηριστικών.

- Η μηχανική του σχεδίου σχετίζεται με την επιλογή των χαρακτηριστικών: ενώ η μηχανική των χαρακτηριστικών δημιουργεί νέα χαρακτηριστικά σε μια προσπάθεια βελτίωσης της μηχανικής μάθησης.

Μηχανική Μάθηση-Επεξεργασία Πραγματικού Χρόνου

Πανεπιστήμιο Δυτικής Μακεδονίας

- Όπως συμβαίνει με την ήδη υπάρχουσα πρόκληση της διαθεσιμότητας δεδομένων, οι παραδοσιακές προσεγγίσεις μηχανικής μάθησης δεν σχεδιάζονται για να χειρίζονται σταθερές ροές δεδομένων.
- Αυτό οδηγεί σε μια άλλη πρόκληση την ανάγκη επεξεργασίας σε πραγματικό χρόνο.
- Αυτό είναι διαφορετικό από την πρόκληση της διαθεσιμότητας δεδομένων και αποτελεί διαφορετική πρόκληση.

- Η διαθεσιμότητα των δεδομένων αναφέρεται στην ανάγκη να ενημερωθούν τα μοντέλα μνήμης.

- Οι επεξεργασίες σε πραγματικό χρόνο σχετίζονται με την χρονική επεξεργασία δεδομένων ταχείας λήψης.

- Οι προγραμματιστές αλγορίθμων, είναι αυτοί που διαχειρίζονται τα συστήματα ανίχνευσης απάτης και συστήματα παρακολούθησης, είναι αυτοί που ψάχνουν να δώσουν λύση στο πρόβλημα.

- Η σημασία της επεξεργασίας σε πραγματικό χρόνο στη σημερινή εποχή των αισθητήρων, των κινητών συσκευών και του IoT έχει οδηγήσει στην εμφάνιση πολλών συστημάτων ροής που ασχολούνται με τα μεγάλα δεδομένα.

- Έχουμε ως παραδείγματα το Twitter και το Yahoo.

- Παρόλο που τα συστήματα αυτά έχουν μεγάλη επιτυχία στην πραγματική επεξεργασία του χρόνου, δεν περιλαμβάνουν πολύπλοκα διαφορετικά μοντέλα μηχανικής μάθησης.

- Ωστόσο μπορούν να προσφέρουν τα χαρακτηριστικά της μηχανικής μάθησης χρησιμοποιώντας εξωτερικές γλώσσες ή εργαλεία.
- Υπάρχει η ανάγκη να συγχωνευθούν αυτές οι λύσεις συνεχούς ροής με μηχανικούς υπολογιστές που θα παράσχουν στιγμιαία αποτελέσματα.
- Ωστόσο, η πολυπλοκότητα τέτοιων αλγορίθμων και η αραιή διαθεσιμότητα λύσεων μηχανικής μάθησης το καθιστούν ένα δύσκολο καθήκον.

Big data στη βιομηχανία

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η ιδέα των μεγάλων δεδομένων στο τομέα της βιομηχανίας δεν είναι σίγουρα κάτι νέο.

- Όσο οι βιομηχανίες και οι επιχειρήσεις συλλέγουν πληροφορίες σχετικά με τις επιχειρηματικές τους διαδικασίες, τους πελάτες τους, τις προοπτικές τους και τα προϊόντα τους, υπάρχει ένα μεγάλο πρόβλημα στη διαχείριση των δεδομένων.

- Δεν ήταν απλώς οικονομικό ή πρακτικό για τις εταιρείες να είναι σε θέση να διαχειρίζονται αποτελεσματικά όλα τα δεδομένα στις οργανώσεις τους.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ως εκ τούτου, τα τελευταία 30 χρόνια, οι εταιρείες έπρεπε να κάνουν συμβιβασμούς.

- Είτε οι επαγγελματίες διαχείρισης δεδομένων θα πρέπει να συμβιβαστούν αποθηκεύοντας μόνο στιγμιότυπα δεδομένων ή θα πρέπει να δημιουργήσουν ξεχωριστές βάσεις δεδομένων για την αποθήκευση τμημάτων δεδομένων.

- Οι εταιρείες έχουν δοκιμάσει πολύπλοκες λύσεις για να προσπαθήσουν να ενσωματώσουν δεδομένα και με αυτό τον τρόπο να βελτιώσουν τη λήψη των επιχειρηματικών αποφάσεων.

- Αυτό συχνά απαιτούσε από τους προγραμματιστές να αναπτύξουν πολύπλοκα προγράμματα για να δημιουργήσουν τη σωστή επιχειρησιακή διαχείριση των δεδομένων.

- Οι παράγοντες που παρεμποδίζουν τη διατήρηση της επιχειρηματικής αξίας από τα δεδομένα τους ήταν ποικίλες και περίπλοκες.

- Έτσι έπρεπε να βρεθούν άμεσες λύσεις για την καλύτερη διαχείριση τους.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι παράγοντες που παρεμποδίζουν τη διατήρηση της επιχειρηματικής αξίας από τα δεδομένα είναι ποικίλοι και περίπλοκοι.

- Αυτοί οι παράγοντες περιλάμβαναν:

1. *Το κόστος της αγοράς αρκετών συστημάτων και αποθήκευσης για τη φυσική συγκράτηση των δεδομένων.*
2. *Το πρόβλημα της διαχείρισης μιας βάσης δεδομένων που ήταν υπερβολικά μεγάλη ώστε να μπορεί να διαχειρίζεται, να υποστηρίζεται ή να αναζητά.*
3. *Η ήδη υπάρχουσα τεχνολογία για τη διαχείριση της ποικιλίας των δεδομένων με τη σωστή ταχύτητα.*

Πανεπιστήμιο Δυτικής Μακεδονίας

- Διαφορετικές εταιρείες σε διαφορετικές βιομηχανίες πρέπει να διαχειρίζονται τα δεδομένα τους με διαφορετικό τρόπο.

- Ωστόσο, ορισμένα κοινά επιχειρηματικά ζητήματα βρίσκονται στο επίκεντρο του τρόπου με τον οποίο τα μεγάλα δεδομένα θεωρούνται σημαντικά για την επιχειρηματική στρατηγική.

- Ενώ οι περισσότερες επιχειρήσεις διαθέτουν μηχανισμούς για την παρακολούθηση των αλληλεπιδράσεων των πελατών, είναι πολύ πιο δύσκολο να προσδιοριστούν οι σχέσεις μεταξύ πολλών πηγών δεδομένων για να κατανοήσουν τις μεταβαλλόμενες απαιτήσεις των πελατών.

- Η μεγαλύτερη πρόκληση για την επιχείρηση είναι να είναι σε θέση να εξετάσει το μέλλον και να προβλέψει τι μπορεί να αλλάξει και γιατί.
- Οι εταιρείες θέλουν να είναι σε θέση να λαμβάνουν αποφάσεις με ταχύτερο και αποτελεσματικότερο τρόπο.
- Η επιχείρηση θέλει να εφαρμόσει αυτή τη γνώση για να αναλάβει δράση που μπορεί να αλλάξει τα αποτελέσματα των επιχειρήσεων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι ηγέτες πρέπει επίσης να κατανοήσουν τις επιπτώσεις των επιχειρηματικών σχεδίων τους που αφορούν τα προϊόντα τους.
- Οι επιχειρήσεις υιοθετούν μια ολιστική προσέγγιση στα δεδομένα.
- Τέσσερα στάδια αποτελούν μέρος της διαδικασίας σχεδιασμού που εφαρμόζεται στα μεγάλα δεδομένα: *σχεδιασμός, εκτέλεση, έλεγχος και δράση.*

Πανεπιστήμιο Δυτικής Μακεδονίας

1. *Προγραμματισμός των δεδομένων*:

- Με τον όγκο των δεδομένων που είναι διαθέσιμα στην επιχείρηση, υπάρχουν κίνδυνοι για την πραγματοποίηση υποθέσεων βάσει μιας ενιαίας προβολής δεδομένων.

- Ο μόνος τρόπος για να γίνει αυτό είναι ότι οι ηγέτες των επιχειρήσεων να να έχουν μια συνολική εικόνα για τα στοιχεία της επιχείρησης και να έχουν μια σαφή κατανόηση του τρόπου με τον οποίο αυτές οι πηγές δεδομένων σχετίζονται.

Πανεπιστήμιο Δυτικής Μακεδονίας

2. *Διεξαγωγή της ανάλυσης*:

- Αφού η επιχείρηση κατανοήσει τους επιχειρηματικούς στόχους, είναι καιρός να ξεκινήσει η ανάλυση των δεδομένων ως μέρος της διαδικασίας σχεδιασμού.

- Αυτό δεν είναι μια αυτόνομη διαδικασία.

- Η εκτέλεση σε μεγάλη ανάλυση δεδομένων απαιτεί την εκμάθηση ενός συνόλου νέων εργαλείων και νέων δεξιοτήτων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Πολλοί οργανισμοί θα πρέπει να προσλάβουν επιστήμονες δεδομένων που μπορούν να κατανοήσουν πώς να διαχειριστούν αυτό το τεράστιο όγκο διαφορετικών δεδομένων και να αρχίσουν να κατανοούν πώς σχετίζονται όλα τα στοιχεία δεδομένων στο πλαίσιο του επιχειρηματικού προβλήματος ή ευκαιρίας.

- Η μεγάλη αγορά αναλυτικών δεδομένων είναι πολύ ιδιαίτερη.

- Έτσι είναι απαραίτητο να βρεθούν εξειδικευμένοι επαγγελματίες και επιστήμονες της πληροφορικής.

Πανεπιστήμιο Δυτικής Μακεδονίας

3) *Έλεγχος των αποτελεσμάτων*:

- Πολλές εταιρείες που χρησιμοποιούν πηγές δεδομένων θα χρειαστούν χρόνο για να ελέγξουν την ποιότητα των μεγάλων δεδομένων.

- Όταν σχεδιάζουν και λαμβάνουν αποφάσεις οι επιχειρήσεις βάσει ανάλυσης, πρέπει να είναι σίγουρες ότι έχουν ισχυρή βάση.

Πανεπιστήμιο Δυτικής Μακεδονίας

4. *Δράση*:

- Αφού ολοκληρωθεί η ανάλυση, είναι καιρός να τεθεί σε εφαρμογή το σχέδιο.
- Ωστόσο, οι ενέργειες πρέπει να αποτελούν μέρος ενός γενικού κύκλου σχεδιασμού που επαναλαμβάνεται ειδικά καθώς οι αγορές γίνονται πιο δυναμικές.
- Κάθε φορά που μια επιχείρηση ξεκινά μια νέα στρατηγική, είναι σημαντικό να δημιουργηθεί συνεχώς ένας μεγάλος κύκλος αξιολόγησης μεγάλων δεδομένων.

Πανεπιστήμιο Δυτικής Μακεδονίας

- Αυτή η προσέγγιση της δράσης με βάση τα αποτελέσματα των μεγάλων αναλυτικών στοιχείων και στη συνέχεια τη δοκιμή των αποτελεσμάτων της εκτέλεσης επιχειρηματικής στρατηγικής είναι το κλειδί της επιτυχίας.

- Τα μεγάλα δεδομένα προσθέτουν το κρίσιμο στοιχείο που είναι σε θέση να αξιοποιήσει πραγματικά αποτελέσματα για να επαληθεύσει ότι μια στρατηγική λειτουργεί όπως είχε προβλεφθεί.

- Μερικές φορές τα αποτελέσματα μιας νέας στρατηγικής δεν ταιριάζουν με τις προσδοκίες της εκάστοτε επιχείρησης.

- Σε ορισμένες περιπτώσεις, αυτό σημαίνει αλλαγή στη στρατηγική.
- Σε άλλες περιπτώσεις, οι ακούσιες συνέπειες θα οδηγήσουν μια εταιρεία σε μια νέα κατεύθυνση που θα μπορούσε να έχει καλύτερα αποτελέσματα.
- Και στις 2 περιπτώσεις η χρήση των μεγάλων δεδομένων είναι ιδιαίτερα σημαντική.

Πανεπιστήμιο Δυτικής Μακεδονίας

Big data και cloud computing

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το Cloud computing είναι μια μέθοδος παροχής μιας σειράς κοινών υπολογιστικών πόρων που περιλαμβάνουν πλατφόρμες εφαρμογών, υπολογιστών, αποθήκευσης, δικτύωσης, ανάπτυξης καθώς και επιχειρηματικές διαδικασίες.

- Το Cloud computing μετατρέπει τα παραδοσιακά δεδομένα με πλούσιο υπολογιστικό υλικό σε κοινόχρηστους συνδυασμούς πόρων που βασίζονται στη λειτουργία του διαδικτύου.

- Στο cloud computing, τα πάντα, από την υπολογιστική ενέργεια έως την υποδομή υπολογιστών από τις εφαρμογές και τις επιχειρηματικές διαδικασίες μέχρι δεδομένα και αναλύσεις, μπορούν να παραδοθούν στο χρήστη ως υπηρεσία.

- Για να είναι λειτουργικό στον πραγματικό κόσμο, το σύννεφο πρέπει να έχει υλοποιηθεί με κοινές τυποποιημένες διαδικασίες και αυτοματοποιημένες μεθόδους.

- Πολλές επιχειρήσεις εκμεταλλεύονται τις υπηρεσίες cloud για τα πάντα.

- Από την δημιουργία αντιγράφων ασφαλείας με τη χρήση ειδικού λογισμικού, μέχρι και υπηρεσίες διαχείρισης σχέσεων με πελάτες.
- Με την ανάπτυξη των φορητών υπολογιστών, περισσότεροι καταναλωτές, επαγγελματίες και εταιρείες δημιουργούν και έχουν πρόσβαση σε δεδομένα με υπηρεσίες που βασίζονται σε cloud.
- Ο μέσος καταναλωτής μπορεί να βρει μια πληθώρα προϊόντων και υπηρεσιών στο cloud να τις συγκρίνει και να κάνει την έρευνα της αγοράς.

- Πολλά σενάρια βασίζονται στην υποδομή υπηρεσιών δεδομένων που βασίζεται στο σύννεφο(cloud).
- Ένα δημοφιλές παράδειγμα των πλεονεκτημάτων του cloud που υποστηρίζει μεγάλα δεδομένα μπορεί να σημειωθεί τόσο στη περίπτωση της Google όσο και στη περίπτωση του Amazon.
- Και οι δύο εταιρείες εξαρτώνται από την ικανότητα διαχείρισης μεγάλων ποσοτήτων μεγάλων δεδομένων για να προωθήσουν τις επιχειρήσεις τους.

- Αυτές οι 2 εταιρίες χρειάστηκαν να βρουν υποδομές και τεχνολογίες που θα μπορούσαν να υποστηρίξουν εφαρμογές μεγάλων δεδομένων σε τεράστια κλίμακα.

- Χαρακτηριστικό παράδειγμα είναι το Gmail και τα εκατομμύρια μηνύματα που επεξεργάζεται η Google την ημέρα ως τμήμα αυτής της υπηρεσίας.

- Η Google μπόρεσε να βελτιστοποιήσει το λειτουργικό σύστημα Linux και το περιβάλλον του λογισμικού για την υποστήριξη του ηλεκτρονικού ταχυδρομείου με τον πιο αποδοτικό τρόπο.

- Ως εκ τούτου, μπορεί να υποστηρίξει εύκολα εκατοντάδες εκατομμύρια χρήστες.

- Ακόμη πιο σημαντικό, η Google είναι σε θέση να καταγράψει και να αξιοποιήσει το τεράστιο όγκο δεδομένων τόσο για τους χρήστες ηλεκτρονικού ταχυδρομείου όσο και για τους χρήστες της μηχανής αναζήτησης για να οδηγήσει στην καλύτερη διαχείριση των δεδομένων τους.

- Ομοίως, η Amazon, με τα κέντρα δεδομένων της, έχει βελτιστοποιηθεί για να υποστηρίξει τους καθημερινούς τεράστιους φόρτους εργασίας .

- Έτσι η Amazon να συνεχίσει να προσφέρει νέες υπηρεσίες και να υποστηρίξει έναν αυξανόμενο αριθμό πελατών με τη χρήση των μεγάλων δεδομένων.

- Για να αναπτυχθούν οι δραστηριότητες της, η Amazon πρέπει να είναι σε θέση να διαχειρίζεται τα μεγάλα δεδομένα σχετικά με τα εμπορεύματά της, τους αγοραστές της και το σύστημα των εμπόρων.

- Αυτές οι εταιρείες έχουν πλέον ενεργό ρόλο σε μια σειρά υπηρεσιών βάση το σύννεφο και τη συσχέτιση με τα μεγάλα δεδομένα.

- Στην πραγματικότητα, ορισμένα χαρακτηριστικά του νέφους το καθιστούν ένα σημαντικό μέρος του μεγάλου οικοσυστήματος δεδομένων:

  - *Επεκτασιμότητα*: Η δυνατότητα επέκτασης σε σχέση με το υλικό αφορά την ικανότητα μετακίνησης από μικρές σε μεγάλες ποσότητες ισχύος επεξεργασίας με την ίδια αρχιτεκτονική.

  - Καθώς οι πόροι υλικού αυξάνονται. Το σύννεφο μπορεί να διαχειριστεί ευκολότερα μεγάλους όγκους δεδομένων.

- Η κατανεμημένη υπολογιστική, που αποτελεί αναπόσπαστο μέρος του μοντέλου του cloud, λειτουργεί πρακτικά ως διαίρεση και αποθήκευση των δεδομένων.

- Έτσι, εάν υπάρχουν τεράστιοι όγκοι δεδομένων, μπορούν να διαμοιραστούν σε διακομιστές cloud.

- Ένα επίσης σημαντικό χαρακτηριστικό είναι ότι μπορεί να κλιμακωθεί δυναμικά.

- Αυτό σημαίνει ότι εάν η υπάρξει ανάγκη για περισσότερους πόρους από τους αναμενόμενους, μπορεί εύκολα να δημιουργηθούν και άλλοι. Αυτό συνδέεται με την έννοια της ελαστικότητας και της προσαρμοστικότητας.

- Η ελαστικότητα αναφέρεται στην ικανότητα επέκτασης ή συρρίκνωσης της ζήτησης υπολογιστικών πόρων σε πραγματικό χρόνο, με βάση την ανάγκη.

- Ένα από τα πλεονεκτήματα του νέφους είναι ότι οι πελάτες έχουν τη δυνατότητα να έχουν πρόσβαση σε υπηρεσία όταν τη χρειάζονται.

- Αυτό μπορεί να είναι χρήσιμο για τα μεγάλα δεδομένα όπου μπορεί να χρειαστεί να επεκταθούν οι πόροι της πληροφορικής που χρειάζονται στην κάθε περίπτωση για να αντιμετωπιστεί ο όγκος και η ταχύτητα των δεδομένων.

- Φυσικά, αυτό το χαρακτηριστικό του cloud το καθιστά ελκυστικό για τους τελικούς χρήστες και επίσης σημαίνει ότι ο πάροχος υπηρεσιών πρέπει να σχεδιάσει μια αρχιτεκτονική πλατφόρμας που είναι βελτιστοποιημένη για αυτό το είδος υπηρεσιών.

- Οι τρεις κύριοι τύποι υπολογιστικού νέφους είναι οι IaaS, PaaS και SaaS.
- Ανάλογα με τις ανάγκες και τις χρήσεις τις κάθε εταιρείας, επιλέγεται ο κάθε τύπος διαφορετικά.
- Τελικά και οι τρεις τύποι υπολογιστικού νέφους έχουν και πλεονεκτήματα και μειονεκτήματα.

- Το PaaS είναι μια ολόκληρη υποδομή που είναι συσκευασμένη έτσι ώστε να μπορεί να χρησιμοποιηθεί για το σχεδιασμό, την υλοποίηση και την ανάπτυξη εφαρμογών και υπηρεσιών σε ένα δημόσιο ή ιδιωτικό περιβάλλον cloud.

- Το PaaS επιτρέπει σε έναν οργανισμό να αξιοποιεί βασικές υπηρεσίες χωρίς να χρειάζεται να αντιμετωπίσει την πολυπλοκότητα της διαχείρισης υλικού και λογισμικού.

Big data στο τομέα της ενέργειας

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η μείωση της κατανάλωσης ενέργειας, η εξεύρεση νέων ανανεώσιμων πηγών ενέργειας και η αύξηση της ενεργειακής απόδοσης αποτελούν σημαντικούς στόχους για την προστασία του περιβάλλοντος και τη διατήρηση της οικονομικής ανάπτυξης.

- Οι μεγάλοι όγκοι δεδομένων παρακολουθούνται όλο και περισσότερο και αναλύονται σε πραγματικό χρόνο για να βοηθήσουν στην επίτευξη αυτών των στόχων.

- Τα μεγάλα δεδομένα έχουν κάνει και εδώ την εμφάνιση τους και μπορεί να αξιοποιηθούν σημαντικά.

- Πολλοί οργανισμοί χρησιμοποιούν μια ποικιλία μέτρων για να εξασφαλίσουν ότι διαθέτουν τους ενεργειακούς πόρους που χρειάζονται τώρα και στο μέλλον.

- Οι μη παραδοσιακές πηγές ενέργειας, όπως οι ανεμογεννήτριες, οι ηλιακές εκμεταλλεύσεις και η ενέργεια των κυμάτων, γίνονται πιο ρεαλιστικές, καθώς η τιμή και η έλλειψη ορυκτών καυσίμων εξακολουθούν να προκαλούν ανησυχίες.

- Αυτοί οι οργανισμοί παράγουν και αποθηκεύουν τη δική τους ενέργεια και χρειάζονται πολλές πληροφορίες σε πραγματικό χρόνο για να ταιριάζουν με την προσφορά στη ζήτηση.

- Χρησιμοποιούν δεδομένα ροής για τη μέτρηση και την παρακολούθηση της ζήτησης και της προσφοράς ενέργειας.
- Επίσης, χρησιμοποιούνται για να βελτιώσουν την κατανόηση των ενεργειακών τους αναγκών και να λάβουν αποφάσεις σε πραγματικό χρόνο σχετικά με την κατανάλωση ενέργειας.
- Πρακτικά οι παρακολούθηση των μεγάλων δεδομένων έχει πολλές εφαρμογές.

- Ένας οργανισμός παρακολουθεί τα μεγάλα δεδομένα σχετικά με την ενεργειακή κατανάλωση και τα ενσωματώνει με δεδομένα καιρού για να κάνει προσαρμογές σε πραγματικό χρόνο στη χρήση και παραγωγή ενέργειας.

- Τα μέλη της επιστημονικής κοινότητας συλλέγουν και αναλύουν τα δεδομένα της ενέργειας. Αυτό επιτρέπει στους οργανισμούς αυτής της κοινότητας να καταναλώνουν ενέργεια πιο αποτελεσματικά και να μειώνουν το ενεργειακό κόστος.

- Τα μεγάλα δεδομένα τους επιτρέπουν να παρακολουθούν την προσφορά και τη ζήτηση και να διασφαλίζουν ότι οι μεταβολές της ζήτησης αναμένονται και διατηρούνται σε ισορροπία με τις μεταβολές της προσφοράς.

Big data στο τομέα της υγείας

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα μεγάλα δεδομένα έχουν τεράστια σημασία για τη βιομηχανία της υγειονομικής περίθαλψης.
- Συμπεριλαμβάνονται σε όλα, από τη γενετική έρευνα μέχρι την προηγμένη ιατρική απεικόνιση και την έρευνα για τη βελτίωση της ποιότητας της περίθαλψης.
- Ενώ η διεξαγωγή της ανάλυσης των μεγάλων δεδομένων σε κάθε έναν από αυτούς τους τομείς είναι σημαντική για την προώθηση της έρευνας, ένα σημαντικό πλεονέκτημα είναι η εφαρμογή αυτών των πληροφοριών στην κλινική ιατρική.

Big data στο τομέα της υγείας

Πανεπιστήμιο Δυτικής Μακεδονίας

- Εάν συγκεντρωθούν αρκετά δεδομένα, αυτά τα δεδομένα μπορούν να εφαρμοστούν πρακτικά και γρήγορα στην κατάλληλη στιγμή για να βοηθήσουν στη διάσωση ζωών.

- Οι κλινικοί ιατροί και οι ερευνητές χρησιμοποιούν τα μεγάλα δεδομένα για να επιταχύνουν τη λήψη αποφάσεων σε νοσοκομειακά περιβάλλοντα και να βελτιώνουν τα αποτελέσματα της υγειονομικής περίθαλψης για τους ασθενείς.

- Είναι πολύ σημαντικός ο ρόλος των μεγάλων δεδομένων στο τομέα της ιατρικής.

Big data στο τομέα της υγείας

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι γιατροί χρησιμοποιούν μεγάλο όγκο δεδομένων που είναι ευαίσθητα στο χρόνο όταν φροντίζουν ασθενείς, συμπεριλαμβανομένων αποτελεσμάτων εργαστηριακών εξετάσεων, αναφορών παθολογίας, ακτινών Χ και ψηφιακής απεικόνισης.

- Χρησιμοποιούν επίσης ιατρικές συσκευές για την παρακολούθηση των ζωτικών σημείων του ασθενούς, όπως η αρτηριακή πίεση, ο καρδιακός ρυθμός και η θερμοκρασία.

- Παρόλο που αυτές οι συσκευές παρέχουν ειδοποιήσεις όταν οι αναγνώσεις ξεπερνούν το κανονικό όριο, σε ορισμένες περιπτώσεις, θα μπορούσε να έχουν προληπτική δράση εάν οι γιατροί είχαν τη δυνατότητα έγκαιρης προειδοποίησης.

- Οι αλλαγές στην κατάσταση ενός ασθενούς είναι συχνά δύσκολο να παρθούν με μια φυσική εξέταση, αλλά θα μπορούσαν να ληφθούν από τις συσκευές παρακολούθησης, εάν υπήρχε ένας τρόπος για να έχουν πιο άμεση πρόσβαση στα δεδομένα.

- Οι συσκευές παρακολούθησης που χρησιμοποιούνται σε μονάδες εντατικής θεραπείας παράγουν χιλιάδες μετρήσεις ανά δευτερόλεπτο.

- Στο παρελθόν, αυτές οι αναγνώσεις γινόντουσαν κάθε 30-60 λεπτά.

- Αυτές οι συσκευές παρακολουθούσαν πολύ μεγάλους όγκους δεδομένων, αλλά λόγω του περιορισμού της τεχνολογίας, μεγάλο μέρος αυτών των δεδομένων δεν ήταν διαθέσιμο για ανάλυση.

- Στη σημερινή εποχή ωστόσο με τη χρήση των μεγάλων δεδομένων και τη γρήγορη διαχείριση και ανάλυση τους καταλαβαίνουμε ότι η ανάλυση των δεδομένων πλέον είναι πιο προσιτή.

- Και είναι σίγουρα πιο εύκολη η έγκαιρη αντιμετώπιση και η διαχείριση καταστάσεων από τους γιατρούς.

Big data στο τομέα της άθλησης

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τα περισσότερα αθλήματα έχουν πλέον εισάγει τις μεγάλες αναλύσεις δεδομένων.
- Έχουμε το εργαλείο της IBM SlamTracker για τουρνουά τένις.
- Η χρήση της ανάλυσης βίντεο που παρακολουθούν την απόδοση κάθε παίκτη σε παιχνίδι ποδοσφαίρου ή μπέιζμπολ και η τεχνολογία αισθητήρων στον αθλητικό εξοπλισμό, όπως μπάλες μπάσκετ ή γκολφ, επιτρέπουν την ανατροφοδότηση μέσω των smartphones και των διακομιστών cloud στο παιχνίδι και συμβάλλει στη βελτίωση της απόδοσης του κάθε παίχτη.

- Πολλές αθλητικές ομάδες παρακολουθούν τους αθλητές έξω από το αθλητικό περιβάλλον χρησιμοποιώντας την έξυπνη τεχνολογία για την ανάλυση της ποιότητας της διατροφής και του ύπνου.

- Από τα μεγάλα δεδομένα που παίρνουν σε πραγματικό χρόνο γίνεται καλύτερη προετοιμασία των αθλητών και αυξάνονται οι αποδώσεις τους.

- Γενικά, μπορούν να συλλέξουν δεδομένα από κάθε αντίπαλη ομάδα και να καταστρώσουν διαφορετική στρατηγική.

Big data στο τομέα της άθλησης

Πανεπιστήμιο Δυτικής Μακεδονίας

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η ιδέα της αποθήκευσης των δεδομένων προέκυψε σχεδόν πριν από 30 χρόνια.
- Η αποθήκευση των δεδομένων προοριζόταν για την επίλυση ενός μεγάλου προβλήματος για πελάτες που απαιτούσαν λύσεις σε προβλήματα.
- Όλο και περισσότερο, οι εταιρείες θέλησαν να αντικαταστήσουν τα αναποτελεσματικά συστήματα λήψης των αποφάσεων με ένα πιο εξορθολογισμένο μοντέλο.

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι εταιρείες ήθελαν να είναι σε θέση να έχουν ένα ενιαίο αρχιτεκτονικό μοντέλο που θα έκανε πολύ πιο εύκολη τη λήψη επιχειρηματικών αποφάσεων.
- Αυτή η προσέγγιση, απαιτούσε την αποθήκευση των μεγάλων δεδομένων.
- Ωστόσο, με την εμφάνιση μεγάλων δεδομένων, η έννοια της αποθήκευσης των δεδομένων αλλάζει, ώστε να μπορεί να εφαρμοστεί στη σύγχρονη πραγματικότητα.

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η παραδοσιακή αποθήκευση των μεγάλων δεδομένων θα συνεχίσει να επιβιώνει και να ευδοκιμεί επειδή είναι πολύ χρήσιμη στην ανάλυση των δεδομένων για τη λήψη αποφάσεων.

- Ωστόσο, οι τύποι αποθήκευσης των δεδομένων θα προσαρμόζονται στα μεγάλα δεδομένα.

- Γενικά, υπάρχει μια μετάβαση και ένας εκσυγχρονισμός του συστήματος αποθήκευσης.

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Σε αντίθεση με τα παραδοσιακά λειτουργικά συστήματα και τις εφαρμογές των βάσεων δεδομένων, η αποθήκευση των δεδομένων χρησιμοποιήθηκε κυρίως από τις επιχειρήσεις.

- Ενεργό ρόλο όμως είχαν και οι οικονομικοί αναλυτές για να βοηθήσουν στη λήψη αποφάσεων σχετικά με την κατεύθυνση μιας επιχειρηματικής στρατηγικής.

- Τα δεδομένα έπρεπε να συγκεντρωθούν από πολλές πηγές σχεσιακών βάσεων δεδομένων και έπειτα να διασφαλιστούν ότι τα μεταδεδομένα ήταν ακριβή.

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Τυπικά μπορούμε να πούμε ότι έχει καθιερωθεί μια σειρά αρχών της αποθήκευσης των μεγάλων δεδομένων, η οποία περιλαμβάνει τα ακόλουθα χαρακτηριστικά:

1. *Θα πρέπει να είναι προσαρμοσμένη στο θέμα.*
2. *Θα πρέπει να οργανώνεται έτσι ώστε οι σχετικές ενότητες να συνδέονται μεταξύ τους.*
3. *Οι πληροφορίες πρέπει να είναι σωστά διεκπεραιωνόμενες ώστε να μην μπορούν να αλλάξουν κατά λάθος.*

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η αποθήκευση των δεδομένων υποστηρίζει συνήθως δομημένα δεδομένα και έχουν στενά συνδεθεί με τα επιχειρησιακά συστήματα της επιχείρησης.

- Ωστόσο αυτά τα συστήματα βρίσκονται τώρα στο επίκεντρο σημαντικών αλλαγών καθώς οι οργανισμοί προσπαθούν να επεκτείνουν και να τροποποιήσουν την αποθήκευση των δεδομένων έτσι ώστε να εκσυγχρονίζονται στον νέο κόσμο των μεγάλων δεδομένων.

Παραδοσιακή αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

Αρχιτεκτονική αποθήκευσης

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η αρχιτεκτονική μιας αποθήκευσης δεδομένων είναι το κλειδί για την ανάλυση των δεδομένων και την αποθήκευση τεράστιου όγκου πληροφοριών.

- Η αρχιτεκτονική αποθήκευσης μεγάλων δεδομένων αποτελείται από τα ακόλουθα:

  - *Συστήματα προέλευσης*, που πρακτικά αυτά αντιπροσωπεύουν τις διάφορες πηγές για την αποθήκευση των δεδομένων.

Αρχιτεκτονική αποθήκευσης

Πανεπιστήμιο Δυτικής Μακεδονίας

- *Μετακίνηση δεδομένων*: πρακτικά αυτό αντιπροσωπεύει τις κύριες τεχνικές προγραμματισμού για τη μετακίνηση δεδομένων εντός του οικοσυστήματος.

- *Βάσεις δεδομένων*: υπάρχουν διάφορες βάσεις δεδομένων που δημιουργούνται και αναπτύσσονται μέσα στην αποθήκευση των δεδομένων.

- *Στάδια προγραμματισμού*: αυτές οι βάσεις δεδομένων αποτελούν τον κύριο ρόλο της επεξεργασίας και της προ επεξεργασίας για όλα τα δεδομένα που πρέπει να μεταφερθούν στην αποθήκευση των δεδομένων.

- Η λειτουργική αποθήκευση δεδομένων αντιπροσωπεύει τη δομή της βάσης δεδομένων και χρησιμοποιείται ως εργαλείο για την καθημερινή επεξεργασία δεδομένων.

- *Datamarts*: αυτές είναι ειδικές βάσεις δεδομένων που σχεδιάζονται και αναπτύσσονται για χρήση από συγκεκριμένες επιχειρησιακές μονάδες.

- Οι επιχειρήσεις κατασκευάζουν πολλαπλές βάσεις datamarts και τις ενσωματώνουν στην αποθήκευση των δεδομένων.

Αρχιτεκτονική αποθήκευσης

Πανεπιστήμιο Δυτικής Μακεδονίας

- *Αναλυτική βάση δεδομένων*: αυτές είναι βάσεις δεδομένων που εξάγουν ή αντιγράφουν δεδομένα από την αποθήκευση των δεδομένων και υποστηρίζουν αναλυτικές πλατφόρμες για την εξόρυξη των δεδομένων.

- Οι αναλυτικές βάσεις δεδομένων αναπτύσσονται σε συνδυασμό με τις τεχνολογίες υποδομής, όπως:
  - Τεχνολογίες βάσεων δεδομένων που τυπικά είναι οι παραδοσιακές μέθοδοι αποθήκευση των δεδομένων και αναπτύσσονται σε τεχνολογίες RDBMS όπως Oracle, SQL Server, DB2 και Teradata.

- *Δίκτυα*: Εταιρικές συνδέσεις δικτύου βασισμένες σε τεχνολογία οπτικών ινών
- Hardware αποθήκευσης:
  - Το SAN (δικτυακό χώρος αποθήκευσης) είναι ο πιο κοινός τρόπος αποθήκευσης.
  - Όλα τα δεδομένα μοιράζονται συνήθως με το SAN σε μια κατανομή πληροφοριών.
  - Μικρότερες αποθήκες δεδομένων μπορούν να αποθηκευτούν σε NAS.

Νέα προσέγγιση στην αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Στην αποθήκευση των δεδομένων, συχνά εντοπίζεται ένα συνδυασμός από πίνακες σχεσιακών βάσεων δεδομένων, αρχεία και μια μεγάλη ποικιλία από πηγές.

- Μια άρτια κατασκευασμένη μέθοδος αποθήκευσης μεγάλων δεδομένων, σχεδιάζεται έτσι ώστε τα δεδομένα να μετατραπούν σε μια κοινή μορφή, επιτρέποντας την επεξεργασία των ερωτημάτων με ακρίβεια και συνέπεια.

- Τα εξαγόμενα αρχεία πρέπει να μετατραπούν ώστε να ταιριάζουν με τους κανόνες και τις διαδικασίες που έχουν σχεδιαστεί για την αποθήκευση των δεδομένων.

Νέα προσέγγιση στην αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Είναι προτιμότερο να υπάρχουν διαδικασίες και κανόνες εντός της αποθήκευσης των δεδομένων για να επιβεβαιωθεί ότι οι υπολογισμοί είναι ακριβείς.

- Ενώ αυτές οι ιδέες είναι θεμελιώδεις για την αποθήκευση των δεδομένων, είναι επίσης μια βασική αρχή των κανόνων αποθήκευσης των μεγάλων δεδομένων.

- Τα δεδομένα πρέπει να εξαχθούν από τις μεγάλες πηγές δεδομένων, ώστε αυτές οι πηγές να μπορούν να συνεργαστούν με ασφάλεια και να παράγουν ουσιαστικά αποτελέσματα.

Νέα προσέγγιση στην αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η αποθήκευση των πληροφοριών των μεγάλων δεδομένων είναι διαφορετική από αυτή που υπάρχει σε ένα παραδοσιακό μοντέλο αποθήκευσης δεδομένων.
- Στην κλασσική αποθήκευση δεδομένων, μετά την κωδικοποίηση των δεδομένων, αυτά δεν αλλάζουν ποτέ.
- Μια τυπική αποθήκευση δεδομένων παράγει δεδομένα με βάση την ανάγκη να αναλυθεί ένα συγκεκριμένο ζήτημα που απαιτεί παρακολούθηση, ανάλυση και λύση.

Νέα προσέγγιση στην αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η αποθήκευση των πληροφοριών μπορεί να διαφέρει δραματικά στα μεγάλα δεδομένα.
- Η κατανεμημένη δομή των μεγάλων δεδομένων οδηγεί συχνά τους χρήστες να αποθηκεύσουν πρώτα τα δεδομένα και στη συνέχεια να εκτελέσουν την εξαγωγή και τον μετασχηματισμό τους ανάλογα με την ανάγκη τους.
- Οπότε απαιτείται διαφορετική αντιμετώπιση όσων αφορά την αποθήκευση τους.

Νέα προσέγγιση στην αποθήκευση των δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

Εκσυγχρονισμός της αποθήκευσης των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Είναι χρήσιμο να σκεφτούμε τις ομοιότητες και τις διαφορές μεταξύ του τρόπου διαχείρισης των δεδομένων στην παραδοσιακή αποθήκευση δεδομένων, όταν η αποθήκευση συνδυάζεται με μεγάλα δεδομένα.

- Οι ομοιότητες μεταξύ των δύο μεθόδων διαχείρισης δεδομένων περιλαμβάνουν:

1. *Απαιτήσεις για κοινούς ορισμούς δεδομένων.*
2. *Απαιτήσεις για την εξαγωγή και μετατροπή βασικών πηγών δεδομένων.*

- Οι διαφορές μεταξύ της παραδοσιακής αποθήκευσης των δεδομένων και των μεγάλων δεδομένων περιλαμβάνουν:

1. *Το κατανεμημένο υπολογιστικό μοντέλο μεγάλων δεδομένων που είναι απαραίτητο για την αποθήκευση των δεδομένων.*
2. *Η ανάλυση των μεγάλων δεδομένων αποτελεί το επίκεντρο των νέων τεχνολογιών, ενώ η παραδοσιακή αποθήκευση των μεγάλων δεδομένων χρησιμοποιείται για να προσθέσει ιστορικό πλαίσιο και τις βάσεις για τον εκσυγχρονισμό των μεθόδων αποθήκευσης.*

Εκσυγχρονισμός της αποθήκευσης των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Με την εμφάνιση των μεγάλων δεδομένων, τα μοντέλα ανάπτυξης και διαχείρισης δεδομένων αλλάζουν.
- Η παραδοσιακή αποθήκευση δεδομένων τυπικά υλοποιείται σε ένα και μόνο μεγάλο σύστημα εντός της βάσης δεδομένων.
- Το κόστος αυτού του μοντέλου οδήγησε τους οργανισμούς να βελτιστοποιήσουν αυτές τις μεθόδους αποθήκευσης και να περιορίσουν το μέγεθος των δεδομένων που διαχειρίζονται.

Εκσυγχρονισμός της αποθήκευσης των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ωστόσο, όταν οι οργανισμοί θέλουν να εκμεταλλευτούν την τεράστια ποσότητα πληροφοριών που παράγονται από τις μεγάλες πηγές δεδομένων, τα παραδοσιακά μοντέλα δεν λειτουργούν πλέον.

- Ως εκ τούτου, η αποθήκευση των δεδομένων έχει γίνει μια μέθοδος δημιουργίας ενός βελτιστοποιημένου περιβάλλοντος για τη στήριξη της μετάβασης στη διαχείριση νέων πληροφοριών.

- Επιπλέον η διαχείριση των νέων πληροφοριών και ο όγκος της πληροφορίας απαιτεί νέα προσέγγιση στην αποθήκευση των δεδομένων.

Εκσυγχρονισμός της αποθήκευσης των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Όταν οι επιχειρήσεις πρέπει να συνδυάσουν τη δομή της αποθήκευσης δεδομένων με τα μεγάλα δεδομένα, το μοντέλο του υλικού αποθήκευσης αποτελεί συνήθως τη λύση.

- Συνήθως, το υλικό αποθήκευσης είναι ένα ολοκληρωμένο σύστημα που ενσωματώνει το υλικό (hardware) το οποίο είναι βελτιστοποιημένο για την αποθήκευση και τη διαχείριση δεδομένων .

- Αυτό πρακτικά μπορεί να είναι ένα δωμάτιο με servers η κάποιο άλλο μέσω αποθήκευσης και ανάλυσης των μεγάλων δεδομένων.

Εκσυγχρονισμός της αποθήκευσης των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Επειδή αναφερόμαστε σε υλικό, οι συσκευές μπορούν να είναι σχετικά εύκολο και γρήγορο να εφαρμοστούν, καθώς και να προσφέρουν χαμηλότερο κόστος λειτουργίας και συντήρησης.

- Επίσης το υλικό μπορεί να είναι προ φορτωμένο με μια σχεσιακή βάση δεδομένων, όπως το πλαίσιο του Hadoop, MapReduce και πολλά από τα εργαλεία που βοηθούν στη λήψη και οργάνωση δεδομένων από διάφορες πηγές.

- Μπορεί να περιλαμβάνει επίσης τις αναλυτικές μηχανές και τα εργαλεία για την την διαδικασία ανάλυση των δεδομένων από τις πολλαπλές πηγές.

Εκσυγχρονισμός της αποθήκευσης των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

Το cloud στην αποθήκευση των big data

Πανεπιστήμιο Δυτικής Μακεδονίας

- Το νέφος (cloud) γίνεται μια συναρπαστική πλατφόρμα για τη διαχείριση των μεγάλων δεδομένων και μπορεί να χρησιμοποιηθεί σε μια πληθώρα εφαρμογών αποθήκευσης.

- Ορισμένες από τις νέες καινοτομίες στην αποθήκευση και τη μεταφορά των δεδομένων ορίζουν το cloud ως μια από τις μεγαλύτερες πλατφόρμες αποθήκευσης δεδομένων.

- Η ευκολία στην πρόσβαση σε πραγματικό χρόνο και χωρίς να απαιτείται φυσικός χώρος ή υλικό αποθήκευσης το καθιστά ιδιαιτέρως δημοφιλές.

- Η Aspera, μια εταιρεία που ειδικεύεται στη γρήγορη μεταφορά δεδομένων μεταξύ δικτύων, συνεργάζεται με την Amazon για να προσφέρει υπηρεσίες διαχείρισης δεδομένων στο cloud.

- Άλλες εταιρίες όπως η FileCatalyst και η Data Expedition επικεντρώνονται επίσης στην τεχνολογία του cloud.

- Στην ουσία, αυτή η κατηγορία τεχνολογίας αξιοποιεί το δίκτυο και το βελτιστοποιεί με σκοπό τη μετακίνηση των αρχείων.

Το μέλλον της αποθήκευσης των μεγάλων δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η τεχνολογία της αποθήκευσης των δεδομένων έχει πράγματι αρχίσει να αλλάζει και να εξελίσσεται με την έλευση των μεγάλων δεδομένων.

- Στο παρελθόν, δεν ήταν οικονομικό για τις επιχειρήσεις να αποθηκεύουν το τεράστιο όγκο δεδομένων από ένα μεγάλο αριθμό πηγών σε φυσικούς χώρους αποθήκευσης όπως σκληρούς δίσκους.

- Η έλλειψη αποδοτικών και πρακτικών αρχιτεκτονικών κατανεμημένων μεθόδων αποθήκευσης σήμαινε ότι πρέπει να σχεδιαστεί μια νέα μέθοδος αποθήκευσης των δεδομένων, ώστε να μπορεί να βελτιστοποιηθεί και να εξελιχθεί το σύστημα της αποθήκευσης.

Το μέλλον της αποθήκευσης των μεγάλων δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Οι τωρινές μέθοδοι αποθήκευσης των δεδομένων δημιουργήθηκαν με σκοπό μόνο την αποθήκευση των δεδομένων και τον έλεγχο τους.
- Επιπλέον, η αποθήκευση των δεδομένων απαιτούσε τον προσεχτικό έλεγχο ώστε τα δεδομένα να προσδιοριστούν και να διορθωθούν με ακρίβεια.
- Αυτή η προσέγγιση έχει καταστήσει την αποθήκευση των δεδομένων ακριβής και χρήσιμη αλλά εντούτοις στάσιμη σαν τεχνολογία.

Το μέλλον της αποθήκευσης των μεγάλων δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Ωστόσο, το ίδιο επίπεδο ελέγχου και ακρίβειας κατέστησε δύσκολη την δημιουργία ενός περιβάλλοντος που μπορεί να αξιοποιήσει πολύ πιο δυναμικές μεγάλες πηγές δεδομένων.
- Η αποθήκευση των μεγάλων δεδομένων εξελίσσεται αργά.
- Θα χρειαστεί αρκετό χρόνο και διαρκής παρακολούθηση των νέων τεχνολογιών.

Το μέλλον της αποθήκευσης των μεγάλων δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- Η αποθήκευση των δεδομένων και οι βάσεις δεδομένων θα συνεχίσουν να βελτιστοποιούνται.
- Ωστόσο, οι υπάρχουσες τεχνολογίες και οι μέθοδοι συνδυάζουν τα δομημένα συστήματα δεδομένων με διαφορετικές προσεγγίσεις και μεθόδους.
- Η αποθήκευση πλέον θα παρέχει τη δυνατότητα ανάλυσης τεράστιων όγκων δεδομένων σε σχεδόν πραγματικό χρόνο.

Το μέλλον της αποθήκευσης των μεγάλων δεδομένων

Πανεπιστήμιο Δυτικής Μακεδονίας

- P. B. Dongre and L. G. Malik, ‘‘A review on real time data stream classification and adapting to various concept drift scenarios,’’ in Proc. IEEE Int. Adv. Comput. Conf. (IACC).
- J. D. D. Lavaire, A. Singh, M. Yousef, S. Singh, and X. Yue, ‘‘Dimensional scalability of supervised and unsupervised concept drift detection: An empirical study,’’ in Proc. IEEE Int. Conf. Big Data (Big Data), Oct. 2015.
- A. Clauset, ‘‘A brief primer on probability distributions,’’ Santa Fe Inst, Santa Fe, NM, USA 2014.
- Z. Ghahramani, ‘‘Probabilistic machine learning and artificial intelligence,’’ Nature 2013.

Βιβλιογραφικές Αναφορές-Πηγές

Πανεπιστήμιο Δυτικής Μακεδονίας

Βιβλιογραφικές Αναφορές-Πηγές

- M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao, ‘‘Learning classifiers when the training data is not IID,’’ in Proc. 20th Int. joint Conf. Artif. Intell. (IJCAI), 2007.
- J. Wang, D. Crawl, S. Purawat, M. Nguyen, and I. Altintas, ‘‘Big data provenance:Challenges,stateoftheartandopportunities,’’inProc.IEEE Int. Conf. Big Data (Big Data), Oct. 2015.
- P. Buneman, S. Khanna, and W.-C. Tan, ‘‘Data provenance: Some basic issues,’’ in FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science. Berlin, Germany: Springer, 2000.
- H. Park, R. Ikeda, and J. Widom, ‘‘RAMP: A system for capturing and tracing provenance in MapReduce workflows,’’ Proc. VLDB Endowment, vol. 4. 2008

Πανεπιστήμιο Δυτικής Μακεδονίας

Βιβλιογραφικές Αναφορές-Πηγές

- Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman - Big Data For Dummies-Wiley (2013).
- RCoreTeam, R:A Language and Environment for Statistical Computing, Vienna, Austria, 2015, vol. 1.
- MATLAB, The MathWorks Inc., Natick, MA, USA, 2016.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, ‘‘The WEKA data mining software: An update,’’ ACM SIGKDD Explorations Newslett 2016.
- A.Labrinidis and H.V.Jagadish,‘‘ Challenges and opportunities with big data,’’ Proc. VLDB Endowment 2013.

Πανεπιστήμιο Δυτικής Μακεδονίας