Ανάκτηση Πληροφορίας

Φροντιστήριο 4

Τσιράκης Νίκος

Περιεχόμενα

- Ροές Δεδομένων

Ροές Δεδομένων

Εισαγωγή

- Οι παραδοσιακές βάσεις δεδομένων αποθηκεύουν στατικά δεδομένα χωρίς την αίσθηση του χρόνου, με εξαίρεση περιπτώσεις όπου ο χρόνος αποτελεί τμήμα επιπρόσθετο στα δεδομένα.
- Με την πάροδο των χρόνων και καθώς ο όγκος της πληροφορίας αυξανόταν γεωμετρικά υπήρξε ανάγκη για ανάλυση δεδομένων σε πραγματικό χρόνο.

“πλημμύρα” από δεδομένα

- Παράγονται όλο και περισσότερα δεδομένα:
  - Τραπεζικά, τηλεπικοινωνιακά,
  - …
  - Επιστημονικά δεδομένα: αστρονομικά, βιολογικά κλπ.
  - Κείμενα στο web κ.α.
- Αποθηκεύονται όλο και περισσότερα δεδομένα:
  - Γρήγορη και φθηνή τεχνολογία αποθήκευσης
  - Ικανά ΣΔΒΔ για μεγάλες ΒΔ

Παραδείγματα

- Το ευρωπαϊκό Very Long Baseline Interferometry (VLBI) διαθέτει 16 τηλεσκόπια, καθένα από τα οποία παράγει **1 Gigabit/second** αστρονομικά δεδομένα σε συνόδους παρατήρησης των 25 ημερών
  - η αποθήκευση και ανάλυση τέτοιου όγκου δεδομένων είναι πρόβλημα
- Ο τηλεπικοινωνιακός κολοσσός AT&T χειρίζεται δισεκατομμύρια κλήσεις/ μέρα
  - τόσο μεγάλος είναι ο όγκος των δεδομένων που αυτά **δεν** αποθηκεύονται – η ανάλυση γίνεται «στον αέρα» (on the fly)
- Η ΒΔ της επιχείρησης λιανεμπορίου Wal-Mart είναι της τάξης των 24 Tbytes
- Το UC Berkeley έκανε την εκτίμηση ότι μέσα στο 2002 παρήχθησαν 5 Exa-bytes (5 εκατομμύρια TBytes) δεδομένων !!!

Τάσεις ανάπτυξης

- Ο νόμος του Moore
  - Η ταχύτητα των υπολογιστών διπλασιάζεται κάθε 18 μήνες
- Ο νόμος της αποθήκευσης
  - Τα δεδομένα που αποθηκεύονται διπλασιάζονται κάθε 9 μήνες
- Κατά συνέπεια …
  - πολύ λίγα από αυτά τα δεδομένα μπορεί να κοιτάξει (και να αναλύσει) ο άνθρωπος
- **Άρα** χρειάζεται η ανακάλυψη γνώσης μέσα από τα δεδομένα (Knowledge Discovery in Data - KDD) για να δώσει νόημα και χρήση στα δεδομένα

Εφαρμογές Data Mining

- Market analysis and management
  - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
- Risk analysis and management
  - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
- Fraud detection (attacks) and management

- αλλά και ...
  - Intelligent query answering
  - Text / Web mining (news group, email, documents)

Ροές δεδομένων

- Οι ροές δεδομένων (data streams) είναι δεδομένα τα οποία αλλάζουν συνεχώς και με γρήγορο ρυθμό με την πάροδο του χρόνου.
- Είναι αδύνατον να γίνει έλεγχος της διάταξης με την οποία έρχονται τα αντικείμενα και επίσης να αποθηκευτεί μια ροή ολόκληρη μαζί στον ίδιο χώρο.

Ροές δεδομένων

- Ερωτήματα σε ροές μπορούν να εφαρμοστούν συνεχόμενα σε μια περίοδο χρόνου και αυξητικά επιστρέφουν νέα αποτελέσματα καθώς νέα δεδομένα έρχονται από την ροή αυτή.
  - Αυτά τα ερωτήματα είναι γνωστά και ως long-running, συνεχόμενα, standing, και μόνιμα ερωτήματα

Γενικά

- Μια **ροή δεδομένων** είναι μια ακολουθία από ψηφιακά κρυπτογραφημένα σήματα που χρησιμοποιούνται για αναπαραστίσουν την πληροφορία που μεταδίδεται (με γρήγορο ρυθμό)
- Γρήγορος ρυθμός σημαίνει πως είναι δύσκολο να:
  - Μεταδοθεί (T) η είσοδος στο πρόγραμμα
  - Υπολογιστούν (C) συναρτήσεις και διεργασίες σε μεγάλο όγκο δεδομένων εισόδου σε γρήγορο ρυθμό και να
  - Αποθηκευθούν (S)
- Είναι αναγκαίες λοιπόν οι λεγόμενες TCS απαιτήσεις

Το φαινόμενο των ροών δεδομένων

- Όταν έχουμε μετάδοση δεδομένων, αν η επικοινωνία παρουσιάζει σφάλματα τότε από τη μια πλευρά τα δεδομένα φτάνουν ορθά αλλά από την άλλη έχουν καθυστέρηση.
- Αν η υπολογιστική ισχύς είναι μικρή ή το πρόγραμμα έχει μεγάλη πολυπλοκότητα, απαιτείται χρόνος για την επεξεργασία των δεδομένων.

Παραδείγματα

- **Transactional** data streams: log interactions between entities
  - Credit card: purchases by consumers from merchants
  - Telecommunications: phone calls by callers to dialed parties
  - Web: accesses by clients of resources at servers
- **Measurement** data streams: monitor evolution of entity states
  - IP network: traffic at router interfaces
  - Sensor networks: physical phenomena, road traffic
  - Earth climate: temperature, moisture at weather stations

Βάσεις δεδομένων και Ροές Δεδομένων

Συστήματα βάσεων δεδομένων

- Μοντέλο: μόνιμες αλληλεξαρτήσεις
- Ανανέωση δεδομένων: τροποποιήσεις
- Ερωτήματα: προσωρινά
- Απαντήσεις ερωτήσεων: ακριβείς
- Αποτίμηση ερωτήσεων: αυθαίρετη
- Πλάνο ερωτημάτων: αμετάβλητο

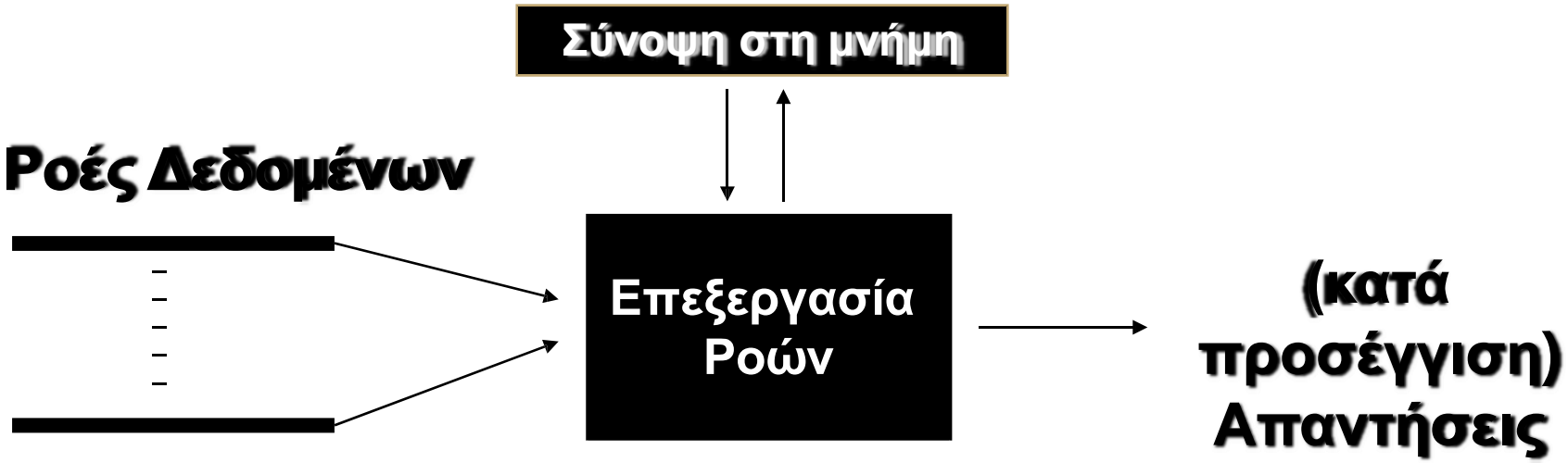
Συστήματα ροών δεδομένων

- Μοντέλο: προσωρινές transient αλληλεξαρτήσεις
- Ανανέωση δεδομένων: συμπληρώματα
- Ερωτήματα: μόνιμα
- Απαντήσεις ερωτήσεων : προσεγγιστικές
- Αποτίμηση ερωτήσεων: ενός περάσματος
- Πλάνο ερωτημάτων: προσαρμοστικό

Μια νέα ομάδα εφαρμογών

- Ένα λογισμικό που λειτουργεί σαν ενδιάμεσο μέσο (middleware) υποστηρίζοντας αποτελεσματικά εξόρυξη δεδομένων από ροές δεδομένων, και επιτρέπει πολύπλοκα ερωτήματα μειώνοντας την επιβάρυνση σε χρόνο.

Η βασική ιδέα



Δυο πεδία ανάπτυξης

- **Ανάγκη αυτόματης δημιουργίας μεγάλου όγκου δεδομένων με συνεχής ενημερώσεις**
  - Αυτό έχει ξεκινήσει τις τελευταίες δεκαετίες με πρώτο αντικείμενο δίκτυα τα οποία καταμετρούσαν τις τραπεζικές συναλλαγές και αυτές των καρτών.
  - Το διαδίκτυο είναι και αυτό ένα μεγάλο δίκτυο το οποίο έχει κατανεμημένα τόσο τις πηγές δεδομένων όσο και τους πελάτες του.Γενικά συμπεραίνουμε πως από τις διάφορες συναλλαγές δημιουργούνται πολλαπλά stream δεδομένων.

Δυο πεδία ανάπτυξης

- **Ανάγκη για ειδική ανάλυση ανανεωμένων ροών σε πραγματικό χρόνο**
  - Η ανάλυση με μια κλασική μέθοδο ανανέωσης δεδομένων είναι απλή γιατί με εφαρμογή ενός ερωτήματος πολύ απλά παίρνουμε την νέα τιμή, αυτό είναι εφαρμόσιμο ειδικά σε τραπεζικές συναλλαγές και αυτές των καρτών.
  - Σε πιο πολύπλοκες αναλύσεις όπως είναι trend analysis, forecasting κ.τ.λ. η ανάλυση γίνεται offline σε warehouses.
  - Σήμερα όμως δεν μας αρκεί αυτό γιατί υπάρχει πληθώρα από αυτόματα data feeds σε πολλούς τομείς.
  - Απαιτείται ειδική ανάλυση και πολύπλοκες εργασίες πρέπει να γίνουν. Αυτές εξαρτώνται άμεσα από το χρόνο και απαιτούν ανάλυση πραγματικού χρόνου.

Μοντέλα ροών δεδομένων

- Θεωρούμε μια ροή εισόδου a1,a2,..... η οποία φτάνει ακολουθιακά και περιγράφει το σήμα A, μια μονοδιάστατη συνάρτηση A: [1....N]→R^2. Τα μοντέλα διαφέρουν στο πως τα a_i περιγράφουν το A.
  - ***Time series Model***
  - ***Cash register Model***
  - ***Turnstile Model***

Μοντέλα ροών δεδομένων

- Απώτερος στόχος μας είναι να υπολογίζουμε διάφορες συναρτήσεις ενός σήματος A σε διαφορετικές χρονικές στιγμές όσο βλέπουμε τη ροή.
- Για αυτό το σκοπό υπάρχουν διαφορετικές μετρικές απόδοσης.
  - Χρόνος υπολογισμού ανα στοιχείο ai στην ροή (Proc. Time).
  - Ο χώρος που χρησιμοποιήθηκε για την αποθήκευση της δομής του At σε χρόνο t (Storage).
  - Ο χρόνος που χρειάστηκε για τον υπολογισμό των συναρτήσεων στο A (Compute Time).

Ερωτήματα σε ροές δεδομένων

- Τα ερωτήματα σε ροές δεδομένων έχουν πολλά κοινά με αυτά που εφαρμόζονται στα παραδοσιακά συστήματα βάσεων δεδομένων.
- Στα συστήματα ροών δεδομένων υπάρχουν δυο ιδιαίτερες διακρίσεις:
  - Η πρώτη είναι μεταξύ των στιγμιαίων και των συνεχόμενων ερωτημάτων.
  - Η δεύτερη διάκριση είναι μεταξύ των προκαθορισμένων ερωτημάτων και των “κατά περίπτωση” ερωτημάτων (ad hoc queries).

Ερωτήματα σε ροές δεδομένων – Διάκτιση Ι

- Τα στιγμιαία είναι ερωτήματα που εφαρμόζονται μια φορά σε μια χρονική περίοδο και θεωρούν τα δεδομένα σαν ένα στιγμιότυπο επιστρέφοντας την απάντηση στο χρήστη.
- Τα συνεχόμενα ερωτήματα εφαρμόζονται συνεχώς καθώς η ροή δεδομένων έρχεται.
- Τα ερωτήματα αυτά είναι πιο ενδιαφέροντα στην κατηγορία των ροών δεδομένων.
- Η απάντηση σε ένα τέτοιο ερώτημα δημιουργείται κατά το πέρασμα του χρόνου, και συνεχώς ανταποκρίνεται στη ροή που έχει περάσει μέχρι τη στιγμή της εφαρμογής του ερωτήματος.
- Οι απαντήσεις αυτές μπορούν να αποθηκεύονται και να ανανεώνονται καθώς νέα δεδομένα έρχονται.
- Ανάλογα την περίπτωση γίνεται χρήση είτε των στιγμιαίων είτε των συνεχών ερωτημάτων.

Ερωτήματα σε ροές δεδομένων – Διάκτιση ΙΙ

- Ένα προκαθορισμένο ερώτημα είναι αυτό που εφαρμόζεται σε ένα σύστημα ροών δεδομένων πριν έρθει κάποιο από τα δεδομένα.
- Είναι συνήθως συνεχόμενα ερωτήματα, αλλά και στιγμιαία ερωτήματα μπορούν να είναι προκαθορισμένα.
- Αντίθετα τα κατά περίπτωση ερωτήματα εφαρμόζονται αφού έχει γίνει η άφιξη των δεδομένων και είναι πραγματικού χρόνου ερωτήματα.
- Μπορούν να είναι είτε στιγμιαία είτε συνεχόμενα ερωτήματα.
- Τα κατά περίπτωση ερωτήματα περιπλέκουν το σχεδιασμό ενός συστήματος διαχείρισης ροών δεδομένων και αυτό γιατί δεν είναι γνωστά για τη βελτιστοποίηση των ερωτημάτων τους και γιατί συνήθως απαιτούν αναφορές σε δεδομένα που ήδη έχουν περάσει από το σύστημα.

Ένα σενάριο ροών δεδομένων

- Ο παγκόσμιος ιστός αποτελείται δρομολογητές οι οποίοι είναι συνδεδεμένοι μεταξύ τους και προωθούν IP πακέτα.
- Για να διαχειριστούμε τέτοια δίκτυα απαιτείται πραγματικού χρόνου αναφορά λαθών, χρήση προτύπων και δυνατότητα να αναφέρονται οι ασυνήθιστες συμπεριφορές.
- Αυτό απαιτεί ανάλυση της κίνησης και των εσφαλμένων δεδομένων σε πραγματικό χρόνο.

Ένα σενάριο ροών δεδομένων

- Η κίνηση στους δρομολογητές μπορούμε να την παρατηρήσουμε σε διάφορα επίπεδα:
  - Στο τελευταίο επίπεδο, έχουμε το packet log: κάθε IP πακέτο έχει ένα header που περιέχει τις IP διευθύνσεις προορισμού και αφετηρίας, ports και άλλα.
  - Σε ένα πιο υψηλό επίπεδο επαύξησης, έχουμε το flow log: κάθε ροή είναι μια συλλογή πακέτων με ίδιες τιμές για συγκεκριμένα χαρακτηριστικά όπως η IP διεύθυνση προορισμού και αφετηρίας και το log περιέχει αθροιστικές πληροφορίες σχετικά με τον αριθμό των bytes και των πακέτων που έχουν σταλεί, ώρα εκκίνησης, ώρα τερματισμού, τύπος πρωτοκόλλου κ.α.
  - Στο υψηλότερο επίπεδο, έχουμε το SNMP log: το οποίο είναι τα επαυξημένα δεδομένα του αριθμού των bytes που στάλθηκαν σε κάθε σύνδεσμο κάθε λίγα λεπτά.

Ένα σενάριο ροών δεδομένων

- Πολλά ακόμα log αρχεία μπορούν να δημιουργηθούν από IP δίκτυα.
- Κάθε φορά ανάλογα τις απαιτήσεις υπάρχουν πολλά ακόμα log ειδικού τύπου που μπορούν να χρησιμοποιηθούν.

Ένα σενάριο ροών δεδομένων

- Ορισμένα ερωτήματα που μπορεί κάποιος να κάνει στα IP logs είναι τα παρακάτω.
  - Πόση κίνηση (HTTP) έγινε σε κάποιο συγκεκριμένο σύνδεσμο από ένα ορισμένο φάσμα IP διευθύνσεων;
  - Πόσες μοναδικές IP διευθύνσεις χρησιμοποίησαν ένα συγκεκριμένο σύνδεσμο για να στείλουν την κίνηση τους από την αρχή της μέρας;
  - Ποιες είναι οι k πιο επιβαρημένες ροές κατά τη διάρκεια της μέρας;
  - Πόσες ροές αποτελούν ένα πακέτο μόνο;
  - Πόσο ποσοστό της κίνηση χτες σε δυο δρομολογητές ήταν κοινή ή παρόμοια;
  - Ποιοι είναι οι k σχετιζόμενοι σύνδεσμοι μέσα στην μέρα για ένα δοθέν μέτρο συσχέτισης;
  - Για κάθε IP διεύθυνση και κάθε πέντε λεπτών διακοπή, υπολόγισε τον αριθμό των bytes και των πακέτων που μεταφέρθηκαν.

Κατηγορίες εφαρμογών ροών δεδομένων

1. Δίκτυα αισθητήρων
   - Τα δίκτυα αισθητήρων (sensor networks) μπορούν να χρησιμοποιηθούν σε πολλές εφαρμογές ελέγχου που εκτελούν πολύπλοκες διεργασίες φιλτραρίσματος.
   - Σε αυτές τις εφαρμογές απαιτούνται πολλές φορές συνάθροιση και σύνδεση πολλαπλών ροών δεδομένων για να είναι δυνατή η ανάλυση δεδομένων από διαφορετικές πηγές.

Κατηγορίες εφαρμογών ροών δεδομένων

2. Ανάλυση κίνησης δικτύων

- Σε αυτή την κατηγορία συναντάμε συστήματα που εκτελούν αναλύσεις σχετικά με την κίνηση δικτύων σε πραγματικό χρόνο και χρησιμοποιούνται ήδη για τον υπολογισμό τέτοιων στατιστικών δεδομένων και για τον εντοπισμό ακραίων καταστάσεων (π.χ. μεγάλη συμφόρηση, αρνήσεις υπηρεσιών).
- Ο συνεχής έλεγχος ορισμένων πηγών και των προορισμών τους είναι αρκετά σημαντικός καθώς το μεγαλύτερο μέρος του bandwidth καταναλώνεται από ένα μικρό σύνολο απειλητικών χρηστών.

Κατηγορίες εφαρμογών ροών δεδομένων

3. Οικονομικά Εισιτήρια

- Η συνεχής ανάλυση χρηματιστηριακών δεδομένων εμπεριέχει, ανακάλυψη συσχετίσεων, αναγνώριση τάσεων και δυνατοτήτων κερδοσκοπίας καθώς και πρόβλεψη μελλοντικών τιμών.

Κατηγορίες εφαρμογών ροών δεδομένων

4. Ανάλυση αρχείων καταγραφής δοσοληψιών

- Συστήματα παρακολούθησης δεδομένων δοσοληψιών μέσω διαδικτύου, τηλεφωνικών εγγραφών και συστήματα αυτόματων ταμειακών συναλλαγών έχουν να κάνουν με ροές δεδομένων.
- Σκοπός των μοντέλων διαχείρισης τέτοιων ροών είναι να βρίσκουν πρότυπα συμπεριφοράς χρηστών που παρουσιάζουν ενδιαφέρον, να αναγνωρίζουν συμπεριφορές χρηστών που μπορεί να επιφέρουν απάτες και να μπορούν να προβλέψουν μελλοντικές συμπεριφορές χρηστών.

Συστήματα ροών δεδομένων

- Άμεση επεξεργασία των ροών δεδομένων
  - Το σύστημα αυτό κρατά τμήματα κάθε φορά από τη ροή δεδομένων και συνήθως χρησιμοποιεί γλώσσες προγραμματισμού όπως η C για να διαχειριστεί την πληροφορία

Συστήματα ροών δεδομένων

- Συστήματα που επιτρέπουν την μεγάλη απόδοση στη διαδικασία ανανεώσεων των βάσεων δεδομένων χρησιμοποιώντας τυποποιημένη τεχνολογία
  - Εδώ έχουμε εφαρμογές φτιαγμένες πάνω από μια βάση δεδομένων η οποία βάση διαχειρίζεται τα δεδομένα (SNPM logs – συνάθροιση δεδομένων από bytes που στάλθηκαν σε κάθε link σε ορισμένα λεπτά) και δίνει στην εφαρμογή τη δυνατότητα για δημιουργία traffic patterns σε links μεταξύ των IP routers.

Συστήματα ροών δεδομένων

- Συστήματα βάσεων δεδομένων όπου εσωτερικά είναι διαμορφωμένα να χειρίζονται ροές δεδομένων
  - Εδώ ερευνητικά το θέμα είναι ανοιχτό και μιλάμε για νέους stream operators, SQL extensions, μεθόδους δρομολόγησης. Εδώ έχουμε εξ ολοκλήρου ενα σύστημα διαχείρισης ροών δεδομένων.

Μέθοδοι σε ροές δεδομένων

- Όταν θέλουμε να αναφερθούμε σε εξόρυξη γνώσης μέσα από data streams τότε υπάρχουν πολλές μέθοδοι που χρησιμοποιούνται γι αυτό το σκοπό.
  - Κατηγοριοποίηση
    - Classic operation in machine learning and data mining
  - Συσταδοποίηση (Clustering)
    - Classic area of machine learning and pattern recognition
  - Πρότυπα ακολουθιών (Sequential Patterns)
  - Μείωση διαστάσεων (Reduction of Dimensions)

Μέθοδοι σε ροές δεδομένων

- Κατηγοριοποίηση
  - Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Συνήθως υπάρχει περιορισμός στον αριθμό των κατηγοριών γι αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές τις οποίες κατατάσσουμε σε 2 κατηγορίες
    - δέντρα αποφάσεων
    - νευρωνικά δίκτυα
  - Και οι δύο βασίζονται στην ιδέα της εκπαίδευσης με τη βοήθεια ενός υποσυνόλου δεδομένων που ονομάζεται σύνολο εκπαίδευσης. Έτσι με την εφαρμογή της διαδικασίας εκπαίδευσης καθορίζονται κάποια πρότυπα για τις κατηγορίες δεδομένων.

Μέθοδοι σε ροές δεδομένων

- Συσταδοποίηση
  - Εδώ μιλάμε για την εργασία καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων. Εδώ δεν έχουμε προκαθορισμένες κατηγορίες. Οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους.

Μέθοδοι σε ροές δεδομένων

- Πρότυπα ακολουθιών
  - Εδώ έχουμε εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικών με το χρόνο ή άλλες ακολουθίες.

Μέθοδοι σε ροές δεδομένων

- Μείωση διαστάσεων
  - Οι τεχνικές αυτές υπολογίζουν μια μικρότερη αντιπροσώπευση του αρχικού συνόλου δεδομένων. Εδώ γίνεται προσπάθεια να διατηρηθεί όσο το δυνατόν η αρχική δομή.

Συμπεράσματα

- Χρειαζόμαστε υποδομές σε TCS για να διαχειριστούμε και να επεξεργαστούμε τις ροές δεδομένων.
- Αυτό δημιουργεί νέα ανοιχτά ερευνητικά ζητήματα σε τομείς όπως:
  - Αλγόριθμοι
  - Βάσεις Δεδομένων
  - Δίκτυα
  - Συστήματα
  - Γλώσσες

Τέλος 4ου Φροντιστηρίου