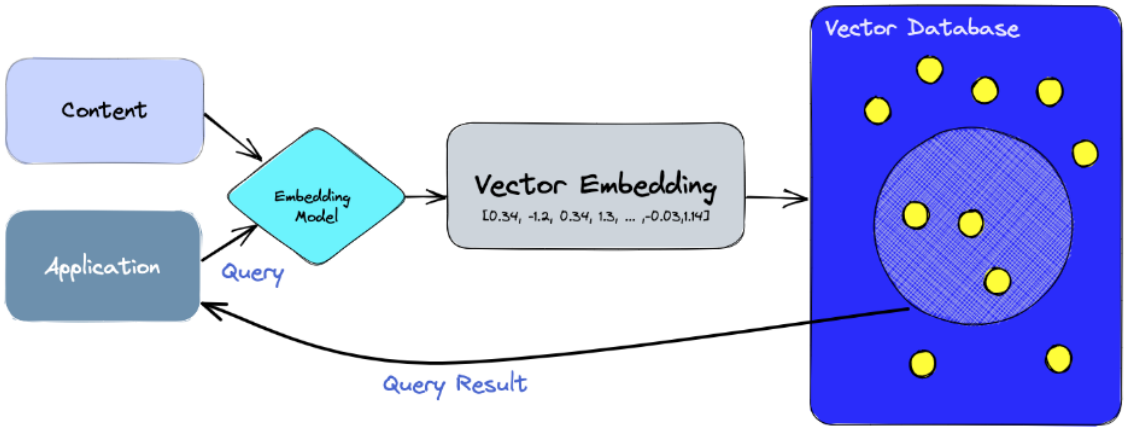
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Προτεινόμενα Θέματα Διπλωματικών Εργασιών

Θέμα 1

Τίτλος (Ελληνικά):	Χρήση Vector Database για επιτάχυνση Αναλυτικής Δεδομένων Όγκου
Τίτλος (Αγγλικά):	Vector Databases to speed up Big Data Analytics
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή:	Γ. Γκούμας, Ν. Κοζύρης
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Machine Learning
Επιθυμητές Γνώσεις:	Προγραμματισμός Συστημάτων, Προηγμένα Θέματα ΒΔ

Σύντομη περιγραφή:



Μια Διανυσματική Βάση Δεδομένων (**vector database**) είναι ένα σύστημα διαχείρισης βάσεων δεδομένων που έχει σχεδιαστεί ειδικά για την αποθήκευση, την ευρετηρίαση και την αποτελεσματική ανάκτηση διανυσματικών δεδομένων υψηλής διάστασης. Τα διανυσματικά δεδομένα (vectors) είναι αναπαραστάσεις αντικειμένων ή οντοτήτων ως διανύσματα σημείων σε έναν πολυδιάστατο χώρο. Αυτά μπορούν να προέρχονται από διάφορες πηγές, όπως η ενσωμάτωση κειμένου, εικόνων ή άλλων δεδομένων (embeddings). Οι διανυσματικές βάσεις δεδομένων παρέχουν: Αναζήτηση ομοιότητας για παρόμοια διανύσματα (π.χ., παρόμοια έγγραφα ή εικόνες), υποστήριξη για εφαρμογές μηχανικής μάθησης και υψηλά κλιμακούμενη απόδοση: Παρέχουν υψηλές επιδόσεις και κλιμακούμενη αναζήτηση ομοιότητας σε τεράστια σύνολα δεδομένων διανυσμάτων μέσω αποδοτικών δομών ευρετηρίασης και

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

στρατηγικών ευρετηρίασης.

Στην παρούσα διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα (ένα ή περισσότερα):

1. την ταχύτητα καθώς και την κλιμακωσιμότητα των μοντέρνων Διανυματικών ΒΔ (π.χ., Milvus, Weaviate, Qdrant, Vespa)
2. Σύγκριση μιας native vector DB σε σχέση με NoSQL datastores που μοπορούν να παίξουν το ρόλο της vector Database (e.g., elasticsearch, MongoDB, redis)
3. Υλοποίηση στοιχείων μιας vector DB σε υπάρχον NoSQL σύστημα (π.χ., HBASE)
4. Σύγκριση με χρήση state-of-the-art similarity benchmarks+algorithms (e.g., https://github.com/google-research/google-research/tree/master/scann , https://github.com/facebookresearch/faiss/wiki/ , https://github.com/erikbern/ann-benchmarks)

Ενδεικτική Βιβλιογραφία:

[1] https://www.pinecone.io/learn/vector-database/
[2] https://www.ibm.com/topics/vector-database

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 2

Τίτλος (Ελληνικά):	Βελτιστοποίηση απόδοσης ΒΔ με χρήση AutoML σε σχέση με υπάρχουσες ML4DB τεχνικές
Τίτλος (Αγγλικά):	DB performance optimization using AutoML vs existing ML4DB techniques
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή:	Γ. Γκούμας, Ν. Κοζύρης
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Machine Learning
Επιθυμητές Γνώσεις:	Προγραμματισμός Συστημάτων, Προηγμένα Θέματα ΒΔ

Σύντομη περιγραφή:

Η αυτοματοποιημένη μηχανική μάθηση (AutoML) είναι η διαδικασία αυτοματοποίησης των εργασιών εφαρμογής μηχανικής μάθησης σε προβλήματα του πραγματικού κόσμου. Το AutoML περιλαμβάνει δυνητικά κάθε στάδιο από την αρχή έως την κατασκευή ενός μοντέλου μηχανικής μάθησης και στοχεύει να επιτρέψει σε μη ειδικούς να κάνουν χρήση μοντέλων και τεχνικών μηχανικής εκμάθησης. Η αυτοματοποίηση της διαδικασίας προσφέρει επιπλέον τα πλεονεκτήματα της παραγωγής απλούστερων λύσεων, ταχύτερης δημιουργίας αυτών των λύσεων και μοντέλων που συχνά υπερτερούν των μοντέλων που έχουν σχεδιαστεί/επιλεγεί με μη αυτόματο τρόπο.

Το ML4DB (Machine Learning για Βάσεις Δεδομένων) είναι ένας αναδυόμενος τομέας που στοχεύει στην αξιοποίηση τεχνικών μηχανικής μάθησης για τον αυτοματισμό και τη βελτιστοποίηση διαφόρων πτυχών των συστημάτων βάσεων δεδομένων. Η βασική ιδέα πίσω από το ML4DB είναι η χρήση μοντέλων και αλγορίθμων μηχανικής μάθησης για να αντικαταστήσουν ή να ενισχύσουν τις παραδοσιακές προσεγγίσεις που βασίζονται σε κανόνες για τη διαχείριση και βελτιστοποίηση βάσεων δεδομένων. Η έννοια του ML4DB μπορεί να εφαρμοστεί σε διάφορα στάδια όπως: *Ρύθμιση Βάσεων Δεδομένων*, *Επιλογή Ευρετηρίων και Φυσικός Σχεδιασμός*, *Βελτιστοποίηση Ερωτημάτων*, *Πρόβλεψη Φορτίου Εργασίας και Διαχείριση Πόρων*, κλπ.

Μια πολύ αναλυτική παρουσίαση και σύγκριση λύσεων ML4DB έχει δημοσιευτεί πρόσφατα [2] με κώδικα και δεδομένα σύγκρισης ανοικτά [3]. Σε αυτή τη διπλωματική, καλείστε να χρησιμοποιήσετε μέρος της αξιολόγησης αυτής ώστε να συγκρίνετε υπάρχουσες λύσεις με τη δική σας που θα χρησιμοποιεί AutoML, σε ένα ή περισσότερα στάδια βελτιστοποίησης (π.χ., index selection, cardinality estimation, cost estimation, etc).

Ενδεικτική Βιβλιογραφία:

[1] https://www.automl.org/automl/
[2] https://dl.acm.org/doi/abs/10.14778/3636218.3636235?casa_token=56rG_HUGdHYAAAAA:M9L-JKpcLKzLz6uJCEtnXko1FvfsLpBRY1teTwl1iuiSeslBnKWLw9J4nSU6m6qF0AbuvLTX_BWX
[3] https://github.com/zhaoyue-ntu/qp_evaluation

Θέμα 3

Τίτλος (Ελληνικά):	Υλοποίηση Συστήματος Feature Store με χρήση Data-Lake
Τίτλος (Αγγλικά):	Design and Implementation of a modern Feature Store using Data Lake technologies
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή:	Γ. Στάμου, Α. Βουλόδημος
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Κατανεμημένα Συστήματα

Σύντομη περιγραφή:

Ένα Feature Store είναι ένα κεντρικό αποθετήριο για την αποθήκευση, διαχείριση και εξυπηρέτηση χαρακτηριστικών (features) Μηχανικής Μάθησης (ML). Η κύρια λειτουργία ενός Feature Store είναι να παρέχει έναν συνεπή και αξιόπιστο τρόπο οργάνωσης, κοινής χρήσης και επαναχρησιμοποίησης features σε έναν οργανισμό. Ορισμένες βασικές λειτουργίες και πλεονεκτήματα των Feature Stores περιλαμβάνουν την ανακάλυψη χαρακτηριστικών, την διατήρηση εκδόσεων και την παρακολούθηση γενεαλογίας τους.

Μια λίμνη δεδομένων (data lake) είναι ένα κεντρικό αποθετήριο που μας επιτρέπει να αποθηκεύουμε όλα τα δομημένα και μη δομημένα δεδομένα σε οποιαδήποτε κλίμακα. Μπορούμε να αποθηκεύσουμε τα δεδομένα ως έχουν, ωστόσο τα μοντέρνα data lakes επιτρέπουν την εκτέλεση απλών ETL επάνω στα raw data αυτά.

Στην παρούσα διπλωματική, θα σχεδιάσουμε και θα αναπτύξουμε ένα σύστημα που συνδυάζει feature store με έξυπνο και αποδοτικό τρόπο με μια λίμνη δεδομένων. Συγκεκριμένα, θα μελετηθούν τρόποι και αλγόριθμοι που θα επιτρέπουν να μεταφέρεται μέρος του υπολογισμού που απαιτείται για τον υπολογισμό ενός feature από την data lake αντί από μια μηχανή όπως το Spark ή το Ray.

Ενδεικτική Βιβλιογραφία:

[1] https://www.featurestore.org/
[2] https://feast.dev/
[3] https://oss.redis.com/redisml/
[4] https://delta.io/

Θέμα 4-5

Τίτλος (Ελληνικά):	...
Τίτλος (Αγγλικά):	...
Επιβλέπων:	Δημήτριος Τσουμάκος (Laurent d'Orazio, Univ Rennes, CNRS, IRISA)
Επιτροπή:	Δ. Πνευματικάτος, Γ. Γκούμας
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Κατανεμημένα Συστήματα

Σύντομη περιγραφή:

Digital transformation has lead to unprecedented data generation creating both opportunities and challenges. In particular, the convergence of cloud computing and Big Data technologies have attracted increasing attention during the last decades. According to the market research organization MarketsandMarkets, the global big data market was estimated to be valued at $162.6 billion in revenue in 2021 and is projected to reach $273.4 billion by 2026, growing at a CAGR of 11.0% from 2021 to 2026.
Domains of application include the Internet, social networks, healthcare, smart cities and in particular vehicles/transports, smart agriculture, telecommunication or security monitoring.

Problem
This thesis aims to propose strategies for Big Data management with respect to multi-objective optimization, in particular considering quality.

Direction 1:
Semantic caching: semantic caching makes it possible to add knowledge in a cache and thus increase the usability of its content. In the context of cloud computing in general and fog/edge computing in particular, semantic caching may be promising in order to balance the load over the global environment. Unfortunately, semantic caching may induce overhead and/or sometime be not necessary, in particular if part of the required data are absent from the cache. Nevertheless, existing results may be enough in some cases. This thesis will address this problem, trying to define some strategies in particular when, even if answers are not complete, the current quality is enough to not trigger additional processing on some data centres and/or edges.

Direction 2:
Fuzzy joins: we have proposed filters for similarity joins in a large scale environment. Our experimental results have shown that there is a clear tradeoff between the flexibility the user would like to allow in particular the semantic distance he would like to consider and the performance (the longer is the distance, the more expensive is the join process). In this thesis, the goal would be to include the concept quality into the similarity join process.

Validation:
To validate the different contributions, experiments will be conducted on Grid5000 (https://www.grid5000.fr/), a large-scale and flexible testbed for experiment-driven research in all areas of computer science, with a focus on parallel and distributed computing including Cloud, HPC and Big Data and AI. Grid5000 (1) provides access to a large amount of resources: 15000 cores, 800 compute-nodes grouped in homogeneous clusters, and featuring various technologies: PMEM, GPU, SSD, NVMe, 10G and 25G Ethernet, InfiniBand, Omni- Path; (2) is highly reconfigurable and controllable: researchers can experiment with a fully customized software stack thanks to bare-metal deployment features, and can isolate their experiment at the networking layer; (3) provides advanced monitoring and measurement features for traces collection of networking and power consumption, providing a deep understanding of

experiments; (4) is designed to support Open Science and reproducible research, with full traceability of infrastructure and software changes on the testbed and (5) gathers a community of 500+ users supported by a solid technical team.
Ενδεικτική Βιβλιογραφία:

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma