/TRP1¡¡/ca 1/CA 1¿¿

DBMS – PROJECTS

BSDS, Second Year, 2025–2026

Deadline: May 15, 2026

Total: 10 marks

SUBMISSION INSTRUCTIONS

STEP I:

1. Submit a solution sketch in a single file by the deadline. The solution must be self-explanatory.
2. The solution should include the sections (if applicable): Introduction, Related Work, Terminologies and Definitions, Theory, Methods, Results, Conclusion.
3. Include names and roll numbers of all of your group members (at most 3).
4. Naming convention for your submission file (assuming M is your project number): `projM` (.docx, .doc, .pdf, .tex, etc.).
5. To submit a solution file (say `projM.docx`), ensure that it is not password protected and mail to `malaybhattacharyya@isical.ac.in`.

STEP II:

1. Deliver a solution sketch through team presentation on (or immediately after) the deadline.
2. The presentation slides should contain the components (if applicable): Introduction, Related Work, Terminologies and Definitions, Theory, Methods, Results, Conclusion.

NOTE: The contribution must be novel and non-trivial.

Project 1: **[Conformal Approximate Nearest Neighbor Search]** The existing approaches of Approximate Nearest Neighbor (ANN) search suffer from restrictive assumptions about underlying data distributions, which limits their generalizability. A recent work has presented the first framework to provide formal, distribution-free error guarantees for Inverted File (IVF)-based ANN search by employing recent advances in Conformal Risk Control [1]. The proposed ConANN method tightly controls approximation error, provides formal guarantees without requiring expansion of the search space, dynamically adapts the cluster probes required per query, and incurs negligible overheads when compared to existing state-of-the-art baselines.

Some future scopes are suggested in this work. This includes the incorporation of confidence-aware retrieval, Applicability beyond IVF, and metric-agnostic design. Extend the ConANN method in any of these directions.

[1] Sonia Horchidan, Fabian Zeiher, Henrik Boström, and Paris Carbone. ConANN: Conformal Approximate Nearest Neighbor Search. PVLDB, 19(1):29 - 42, 2025.

Project 2: [**JSON Schema Validation**] It is important to guarantee that the semi-structured JSON input provided by clients matches a predefined structure. A recent work has introduced a JSON Schema validator, Blaze, which compiles complex schemas quickly to an efficient representation, adding minimal overhead at build time [1].

Some improvements to this work are suggested in the paper. Extend any one them in your work.

[1] Juan Cruz Viotti and Michael J. Mior. Blaze: Compiling JSON Schema for 10× Faster Validation. PVLDB, 19(2): 279 - 291, 2025.

Project 3: [**JSON Schema Validation in Streaming Setting**] A recent work has proposed an approach for performing streaming JSON validation that relies on a new class of pushdown automata that can process JSON documents in an online fashion [1].

Extend this work by incorporating new features to the proposed model.

[1] Juan Cruz Viotti and Michael J. Mior. Blaze: Compiling JSON Schema for 10× Faster Validation. PVLDB, 19(2): 279 - 291, 2025.

Project 4: [**Query Repair for Aggregate Constraints**] In many real-world scenarios, query results must satisfy domainspecific constraints. For instance, a minimum percentage of interview candidates selected based on their qualifications should be female. These requirements can be expressed as constraints over an arithmetic combination of aggregates evaluated on the result of the query. A recent work has studied how to repair a query to fulfill such constraints by modifying the filter predicates of the query [1],

Suggest improvements to this work and extend accordingly.

[1] Shatha Algarni, Boris Glavic, Seokki Lee, and Adriane Chapman. Efficient Query Repair for Aggregate Constraints. PVLDB, 19(2): 252 - 264, 2025.

Project 5: [**Clustering of Tables**] Database table clustering involves grouping semantically related tables, which simplifies many database management, analysis, and integration tasks. A recent work has presented Schuyler, a system that clusters database tables by combining structural and semantic features of the database [1]. It fine-tunes a large language model in a self-supervised way using triplet-loss to produce high-quality embeddings representing table semantics. Subsequently, these embeddings are clustered to achieve a database table clustering. This approach requires no labeled training data and, thus, is applicable to arbitrary databases.

One can extend Schuyler to further data representations, such as a topical clustering of spreadsheets or non-relational databases. Try either of them or suggest a new improvement and extend it accordingly.

[1] Lukas Laskowski, Fabian Panse, Michael Hladik, Jan Portisch, and Felix Naumann. Schuyler: Self-Supervised Clustering of Tables in Relational Databases. PVLDB, 19(4): 657 - 669, 2025.

Project 6: [**Verification of Database Isolation**] Serializability and snapshot isolation are essential for maintaining data consistency and integrity in databases. Verifying whether a database

upholds its claimed guarantees is a challenging task. A recent paper has presented VERYSTRONG, a fast verifier for strong database isolation [1]. It uses a novel formalism called hyperpolygraphs, which compactly captures both certain and uncertain transactional dependencies in database executions. Using this, they develop sound and complete encodings for verifying both serializability and snapshot isolation.

This approach currently targets key-value databases. Extending it to relational models getting inspirations from recent methods a is a promising direction [2]. Another direction is to unify correctness reasoning for both isolation guarantees and query semantics [3], drawing inspiration from recent advances like [4]. Extend the work accordingly.

[1] Zhiheng Cai, Si Liu, Hengfeng Wei, Yuxing Chen, and Anqun Pan. Fast Verification of Strong Database Isolation. PVLDB, 19(4): 563 - 575, 2025.

[2] Rui Yang, Ziyu Cui, Wensheng Dou, Yu Gao, Jiansen Song, Xudong Xie, and Jun Wei. Detecting Isolation Anomalies in Relational DBMSs. Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis, Article ISSTA076 (June 2025), 23 pages, 2025..

[3] Zijing Yin, Si Liu, and David Basin. Testing Graph Databases with Synthesized Queries. Proceedings of the ACM on Management of Data, 3(4):268, 2025.

[4] Lauren Pick, Amanda Xu, Ankush Desai, Sanjit A. Seshia, and Aws Albarghouthi. Checking Observational Correctness of Database Systems. Proceedings of the ACM on Programming Languages, 9:139, 2025.

Project 7: [**Heavy-Hitters with Optimal State Changes**] A recent work has presented a new algorithm for solving the l_k heavy hitters problem for $k \in [1, 2]$ requiring only $O(1/\epsilon n^{1-1/k}\text{polylog}(n/\epsilon)$ state changes and $O(\text{poly}(1/\epsilon)\log n)$ space [1].

Suggest improvements to this work and extend accordingly.

[1] Swartworth, W. and Woodruff, D.P., 2025. Finding Heavy-Hitters with Optimal State Changes. Proceedings of the ACM on Management of Data, 3(5), pp.1-22.

Project 8: [**Query Answering**] When query evaluation produces too many tuples, a new approach in query answering is to retrieve a diverse subset of them. The standard approach for measuring the diversity of a set of tuples is to use a distance function between tuples, which measures the dissimilarity between them, to then aggregate the pairwise distances of the set into a score (e.g., by using sum or min aggregation). Highlighting limitations of this, a recent work has introduced a novel approach for computing the diversity of tuples based on volume instead of distance [1].

Suggest improvements to this work and extend accordingly.

[1] Arenas, M., Merkl, T.C., Pichler, R. and Riveros, C., 2025. Query answering under volume-based diversity functions. Proceedings of the ACM on Management of Data, 3(5), pp.1-18..

Project 9: [**Cardinality Estimation**] Very recently multiple methods for estimating the output cardinality of single-table queries with a having clause have been introduced [1]. These methods

provide cardinality estimates for predicates using the aggregate functions for queries with and without a where-clause. This work also highlights how to handle conjunctions and disjunctions in the having-clause.

Brainstorm and suggest similar ideas for other important SQL clauses for which no important progress has been made so far. Moreover, you may extend this to multi-table queries.

[1] Guido Moerkotte. Cardinality Estimation for Having-Clauses. Proceedings of the VLDB Endowment, 18(1): 28 - 41, 2024.

Project 10: **[Query Rewrites with LLMs]** Recently Large Language Models (LLMs) are being used for the purpose of effective query rewriting [1]. These approaches significantly improve the query execution efficiency and outperforms the baseline methods. In addition, these methods exhibit high robustness across different datasets.

Suggest some improvements to the said methods. You may also deploy LLMs for other purposes of database management tasks and cost estimation.

[1] Zhaodonghui Li, Haitao Yuan, Huiming Wang, Gao Cong, and Lidong Bing. LLM-R^2: A Large Language Model Enhanced Rule-based Rewrite System for Boosting Query E!ciency. PVLDB, 18(1): 53 - 65, 2024.

Project 11: **[In-Context Learning of Text-to-SQL]** There is currently a significant gap between the performance of fine-tuned models and prompting approaches using Large Language Models (LLMs) on the challenging task of text-to-SQL, as evaluated on datasets such as Spider. To improve the performance of LLMs in the reasoning process, a recent study has shown how decomposing the task into smaller sub-tasks can be effective [1]. In particular, it shows that breaking down the generation problem into sub-problems and feeding the solutions of those sub-problems into LLMs can be an effective approach for significantly improving their performance. The experiments with three LLMs show that this approach consistently improves their simple few-shot performance and the in-context learning beats many heavily fine-tuned models.

Some limitations of this approach are already highlighted in the said paper. Our manually constructed demonstrations are fixed for each query class. Future research may explore adaptive and automated methods for generating demonstrations at finer granularities, which can further enhance the performance of our approach. Additionally, the said approach, characterized by its decomposed and step-by-step structure, incurs a heavy cost. As LLMs continue to advance, these costs and latencies should decrease, but reducing the cost is another possible direction. Brainstorm on these directions and suggest further improvements.

[1] M. Pourreza and D. Rafiei. DIN-SQL: Decomposed in-context learning of text-to-sql with self-correction. Advances in Neural Information Processing Systems, 36, 2024.

Project 12: **[Coresets for Query Optimization]** A recent paper has studied query optimization, where the optimizer is made aware of the uncertainty in cardinality estimates and needs to select the most robust plan [1]. This work proves positive and negative results on the sizes of such coresets for common cost-function classes and presents algorithms that construct workload-aware coresets whose size and performance closely match those of the optimal workload-specific coresets.

Suggest improvements to this work and extend accordingly.

[1] Rahul Raychaudhury, Haibo Xu, Pankaj K. Agarwal, Stavros Sintos, Jun Yang. Coresets for Robust Query Optimization. ACM Symposium on Principles of Database Systems (PODS), 2026.

Project 13: [**Cardinality Estimation in Dynamic Streams**] Some recent works have reduced the respective cardinality estimation problem to continual counting on the difference stream associated with the true function values on the input stream [1]. In such reductions, a change in the original stream can cause many changes in the difference stream, this poses a challenge for applying private continual counting algorithms to obtain optimal error bounds. A recent work has improved the accuracy of several such reductions by studying the associated ℓ_p-sensitivity vectors of the resulting difference streams and isolating their properties [2].

Suggest improvements to this work or find other related applications and extend accordingly.

[1] Jain, P., Kalemaj, I., Raskhodnikova, S., Sivakumar, S. and Smith, A., 2023. Counting distinct elements in the turnstile model with differential privacy under continual observation. Advances in Neural Information Processing Systems, 36, pp.4610-4623.

[2] Andersson, Joel Daniel, Palak Jain, and Satchit Sivakumar. Improved Accuracy for Private Continual Cardinality Estimation in Fully Dynamic Streams via Matrix Factorization. ACM Symposium on Principles of Database Systems (PODS), 2026.

Project 14: [**Fragmentation of B-Trees**] Internal fragmentation in data structures is a challengeing task in database design. A seminal result of Yao in this field shows that evenly splitting the leaves of a B-tree against a workload of uniformly random insertions achieves space utilization of around 69% [1]. However, many database applications perform batched insertions, where a small run of consecutive keys is inserted at a single position. A recent work has developed a generalization of Yao's analysis to provide rigorous treatment of such batched workloads. This approach revisits and reformulates the analytical structure underlying Yao's result in a way that enables generalization and is used to argue that even splitting works well for many workloads in our extended class [2].

Suggest improvements to this work and extend accordingly.

[1] Andrew Chi-Chih Yao. On random 2-3 trees. Acta Informatica, 9:159–170, 1978.

[2] Bender, M.A., Bernstein, A., Cao, N., Conway, A., Farach-Colton, M., Komlós, H., Shechter, Y. and Wein, N. Bounding the Fragmentation of B-trees Subject to Batched Insertions. ACM Symposium on Principles of Database Systems (PODS), 2026.

Project 15: [**Transforming Data Orthogonal to Bias**] Despite substantial research on counterfactual fairness, existing methods remain underdeveloped in addressing the impact of multivariate and continuous sensitive variables on decision-making outcomes. To tackle this gap, a recent work has proposed a novel data pre-processing algorithm, Orthogonal to Bias (OB), which is designed to eliminate the influence of a group of continuous sensitive variables, thereby promoting counterfactual fairness in machine learning applications [1]. This approach is based on the assumption of an elliptical distribution within a structural causal model (SCM). The OB algorithm is model-agnostic, making it applicable to a wide range of machine learning

models and tasks. To enhance numerical stability, they have also introduced a sparse variant that incorporates regularization.

Suggest improvements to this work and extend accordingly.

[1] Chen, S. and Zhu, S., 2025, August. Counterfactual Fairness Through Transforming Data Orthogonal to Bias. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 239-249).

Project 16: **[Unlearning Data Bias]** Continual Learning (CL) primarily aims to retain knowledge to prevent catastrophic forgetting and transfer knowledge to facilitate learning new tasks. Unfortunately, data biases simultaneously reduce CL's ability to retain and transfer knowledge. A recent work has proposed a method that removes erroneous memories caused by biases in CL, enhancing performance in both new and old tasks. This method consists of two modules: Error Identification and Error Erasure. The former learns the probability density distribution of task data in the feature space without prior knowledge, enabling accurate identification of potentially biased samples. The latter ensures only erroneous knowledge is erased by shifting the decision space of representative outlier samples. Additionally, an incremental feature distribution learning strategy is designed to reduce the resource overhead during error identification in downstream tasks.

Suggest improvements to this work and extend accordingly.

[1] Cao, X., Gu, H., Yang, X., Wei, B., Liang, H., Wang, X. and Li, T., 2025, August. Errorеraser: Unlearning data bias for improved continual learning. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 119-130).

Project 17: **[Subset Sampling over Joins]** Unlike the traditional subset sampling task, modern machine learning applications over relational data often require sampling from a set defined implicitly by a relational join. A recent paper has studied the problem of subset sampling over joins where each join result is included independently with some probability [1]. They address the general setting where the probability is derived from input tuple weights via decomposable functions (e.g., product, sum, min, max). They proposed the first efficient algorithms that achieve near-optimal time and space complexity for subset sampling over acyclic joins. Their contributions include (i) a static index for generating multiple (independent) subset samples over joins, (ii) a one-shot algorithm for generating a single subset sample over joins, and (iii) a dynamic index that can support tuple insertions, while maintaining a one-shot sample or generating multiple (independent) samples.

Suggest improvements to this work and extend accordingly. Otherwise, use this method for some interesting application.

[1] Esmailpour, Aryan, Xiao Hu, Jinchao Huang, and Stavros Sintos. Subset Sampling over Joins. ACM Symposium on Principles of Database Systems (PODS), 2026.

Project 18: **[Sketching Trimmed Statistics]** A recent work has presented the first ever linear sketching approaching for estimating trimmed statistics for a n-dimensional frequency vector $\mathbf{x}$, e.g., F_p moment for $p \geq 0$ of the largest k frequencies (i.e. coordinates) of $\mathbf{x}$ by absolute value and of the k-trimmed vector, which excludes both the top and bottom k frequencies [1]. Linear

sketches are powerful tools which increase runtime and space efficiency and are applicable to a wide variety of models including streaming and the distributed setting.

Suggest improvements to this work and extend accordingly. Otherwise, use this method for some interesting application.

[1] Lin, Honghao, Hoai-An Nguyen, and David P. Woodruff. "On Sketching Trimmed Statistics. ACM Symposium on Principles of Database Systems (PODS), 2026.

Project 19: [**Recursive querying of neural networks**] Expressive querying of machine learning models enables their verification and interpretation using declarative languages, thus making learned representations of data more accessible. Motivated by the querying of feedforward neural networks, a recent work has investigated logics for weighted structures by revisiting the functional fixpoint mechanism proposed by Grädel and Gurevich [1].

Suggest improvements to this work and extend accordingly.

[1] Grohe, M., Standke, C., Steegmans, J. and Bussche, J.V.D., 2026. Recursive querying of neural networks via weighted structures. ACM Symposium on Principles of Database Systems (PODS), 2026.

Project 20: [**Approximation of Inference Queries**] A recent work has presented an index-based, domain-agnostic framework for approximating inference on unstructured data [1]. It consists of two main components: a coarse-to-fine framework that can be efficiently constructed using minimal memory, and a novel index adaptation component that makes use of oracle invocations during query execution in order to adaptively produce representative sets that yield high-quality approximate results.

Suggest improvements to this work and extend accordingly.

[1] Papadopoulos, C. C., Simitsis, A. and Pedersen, T. B. HAIDES: Adaptive Approximation of Inference Queries over Unstructured Data. In 2025 IEEE 41st International Conference on Data Engineering (ICDE), pp. 2394-2407. IEEE, 2025.

Project 21: [**Experiments Under Diminishing Marginal Effects**] A recent work has addressed a pair of tasks for online sequential experiments under diminishing marginal effect: (1) Adaptively identify the optimal traffic allocation to maximize the expected cumulative reward, and (2) Construct valid central limit theorem to perform reliable statistical inference for the expected reward under the optimal traffic allocation that is unknown beforehand [1]. They showed that classical algorithms can fail to deliver the second task, especially because the statistical inference task presents its own difficulty. We develop a new online algorithm that leverages an additional smoothness condition on how the marginal effects change to achieve both tasks. We prove that this algorithm obtains the best achievable expected cumulative reward. Further, crucially for online platforms' need to do trustworthy statistical inference, the algorithm is proved to enjoy a valid central limit theorem.

Suggest improvements to this work and extend accordingly.

[1] Xu, Jingxu, Yuhang Wu, Yingfei Wang, Chu Wang, and Zeyu Zheng. A/B test and online experiment under diminishing marginal effects: Regret minimization and statistical

inference. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 3401-3412. 2025.

Project 22: **[Quantum Data Management]** A recent paper has addressed the issue of data management for quantum computing in the noisy intermediate-scale quantum (NISQ) era [1]. After differentiating quantum data from classical data, they have outlined current and future data management paradigms in the NISQ era and beyond. They have also pointed out the data management challenges arising from the emerging demands of near-term quantum computing. Our goal is to chart a clear course for future quantum-oriented data management research, establishing it as a cornerstone for the advancement of quantum computing in the NISQ era.

Suggest improvements to this work and extend accordingly.

[1] Rihan Hai, Shih-Han Hung, Tim Coopmans, Tim Littau, and Floris Geerts. Quantum Data Management in the NISQ Era. PVLDB, 18(6): 1720 - 1729, 2025.

Project 23: **[In-Context-Learning-Based Text-to-SQL Errors]** A recent paper has conducted the first comprehensive study of text-to-SQL errors [1]. This study covers four representative ICL-based techniques, five basic repairing methods, two benchmarks, and two LLM settings. This work finds that text-to-SQL errors are widespread and summarizes 29 error types of 7 categories. It is also observed that existing repairing attempts have limited correctness improvement at the cost of high computational overhead with many mis-repairs. Based on the findings, they propose MapleRepair, a novel text-to-SQL error detection and repairing framework.

Suggest improvements to this work and extend accordingly.

[1] Shen, J., Wan, C., Qiao, R., Zou, J., Xu, H., Shao, Y., Zhang, Y., Miao, W. and Pu, G. A study of in-context-learning-based text-to-sql errors. arXiv preprint arXiv:2501.09310, 2025.

Project 24: **[Open Choice]** You can choose any single database-related research paper from these lists [1, 2, 3, 4] and suggest novel improvements on the method proposed there.

[1] `https://www.vldb.org/pvldb/volumes/19`

[2] `https://2026.sigmod.org/pods_papers.shtml`

[3] `https://kdd2025.kdd.org/research-track-papers-2`

[4] `https://ieee-icde.org/2025/research-papers`