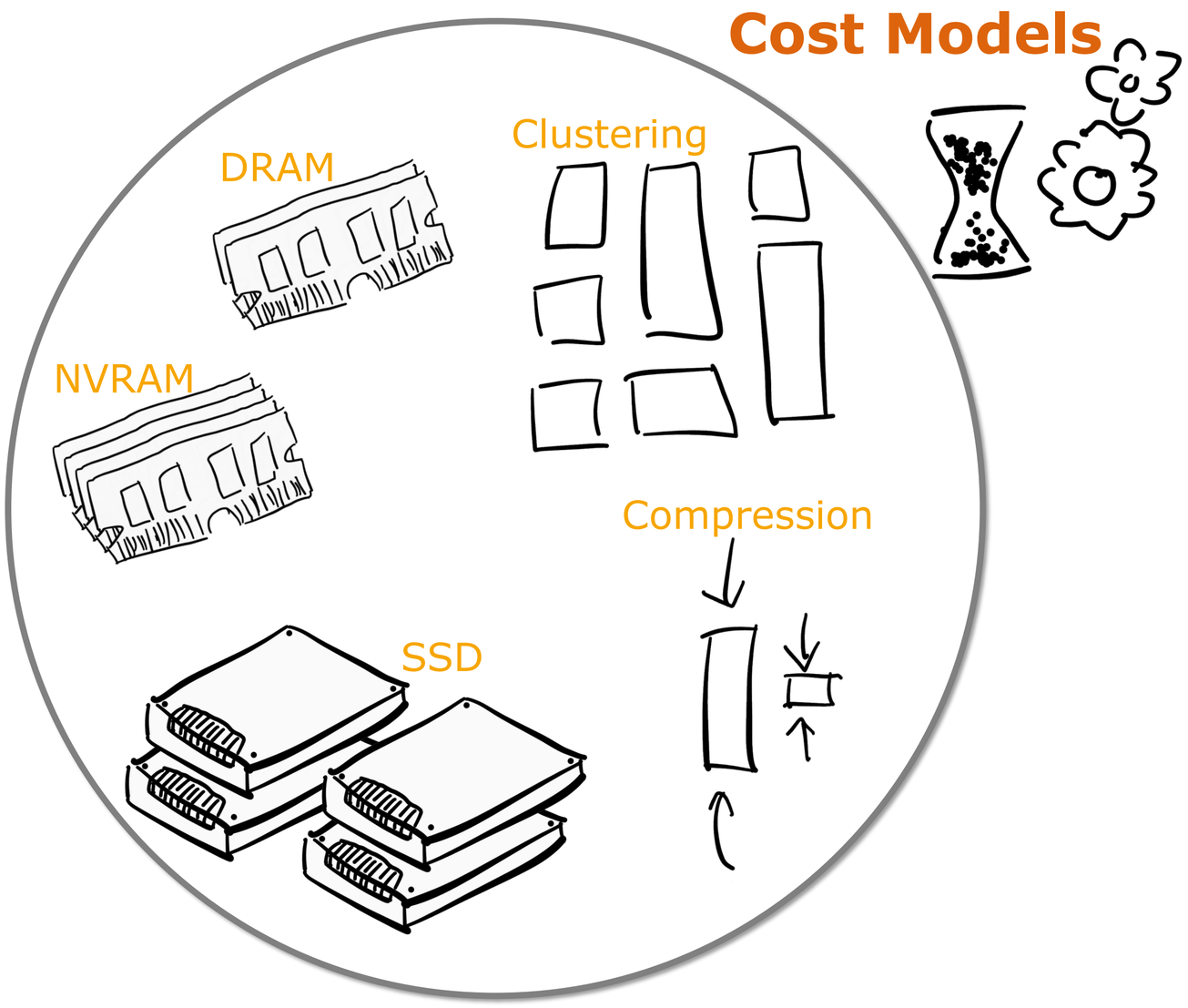
Efficient Cost Models for Joint Physical Database Design

Abstract

Autonomous databases are a relevant field of database research because the need for efficient data analysis increases with the progress of data-intense business processes. Learned cost models are crucial for such systems, as they are the basis on which autonomous database components make decisions.

While most current cost models mainly focus on exact latency estimations, they have the downsides of large amounts of training data, long training duration, and little generalizability to unseen datasets. Autonomous approaches for physical database design usually account for a single aspect, such as tiering or compression. We suggest research on cost models for multiple physical database design aspects that can be reused for new datasets and workloads.



Problem

More and more DB instances operate in cloud environments. Contemporary, analyzing large amounts of data is important for many use cases.

Autonomous databases aim to make design and execution decisions superior to humans [6, 8]. Besides knob-tuning and entirely learned components, autonomous tuning of physical database design helps to optimize performance and operational expenses. This includes, e.g., data tiering, compression, partitioning, or clustering [1, 2, 5].

In the light of high-performance in-memory databases, also considering the performance implications of physical database design is crucial. However, different physical design choices are not isolated problems. Optimizing for one aspect can have a negative impact on another aspect.

So far, autonomous decisions about physical configuration rarely account for multiple design aspects together [8]. The training and learning times of used cost models are long, and the predictions are often bound to single DB instances and workloads [4, 8].

Goal

With the proposed research, we envision efficient training of cost models that enable joint physical database tuning. Different design aspects are considered altogether.

Thus, we can account for large datasets by tiering them to another storage layer, compressing them, and clustering them to ease parallelization and data pruning.

Solution

We propose to develop learned cost models for joint physical database design. These models should be able to provide feasible cost estimates derived from sparse training data. The latter is important as we have to cover the different combinations of design decisions.

Thus, the models must be able to interpolate well to unseen data points. Rather than using identifiers of current datasets, the models have to use data and workload characteristics as features (cf. [3]). Doing so, they can be adapted to new database instances with unseen datasets and workloads. Additionally, we want to balance training overhead and estimation accuracy to a point where we are able to derive good configurations by few training data.

Though there are approaches of joint physical database tuning [7], they do not focus on reusability and efficient learning. Reusable cost models do not yet incorporate physical design and aim to provide accurate estimations [3]. In previous work, we have already shown that imperfect cost estimations enable feasible results for single aspects of physical design, e.g., clustering [5]. Hence, the combination of these approaches is promising.

Daniel Lindner

M. Sc. Student, Data Engineering
Lecture Series on Database Research 2021/22
Hasso Plattner Institute, Potsdam, Germany

E-Mail: daniel.lindner@student.hpi.de

References

[1] Martin Boissier and Max Jendruk. "Workload-Driven and Robust Selection of Compression Schemes for Column Stores". In: *EDBT*. 2019.
[2] Stefan Halfpap and Rainer Schlosser. "Memory-Efficient Database Fragment Allocation for Robust Load Balancing when Nodes Fail". In: *ICDE*. 2021.
[3] Benjamin Hilprecht and Carsten Binnig. "One Model to Rule them All: Towards Zero-Shot Learning for Databases". In: *CIDR*. 2022.
[4] Benjamin Hilprecht, et al. "DBMS Fitting: Why should we learn what we already know?" In: *CIDR*. 2020.
[5] Daniel Lindner, et al. "Learned What-If Cost Models for Autonomous Clustering". In: *ADBIS*. 2021.
[6] Lin Ma, et al. "MB2: Decomposed Behavior Modeling for Self-Driving Database Management Systems". In: *SIGMOD*. 2021.
[7] Keven Richly, et al. "Joint Index, Sorting, and Compression Optimization for Memory-Efficient Spatio-Temporal Data Management". In: *ICDE*. 2021.
[8] Xuanhe Zhou, et al. "Database Meets Artificial Intelligence: A Survey". In: *TKDE* 34.3 (2022).

HPI Hasso Plattner Institut