Learned DBMS Components 2.0: From Workload-Driven to Zero-Shot Learning

by Carsten Binnig (TU Darmstadt)

Database management systems (DBMSs) are the backbone for managing large volumes of data efficiently and thus play a central role in business and science today. For providing high performance, many of the most complex DBMS components such as query optimizers or schedulers involve solving non-trivial problems. To tackle such problems, very recent work has outlined a new direction of so-called learned DBMS components where core parts of DBMSs are being replaced by machine learning (ML) models which has shown to provide significant performance benefits. However, a major drawback of the current workload-driven learning approaches to enable learned DBMS components is that they not only cause very high overhead for training an ML model to replace a DBMS component but that the overhead occurs repeatedly which renders these approaches far from practical. Hence, in this talk we present our vision to tackle the high costs and inflexibility of workload-driven learning called Learned DBMS Components 2.0. First, we introduce data-driven learning where the idea is to learn the data distribution over a complex relational schema. In contrast to workload-driven learning, no large workload has to be executed on the database to gather training data. While data-driven learning has many applications such as cardinality estimation or approximate query processing, many DBMS tasks such as physical cost estimation cannot be supported. We thus propose a second technique called zero-shot learning which is a general paradigm for learned DBMS components. Here, the idea is to train models that generalize to unseen data sets out-of-the-box. The idea is to train a model that has observed a variety of workloads on different data sets and can thus generalize. Initial results on the task of physical cost estimation suggest the feasibility of this approach. Finally, we discuss further opportunities which are enabled by zero-shot learning.