Universität Basel

Research Project

Polypheny-DB: Cost- and Workload-aware Adaptive Data Management

Third-party funded project

Project title Polypheny-DB: Cost- and Workload-aware Adaptive Data Management
Principal Investigator(s) Schuldt, Heiko ;
Project Members Stiemer, Alexander ; Vogt, Marco ;
Organisation / Research unit
Departement Mathematik und Informatik / Databases and Information Systems (Schuldt)
Department
Project Website https://dbis.dmi.unibas.ch/research/projects/polypheny-db/
Project start 01.05.2017
Probable end 30.04.2021
Status Completed

In the last few years, it has become obvious that the “one-size-fits-all” paradigm, according to which database systems have been designed for several decades, has come to an end. The reason is that in the very broad spectrum of applications, ranging from business over science to the private life of individuals, demands are more and more heterogeneous. As a consequence, the data to be considered significantly differs in many ways, for example from immutable data to data that is frequently updated; from highly structured data to unstructured data; from applications that need precise data to applications that are fine with approximate and/or outdated results; from applications which demand always consistent data to applications that are fine with lower levels of consistency. Even worse, many applications feature heterogeneous data and/or workloads, i.e., they intrinsically come with data (sub-collections) with different requirements for data storage, access, consistency, etc. that have to be dealt with in the same system.

This development can either be coped with a large zoo of specialized systems (for each subset of data with different properties), or by a new type of flexible database system that automatically adapts to the –potentially dynamically changing– needs and characteristics of applications. Such behaviour is especially important in the Cloud where several applications need to be hosted on a shared infrastructure, in order to make economic use of the available resources. In addition, the infrastructure costs that incur in a Cloud are made transparent in a very fine-grained way. The Polypheny-DB project will address these challenges by dynamically optimizing the data management layer, by taking into account the resources needed and an estimation of the expected workload of applications. The core will be a comprehensive cost model that seamlessly addresses these criteria and that will be used, in two subprojects, to optimize i.) data storage and access, and ii.) data distribution. Each of the two subprojects will be addressed by one PhD student.

Data Storage and Access: different physical storage media such as spinning discs, flash storage, or main memory come with different properties and performance guarantees for data access, but also significantly differ in terms of the costs for the necessary hardware and the volume of data that can be managed at a certain price. In addition, the access characteristics of applications (e.g., read-only, mostly reads, balanced read/write ratio, high update frequencies) may favour one solution over the other oră even require hybrid solutions. The same also applies to the data model to be used and the choice between structured data, semi-structured data, or schema-less data management. Similarly, applications might require data to be approximated to speed up queries or to be compressed to save storage space.

In Polypheny-DB, we will jointly addressă these choicesă andă alternativesă byă devisingă aă comprehensive cost model for data storage that allows to decide, if necessary at the level of sub-collections, how they can be best stored given an available budget and the requirements for data access. The costs for data storage and access will be continuously monitored, together with an analysis of the expected workload of applications and their storage strategy will be updated dynamically, if necessary.

Data Distribution: a high degree of data availability necessitates that data is replicated across different sites in a distributed system. For read-only applications, this also allows to balance the (read) load across sites. However, in the presence of updates, additional costs for distributed transactions incur to make replicas consistent – unless weaker consistency models can be tolerated. In our previous work, we have designed, implemented, and evaluated a protocol that manages data in a distributed environment either by dynamically replicating data and managing consistency based on the costs that incur and the available budget or by partitioning data, in particular by co-locating data items that are frequently accessed jointly to minimize the number of distributed transactions. While both approaches are very effective, none of them alone can address dynamic workloads. In Polypheny-DB, we will develop a comprehensive cost model that seamlessly combines replication and partitioning of (subsets of) data. Again, the cost model will be used to dynamically re-assess replication and/or partitioning decisions and to adapt data distribution from a cost-effectiveness point of view.

Keywords Cloud data management, Data Partitioning, Data replication, Data storage, Dynamic adaptation
Financed by
Swiss National Science Foundation (SNSF)

Add publication

Add documents

Specify cooperation partners