Universität Basel

Research Project

Polypheny-DDI

Third-party funded project

Project title Polypheny-DDI
Principal Investigator(s) Schuldt, Heiko ;
Project Members Lengweiler, David ; Vogt, Marco ;
Organisation / Research unit
Departement Mathematik und Informatik / Databases and Information Systems (Schuldt)
Department
Departement Mathematik und Informatik
Project start 01.01.2023
Probable end 31.12.2026
Status Active

In recent years, data-driven research has established itself as the fourth pillar in the spectrum of scientific methods, alongside theory, empiric research, and computer-based simulation. In various scientific disciplines, increasingly large amounts of –both structured and unstructured– data are being generated or existing data collections that have originally been isolated from each other are being linked in order to gain new insights. he process of generating knowledge from raw data is called Data Science or Data Analytics.

The entire data analytics pipeline is quite complex, and most work focuses on the actual data analysis (using machine learning or statistical methods), while largely neglecting the other elements of the pipeline. This is particularly the case for all aspects related to data management, storage, processing, and retrieval – even though these challenges actually play an essential role. A Distributed Data Infrastructure (DDI) supports a large variety of data management features as demanded by the data analytics pipeline. However, DDIs are usually very heterogeneous in terms of data models, access characteristics, and performance expectations. In addition, DDIs for integrating, continuously updating, and querying data from various heterogeneous applications need to overcome the inherent structural heterogeneity and fragmentation.

Recently, polystore databases have gained attention because they help overcome these limitations by allowing data to be stored in one system, yet in different formats and data models and by offering one joint query language. In past work, we have developed Polypheny-DB, a distributed polystore that integrates several different data models and heterogeneous data stores. Polypheny-DB goes beyond most existing polystores and even supports data accesses with mixed workloads (e.g., OLTP and OLAP). However, polystores are limited to rather simple object models, static data and exact queries. When individual data items follow a complex inherent structure and consist of several heterogeneous parts between which dedicated constraints exist, when the access goes beyond exact Boolean queries, when data is not static but continuously produced, and/or when objects need to be preserved in multiple versions, then polystores quickly reach their limits.

At the same time, these are typical requirements for data management within a data analytics pipeline. Examples are scientific instruments that continuously produce new data as data streams; social network analysis that requires support for complex object models including structured and unstructured content; data produced by imaging devices that requires sophisticated similarity search support, or frequently changing objects that are subject to time-dependent analyses.

The objective of the Polypheny-DDI project is to seamlessly combine the functionality of a polystore database with that of a distributed data infrastructure to meet the requirements of data science applications. It will focus on i.) supporting complex composite object models and enforcing constraints between the constituent parts; ii.) supporting similarity search in multimedia content, and iii.) supporting continuous data streams and temporal/multiversion data.

Keywords Databases, Multi-model Data Management, Polystore
Financed by
Swiss National Science Foundation (SNSF)

Follow-up project of 3775394 Polypheny-DB: Cost- and Workload-aware Adaptive Data Management

Add publication

Add documents

Specify cooperation partners