Mosaics in Big Data

Stratosphere, Apache Flink, and Beyond

Volker Markl

ABSTRACT

The global database research community has greatly impacted the functionality and performance of data storage and processing systems along the dimensions that define "big data", i.e., volume, velocity, variety, and veracity. Locally, over the past five years, we have also been working on varying fronts. Among our contributions are: (1) establishing a vision for a database-inspired big data analytics system, which unifies the best of database and distributed systems technologies, and augments it with concepts drawn from compilers (e.g., iterations) and data stream processing, as well as (2) forming a community of researchers and institutions to create the Stratosphere platform to realize our vision. One major result from these activities was Apache Flink, an open-source big data analytics platform and its thriving global community of developers and production users. Although much progress has been made, when looking at the overall big data stack, a major challenge for database research community still remains. That is, how to maintain the ease-of-use despite the increasing heterogeneity and complexity of data analytics, involving specialized engines for various aspects of an end-to-end data analytics pipeline, including, among others, graph-based, linear algebra-based, and relational-based algorithms, and the underlying, increasingly heterogeneous hardware and computing infrastructure. At TU Berlin, DFKI, and the Berlin Big Data Center (BBDC), we aim to advance research in this field via the Mosaics project. Our goal is to remedy some of the heterogeneity challenges that hamper developer productivity and limit the use of data science technologies to just the privileged few, who are coveted experts.

CCS CONCEPTS

• **Information Systems → Data Management Systems → Database Management System Engines → Database query processing** → Query optimization; Query operators • **Information Systems → Data Management Systems → Database Management System Engines** → Stream management; Online analytical processing engines; DBMS engine architectures

KEYWORDS

Apache Flink, big data, data science, declarative languages, federation, heterogeneous data management