

Big Data Analytics 2013



Invited talk

Big Data Analytics with Stratosphere

Volker Markl, *TU Berlin*

The talk will present a programming model for big data analytics, with a particular focus on our research in a massively parallel data processor in the Stratosphere project. We will present a new flavor of data processor that goes beyond the popular map/reduce paradigm. We propose a programming model based on second order functions that describe what we call parallelization contracts (PACTs). PACTs are a generalization of the map/reduce programming model, extending it with additional higher order functions and output contracts that give guarantees about the behavior of a function. A PACT program is transformed into a data flow for a massively parallel execution engine, which executes its sequential building blocks in parallel and provides communication, synchronization and fault tolerance. The concept of PACTs allows the system to abstract parallelization from the specification of the data flow and thus enables several types of optimizations on the data flow. The system as a whole is as generic as map/reduce systems, but can provide higher performance through optimization and adaptation of the system to changes in the execution environment. Moreover, it enables the execution of tasks that traditional map/reduce systems cannot execute without mixing data flow program specification and parallelization, like joins, time-series analysis or data mining operations. We will present our research vision and research results that we have achieved during the last year. We will also highlight our research agenda for the upcoming year.