Stratosphere
www.stratosphere.eu

Research Theme
- What are the right system architectures and programming abstractions for “Big Data” analytics? How to benchmark such systems?
- How to compile, optimize, parallelize, and execute data analytics programs in massively parallel and unreliable environments?
- Systems building approach: validate research hypotheses via prototypes and systems, contribute to open source ecosystem.

End-to-end system for data analytics in massively parallel cluster and cloud environments. Combines key technologies of MapReduce and parallel database systems.

Funding organizations

- Alexander S. Alexandrov, Stephan Ewen, Fabian Hueske, Marcus Leich, Sebastian Schelter, Kostas Tzoumas

Ongoing and future research
- How can we support diverse analytical applications from various domains with a single compiler/optimizer and runtime system?
- How can we optimize the system internals for better resource consumption and low latency in modern hardware?
- Robust optimization, optimistic recovery, functional language frontend, domain-specific languages and algebra, ...

Myriad
myriad-toolkit.com


Framework for declarative expression and massively parallel execution of data generation programs.

Exploit skip-ahead random number generation to generate dependent data in independent nodes.

PACT programming model


Specifying complex data analytics jobs using user-defined functions. System interaction using second-order functions that extend MapReduce.

Program optimization


Achieve database logical and physical optimizations in a general data flow programming model via static code analysis of user code.

Stratosphere iterations


Data extraction, transformation, machine learning, graph analysis, post-processing in a single system by “teaching” iterations to data flow engine.

Incremental iterations can exploit sparse computational dependencies without sacrificing dataflow programming abstraction, and match specialized engines in performance.