Analyzing and Linking Big Data with Stratosphere

Kostas Tzoumas

kostas.tzoumas@tu-berlin.de

www.user.tu-berlin.de/kostas.tzoumas/

StratoSphere
Above the Clouds

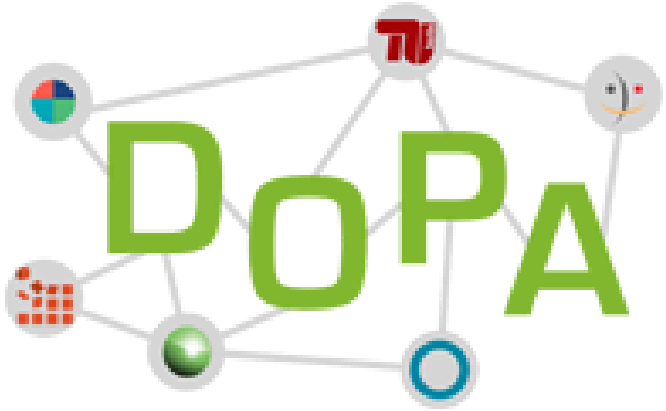
berlin

Big Data

- Driven by cheaper storage and computation
  - Cloud computing further enabling economies of scale
  - Open-source software lowers barrier of entry
- Societal and economical impact
  - Scientific breakthroughs will come from *data exploration*
  - Success in business dictated by the ability to quickly draw insights from data signals
- Big Data Analytics
  - Analytical workloads scale inversely with cost
- Major challenge: Information marketplaces
  - Data as a resource, analytics as a product
  - An “AppStore” for data

The DOPA Vision

- Data Pools as collections of diverse data sets
- Example data pools
  - Evolving history of the Web
  - Financial and statistical data
- Data identification
  - By assigning unique IDs to objects
  - Enables **data linkage**
- Data Analytics
  - Integration and analytics on diverse data sources
  - Using a common framework and language



Some Scenarios

- Sentiment and market analysis
  - An SME producing consumer goods can analyze blogs and social network streams, and link them with customer demographics to perform sentiment analysis and market research
- Green houses
  - A home buyer can find out the energy consumption and distribution over time in houses in a particular area by linking and analyzing energy with demographic data.
- Traffic analysis, transportation and construction planning
  - Linking weather, traffic, road data

Big Data Analytics

- "Big Data" refers to different applications as well as more data
  - Beyond traditional DW queries
- Open-source in data management
  - Enabled primarily by Hadoop popularity
  - Changed research landscape
- New systems are rethinking the complete data management stack at a massively parallel scale
  - As systems mature, need to tackle hard and novel problems

STRATOSPHERE

Stratosphere

- Collaborative Research Project
  - 3 Universities, 5 research groups in the Berlin area
- Infrastructure for Big Data Analytics
- Bridge relational DBMSs and MapReduce worlds
  - Intersection of functional languages and data parallelism
  - Re-architecting data management systems for massive parallelism
- Open-source research platform (Apache)
- Used by a variety of Universities and research institutes

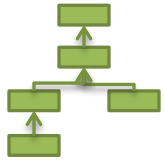
Stratosphere Architecture

StratoSphere
Above the Clouds

Query Processor

```
$res =
  filter $e in $emp
  where
  $e.income > 30000;
```

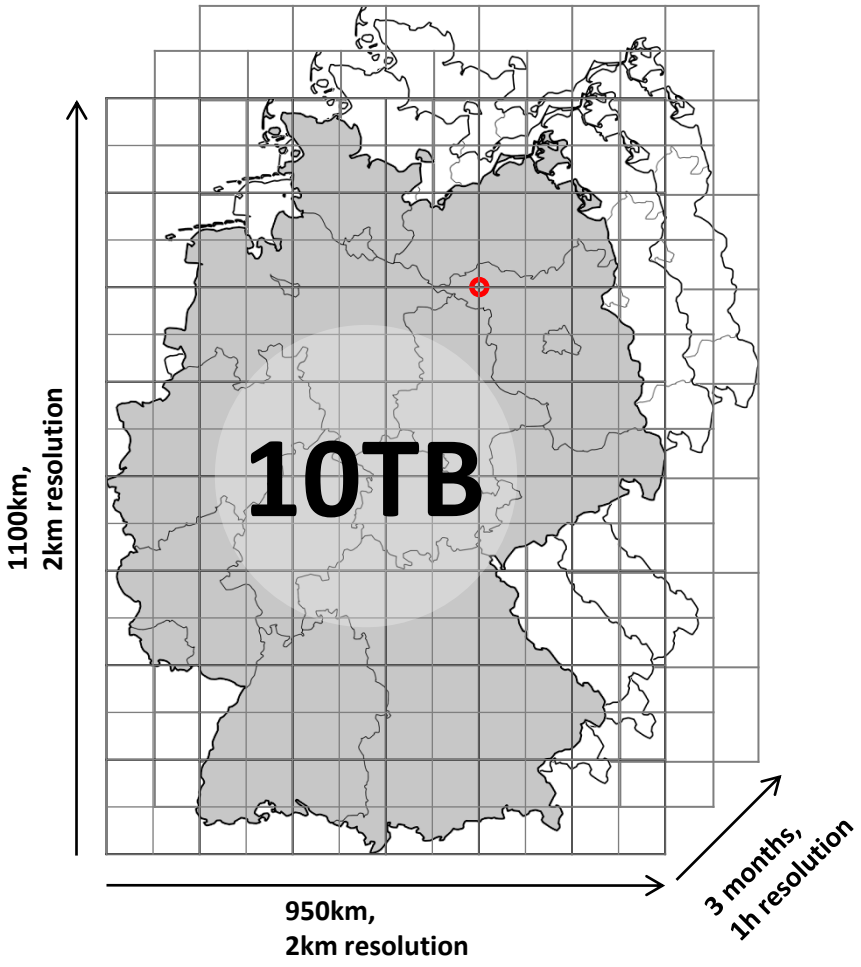
Compiler

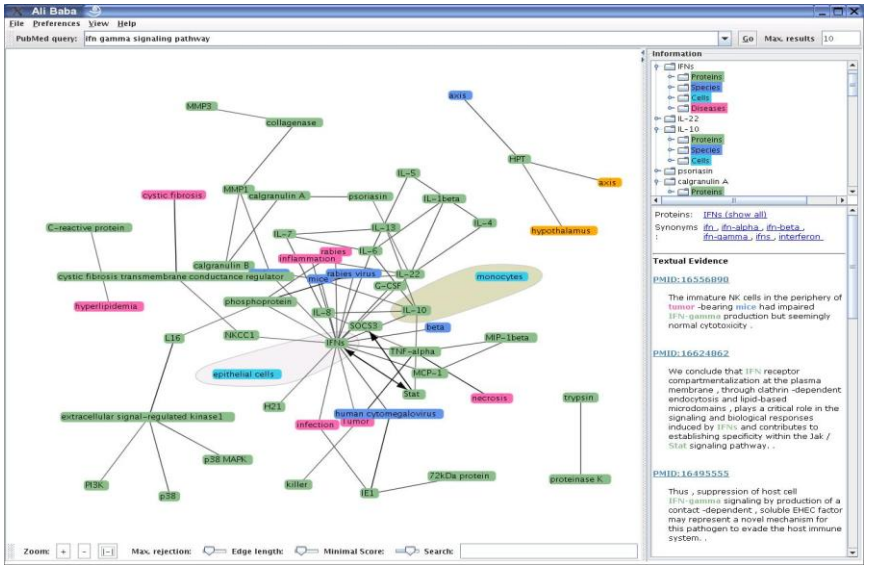


PACT Optimizer

Nephele

Stratosphere Use Cases



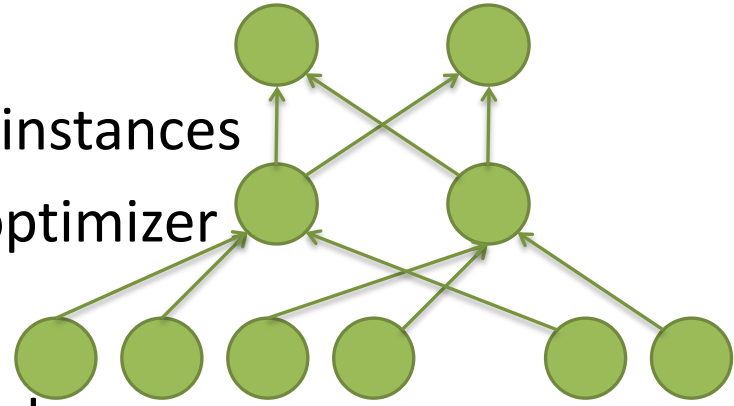


The Meteor Query Language

- Stratosphere declarative front-end inspired by the IBM Jaql language
- Extensible and flexible
  - Easy to add libraries, e.g., for data linkage, cleansing, mining
  - Easy to integrate in language syntax
- Provides operators for Information Extraction and Data Linkage
- Time as a first-class concept

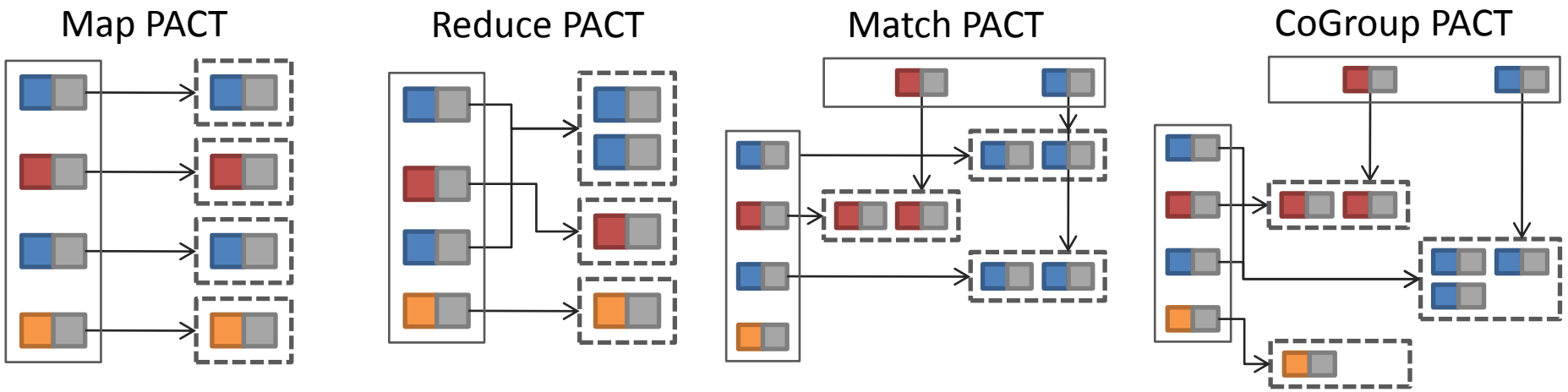
The Nephele Execution Engine

- Executes Nephele schedules
  - DAGs of already parallelized operator instances
  - Parallelization already done by PACT optimizer
- Design decisions
  - Designed to run on top of an IaaS cloud
  - Predictable performance
  - Scalability to 1000+ nodes with flexible fault-tolerance
- Permits network, in-memory (both pipelined), file (materialization) channels



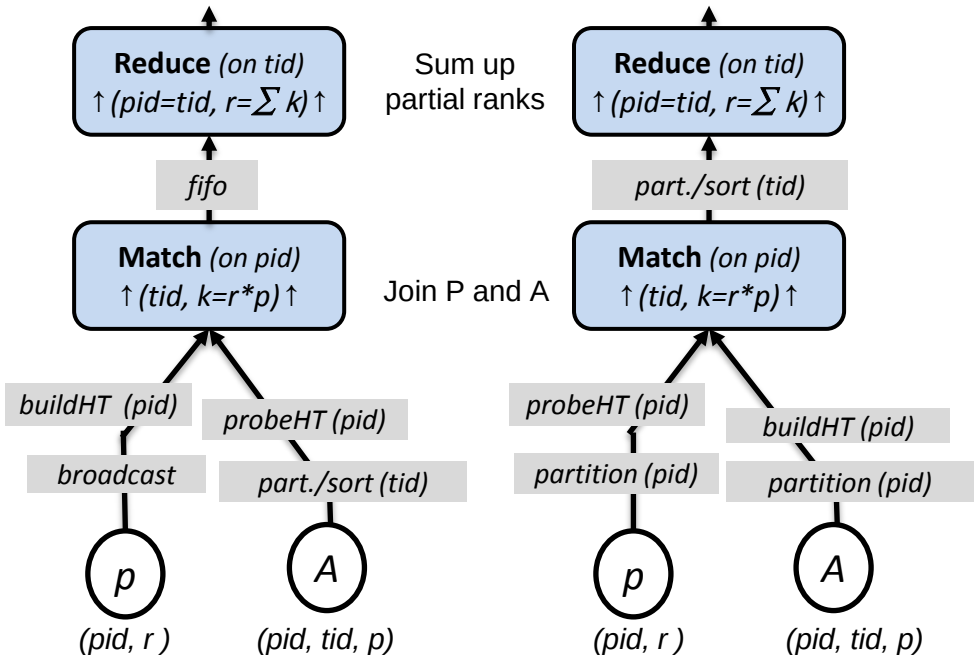
The PACT Programming Model

- Internal Stratosphere programming model
  - Also exposed to the programmer for advanced functionality
- Dataflow, side-effect free programming model enabling massive parallelism
- Centered around the concept of second-order functions
  - Generalization of MapReduce

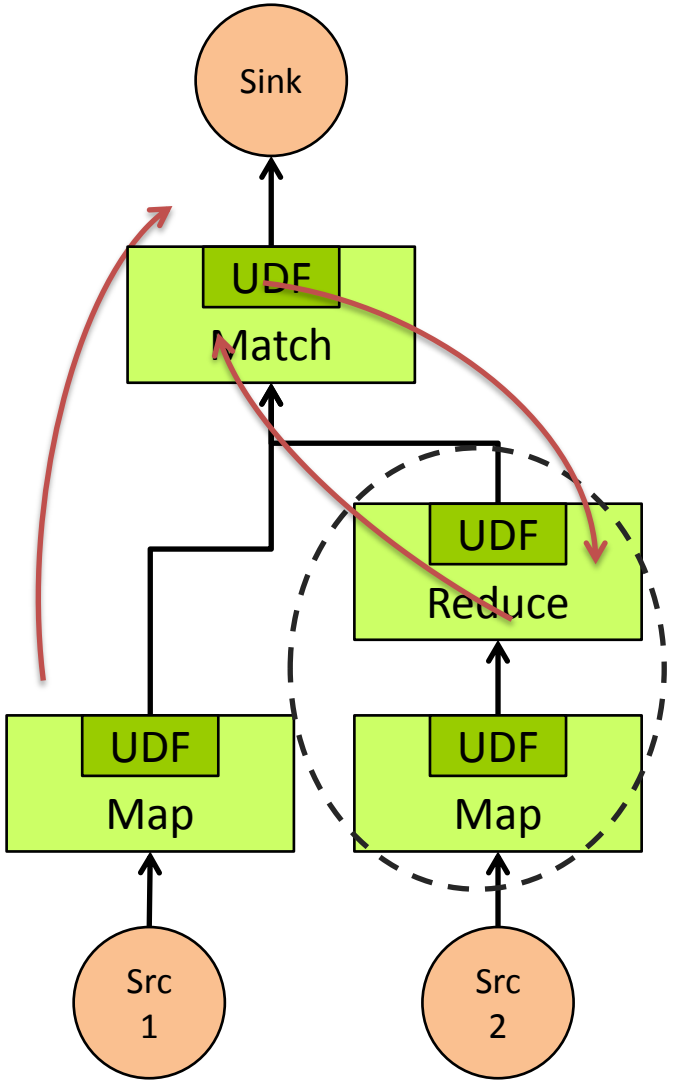


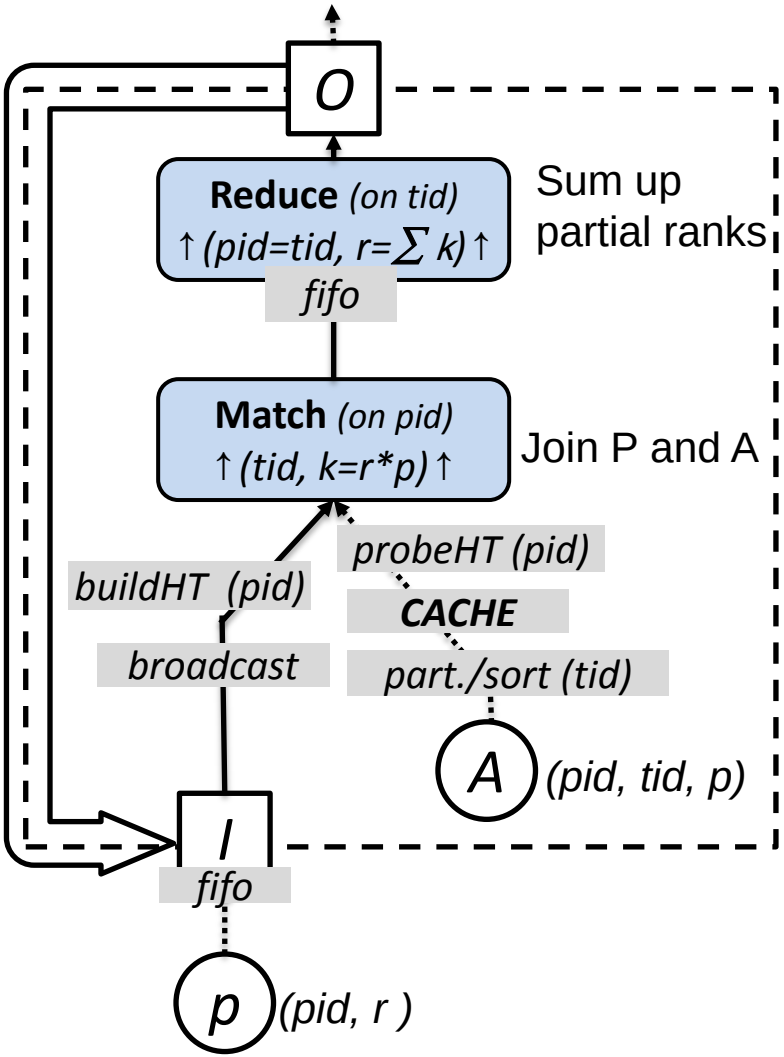
Optimization

- Knowledge of PACT signature permits **automatic optimization** ala Relational DBMSs
- Emulates different hand-crafted MapReduce implementations
- Enables **orders of magnitude** faster programs
- Frees programmer from thinking about execution



Ongoing Research





Summary

- DOPA
  - Bootstrapping the information economy by providing information marketplaces and related business models
  - Brings together heterogeneous data pools
  - Enables easy linkage and analytics across data pools via a flexible programming language
- Stratosphere
  - Technical infrastructure for scalable analytics
  - Pushes the MapReduce paradigm forward
  - Focal point of several research initiatives across Europe and the world

Acknowledgments

- FP7 STREP (DOPA), DFG FOR, EIT (Stratosphere)
- DOPA partners
  - TU Berlin, IMR, DataMarket, OKKAM, Vico, ami
- Stratosphere partners
  - TU Berlin, HU Berlin, HPI
- EIT partners
  - TU Berlin, SICS, TU Delft, Inria, U. Trento, Aalto U., STACZKI

Thank you!

www.stratosphere.eu

@stratosphere_eu

kostas.tzoumas@tu-berlin.de