Big Data Analytics Benchmarking in 2013

Till Westmann
Oracle Labs
What’s the question?

- First generation of MR-based systems well established
  - Hive, Pig, ..
- Next generation of Open-Source BigData systems appears
  - AsterixDB, Spark/Shark, Stratosphere, Impala, …
- How do those new systems compare
  - .. to each other?
  - .. to the first generation?
  - .. to a PRDBMS?
What do we want?

- Who’s faster?
- And under what conditions?
  - Big Memory vs. Small Memory
  - Many Cores vs. Many Boxes

- Ideally in 2013 🎉
What’s available?

- Simple Benchmarks
  - “CALDA”: MR vs. RDBMS (Pavlo et al.)
  - “Big Data Benchmark”: Hive, Redshift, Shark, Impala (UC Berkeley)
- TPC-H
  - Hive vs. SQL server PDW (Floratou et al.)
  - Hive vs. Shark (Xin et al.)
- Application-level benchmark proposals
  - BigBench, …
What’s the plan?

- TPC-H-like queries on TPC-H data
- Different hardware configurations
- On
  - AsterixDB (UC Irvine/UC Riverside)
  - Shark (UC Berkeley)
  - Stratosphere (TU Berlin)
  - Hive
  - Oracle RAC
What are the issues with the plan?

- TPC-H is not really a “Big Data” workload
- Fairness wrt. heterogeneous systems
  - Different architectures/requirements
    - shared disk vs distributed disks
  - Different query languages (SQL, HiveQL, AQL, Meteor)
    - Significant implementation effort
  - Different maturity of optimizers
    - How much manual optimization is “fair”?
Thank you!

Questions / Suggestions ?
Hardware and Software

Engineered to Work Together