Proposal for a Big Data Benchmark Repository

Andries Engelbrecht – HP

May 9, 2012
Big Data Benchmark Repository

Challenges for Big Data Benchmarking

• Benchmarks should reflect customer/industry use-cases & requirements
• Big Data covers a wide range of application domains, data types and use-case scenarios
• Rapidly evolving with an abundance of new workloads and use cases
• Considerably more challenging than narrowly constrained traditional OLTP & Data Warehouse usage patterns
Big Data Benchmark Repository

Big Data Benchmark Repository Proposal
• A repository cataloguing Big Data benchmarks & workloads
• Populated by contributors from the Big Data community & made available to the community
• Less stringent oversight than traditional benchmark councils/committees
• Repository to include the following elements:
  – Name, description and use-case(s) discussion for which the benchmark is intended
  – Access to code, configuration information, workload drivers and applicable monitoring tools/connectors
  – Guidelines to setup, configurations, usage, test objectives, etc.
  – Benchmark results and all typical configuration information & performance metrics (at own discretion)

Benefits
• Provide a common location for easy dissemination for creators of new benchmarks
• Common benchmark/workload repository for all interested parties in the Big Data community
• Benchmarks can evolve and be refined much like open source initiatives today
Thank you