Benchmarking BIG DATA in the Cloud

Dan Koren
Director, Performance Engineering

Friday, May 11, 2012

Assumptions

- **For most data users, BIG DATA will reside in some Cloud**

Applications will be cloud hosted as well

How does one benchmark them?

Classical Benchmarking

Clearly bounded SUT

Known resources and access to all

Single clock

One knows what to measure and how

In the Cloud...

What is the SUT, where is it located, how is it bounded?

Floating resources – with limited or even no access

More than one clock – cannot rely on synchrony

What to do?

Concept...

Components and workloads must measure themselves!

Workloads must report resource utilization and time per unit of work

Benchmark metrics should be computed by aggregating and normalizing data collected by workers

Questions?

Thoughts?

Comments?