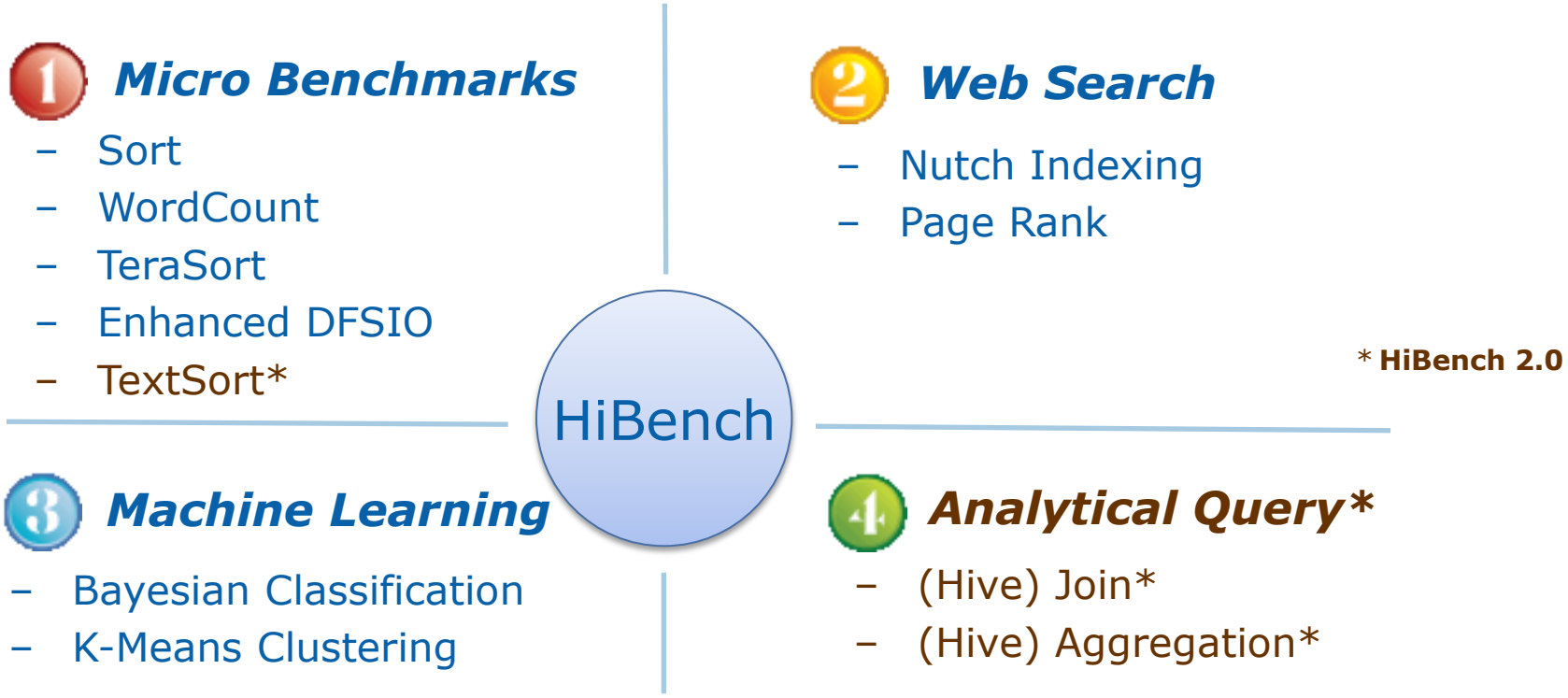
HiBench: A Representative and Comprehensive Hadoop Benchmark Suite

Bhaskar D. Gowda

Enterprise Architect.

DCSG/IASI

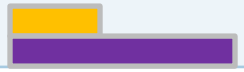
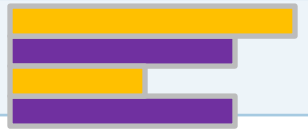
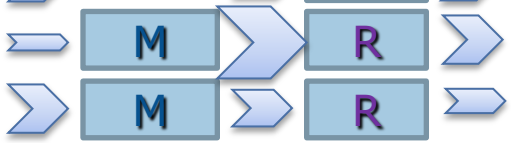
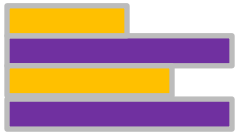
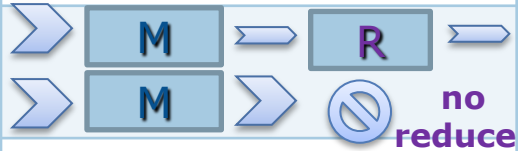
The HiBench benchmark suite



* **HiBench 2.0**

HiBench 1.0 published ("The HiBench Suite: Characterization of the MapReduce-Based Data Analysis") and presented in ICDE'10 workshops

Workload traits drive optimization approach

Workload	System Resource Utilization	Data Access Patterns	Map/Reduce Stage Time Ratio
Sort	I/O bound		
WordCount	CPU bound		
TeraSort	Map stage : CPU-bound; Red stage : I/O-bound		
Nutch Indexing	I/O bound, high CPU utilization in map stage		
Page Rank (1st & 2nd job)	CPU-bound in all jobs		
Bayesian Classification (1st & 2nd job)	I/O bound, with high CPU utilization in map stage in the 1st job		
K-means Clustering	CPU bound in iteration; I/O bound in clustering		
Enhanced DFSIO	I/O-bound	trivial	trivial

data less data even less data compressed

See our whitepaper "Optimizing Hadoop Deployments" (http://communities.intel.com/docs/DOC-5645)

Tradeoffs and Discussions

	GridMix3	SWIM	HiBench
Synthetic workloads	Yes	Yes	No
Ease of scaling	No	Yes	Partially
Mix of jobs	Yes	Yes	Hard
Platform implications	I/O subsystem	I/O subsystem	Full


intel
Amazing things happen with Intel inside®