Caution on "Big Data Benchmark"

- **What changed?**
  - **System scale, complexity, diversity, rapid evolution**
- Opportunity to guide or misguide system design
- Big data encompasses broad scope & diverse solutions/use cases
- For MapReduce, science is not there for "big data benchmark"
- Should be to help solve real life problems, not bragging rights
- Should have multiple workloads with distinct performance metrics
- Should focus on common and underserved use cases
- Should encourage experimentation and acknowledge limits
- Should regularly re-assess the quality of any benchmarks

cloudera

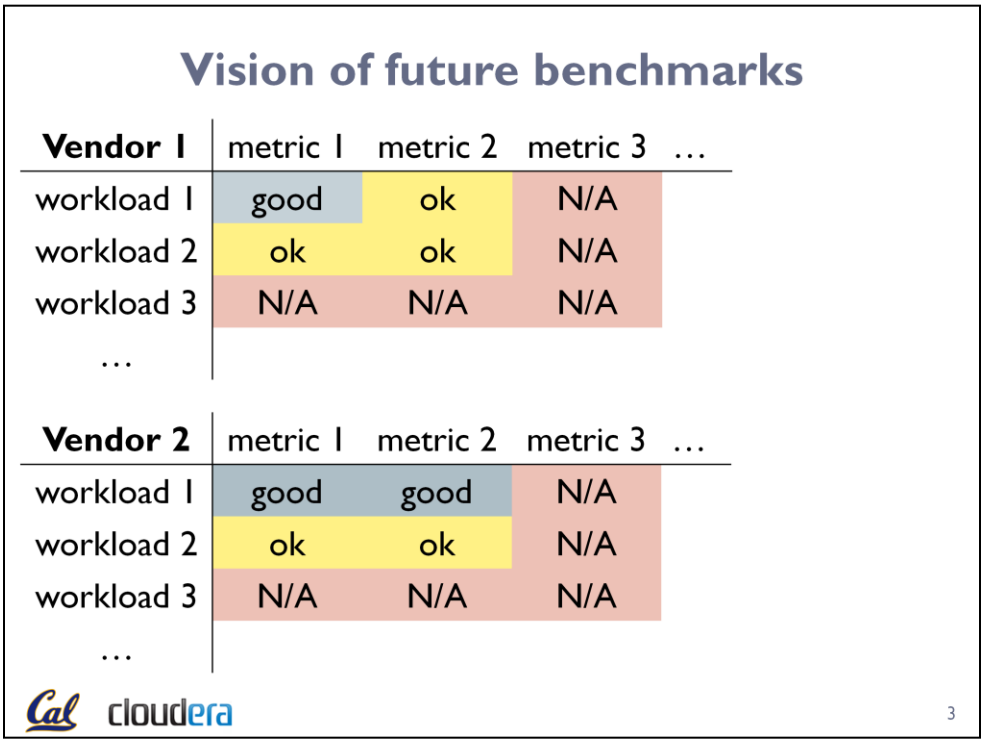
What do real workloads look like

- **Data access patterns**
  - skew in data accesses range between an
  - 80% of data re-accesses occur within
- **Load arrival patterns**
  - offered load over time is
  - peak-to-median ratio range from
- **Common job types**
  - over 90% of all jobs involve data sets of
  - job sizes, shapes, & frequencies are
- **SQL-like programming frameworks**
  - the cluster load from such frameworks are
  - workload analysis at the Hive/Pig/HBase levels is

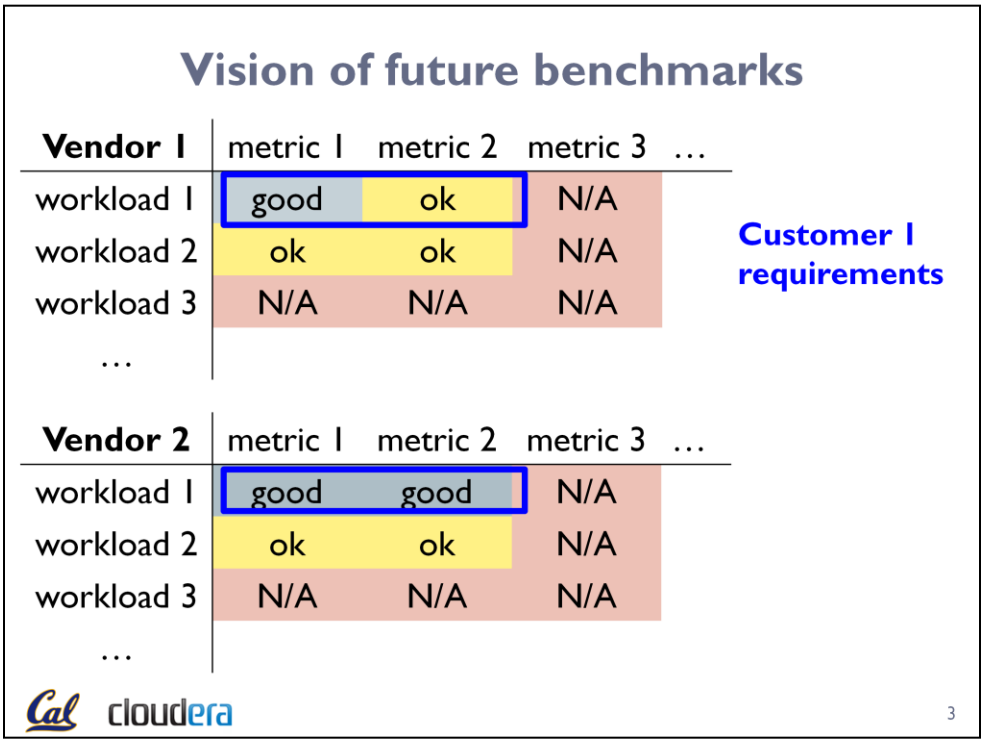
cloudera

What do real workloads look like

- **Data access patterns**
  - skew in data accesses range between an 80-1 and 80-8 rule
  - 80% of data re-accesses occur within minutes to hours
- **Load arrival patterns**
  - offered load over time is bursty and unpredictable
  - peak-to-median ratio range from 9:1 to 260:1
- **Common job types**
  - over 90% of all jobs involve data sets of KB to GB
  - job sizes, shapes, & frequencies are different for different workloads
- **SQL-like programming frameworks**
  - the cluster load from such frameworks are 20-80%
  - workload analysis at the Hive/Pig/HBase levels is non-existent

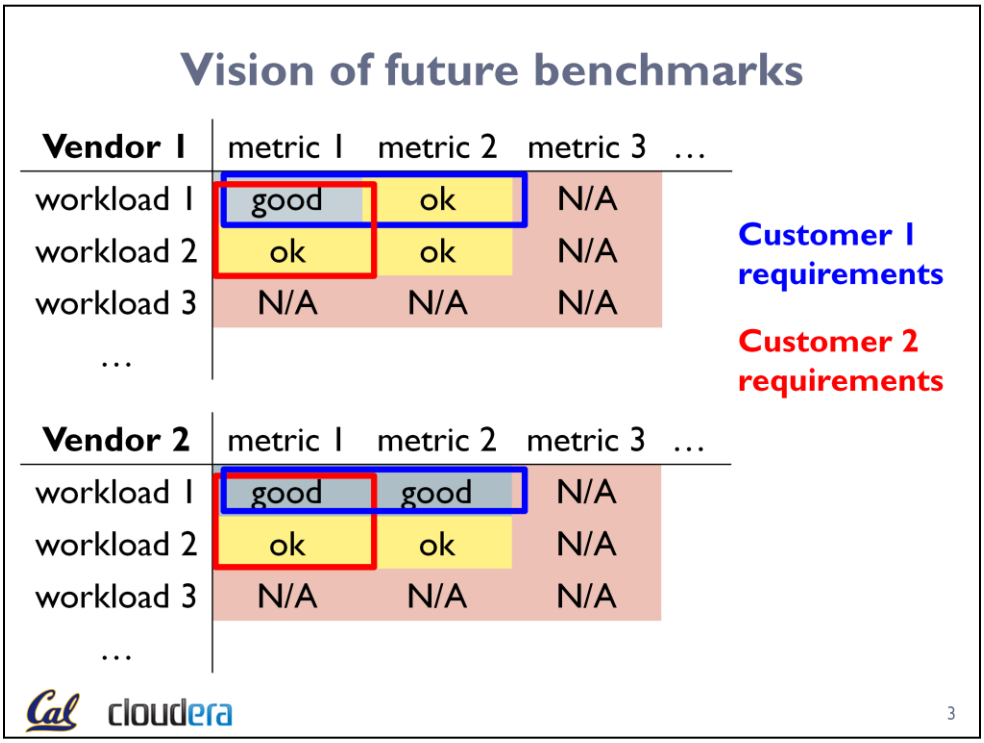


Two vendors run a benchmark of multiple workloads and multiple performance metrics.



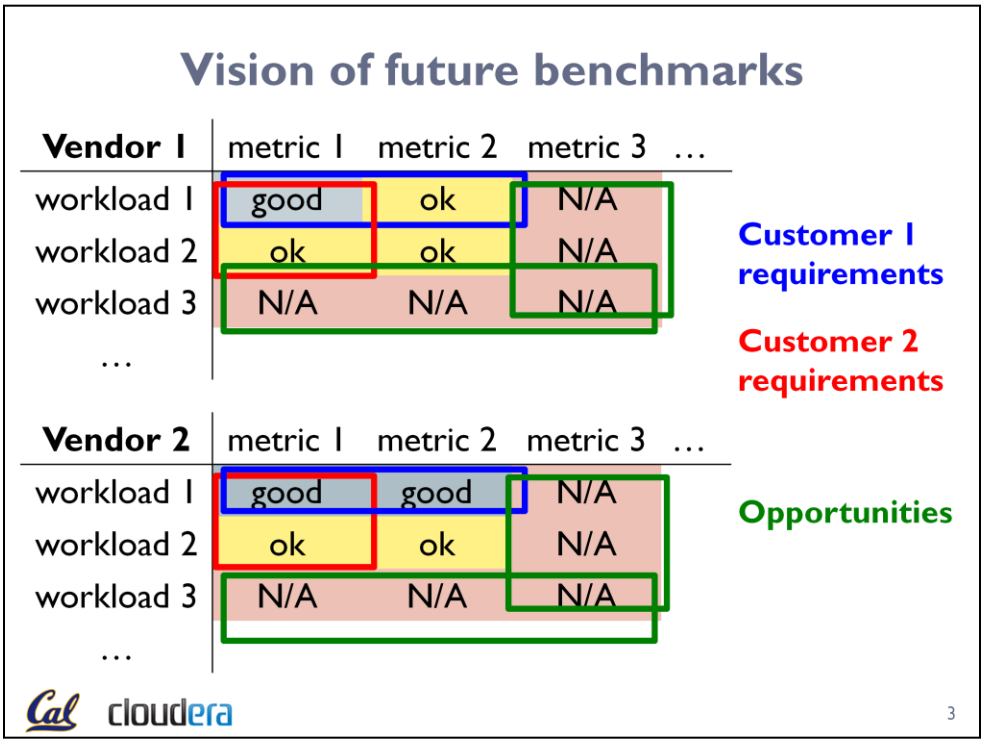
Customer 1 is interested in only Workload 1, but Metrics 1 and 2.

Vendor 2 is better for this customer.



Customer 2 has different requirements, and is interested in Workloads 1 and 2, but only Metric 1. The vendors are equal for Customer 2.

This is a different conclusion than that of Customer 1, and rightly so, since the two have different requirements.



The workloads and metrics that have not been addressed yet represent opportunities.

Such an inclusive benchmark covers diverse “big data” use cases, allows vendors differentiation along different workloads and metrics, help customer clarify their own requirements, and suggest un-addressed opportunities.