



HBA

Hadoop Benchmark Analytics

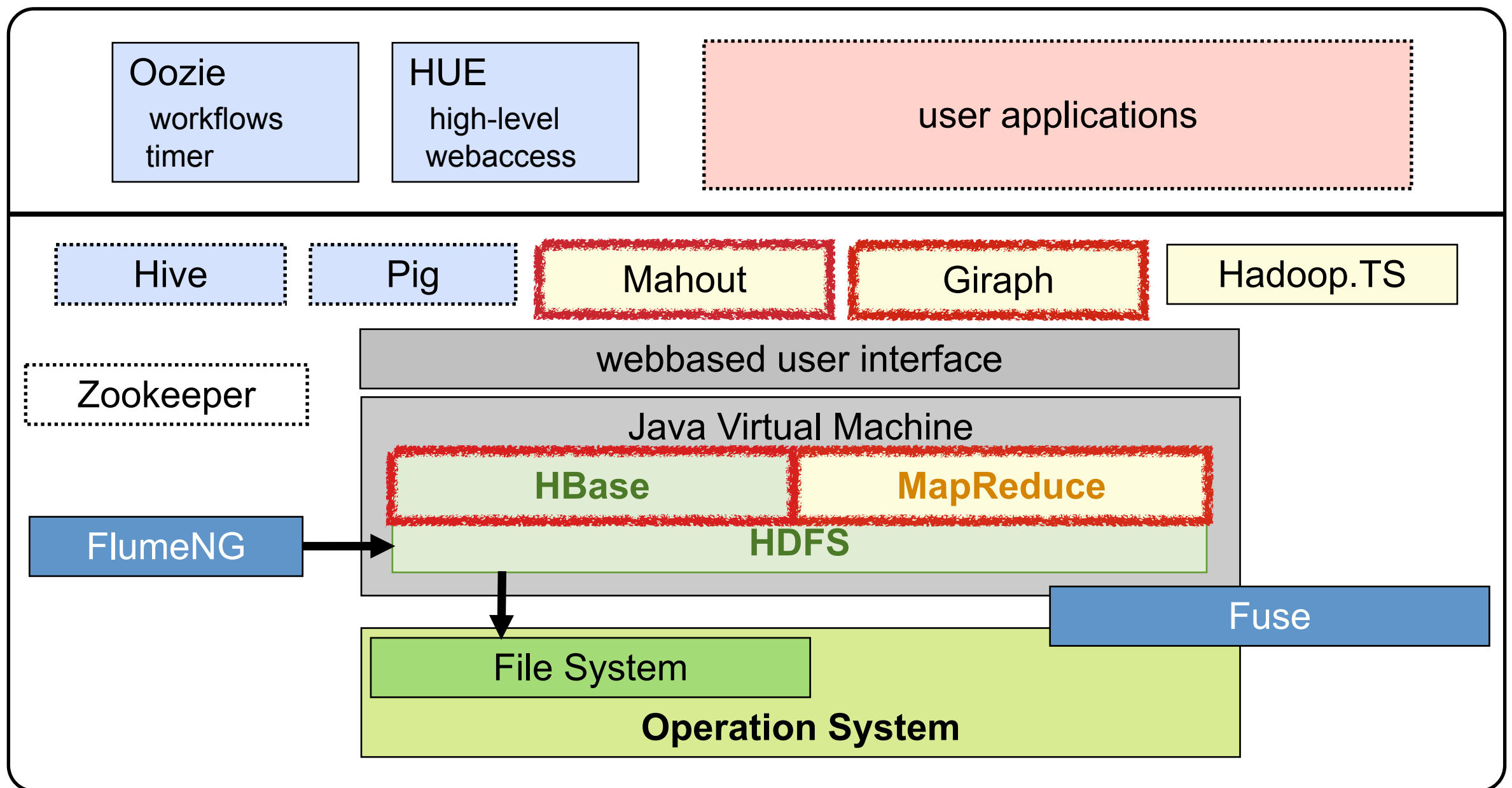
mirko.kaempf@gmail.com



Proposal for a crowd sourced
data collection platform ...

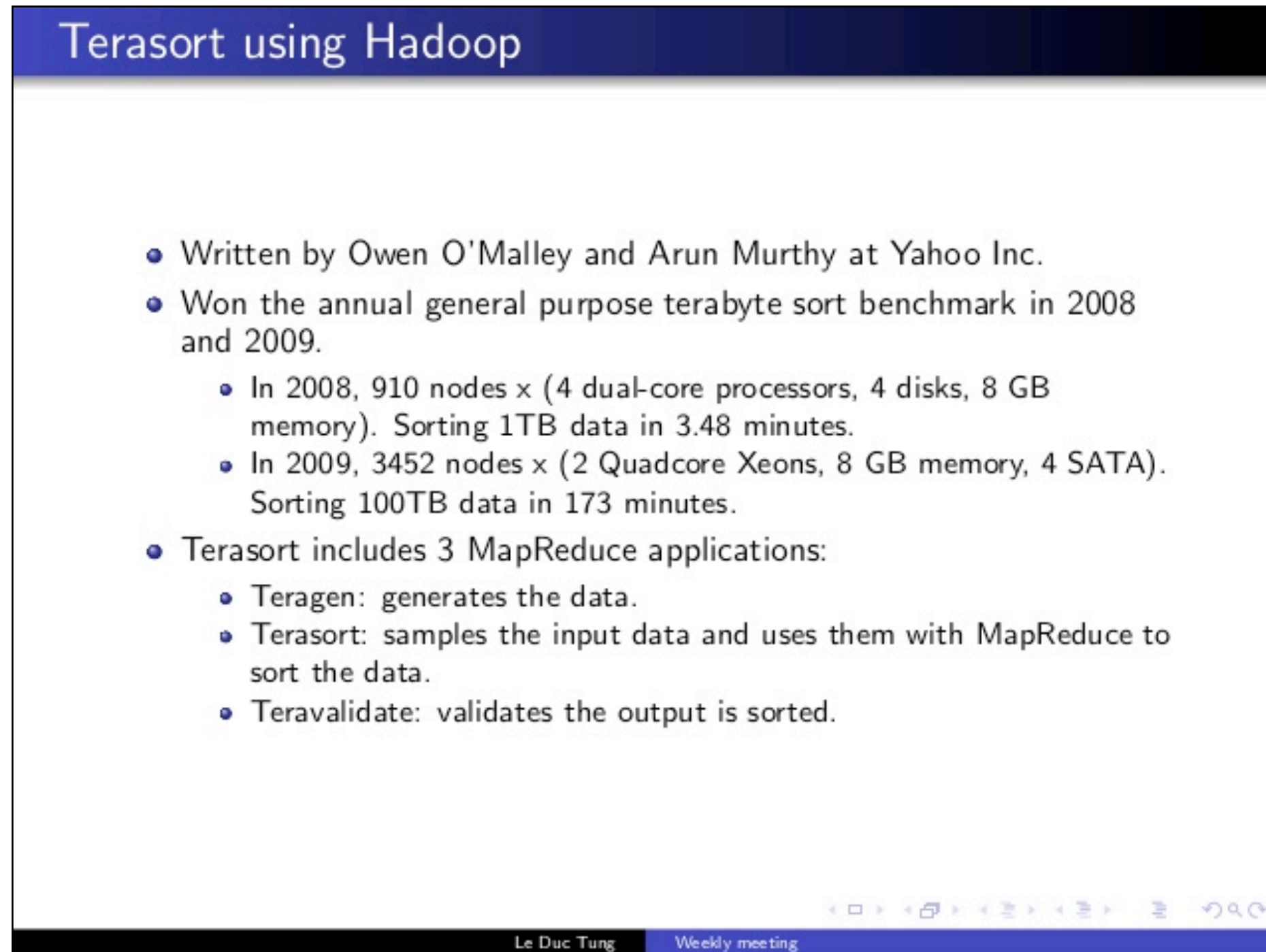
The Hadoop Ecosystem

Where we are?



Terra Sort

- easy benchmark for MR stream processing



The screenshot shows a presentation slide with a blue header bar containing the text "Terasort using Hadoop". The main content area is white and contains a bulleted list of information about Terasort. At the bottom of the slide, there is a black footer bar with the text "Le Duc Tung" and "Weekly meeting".

Terasort using Hadoop

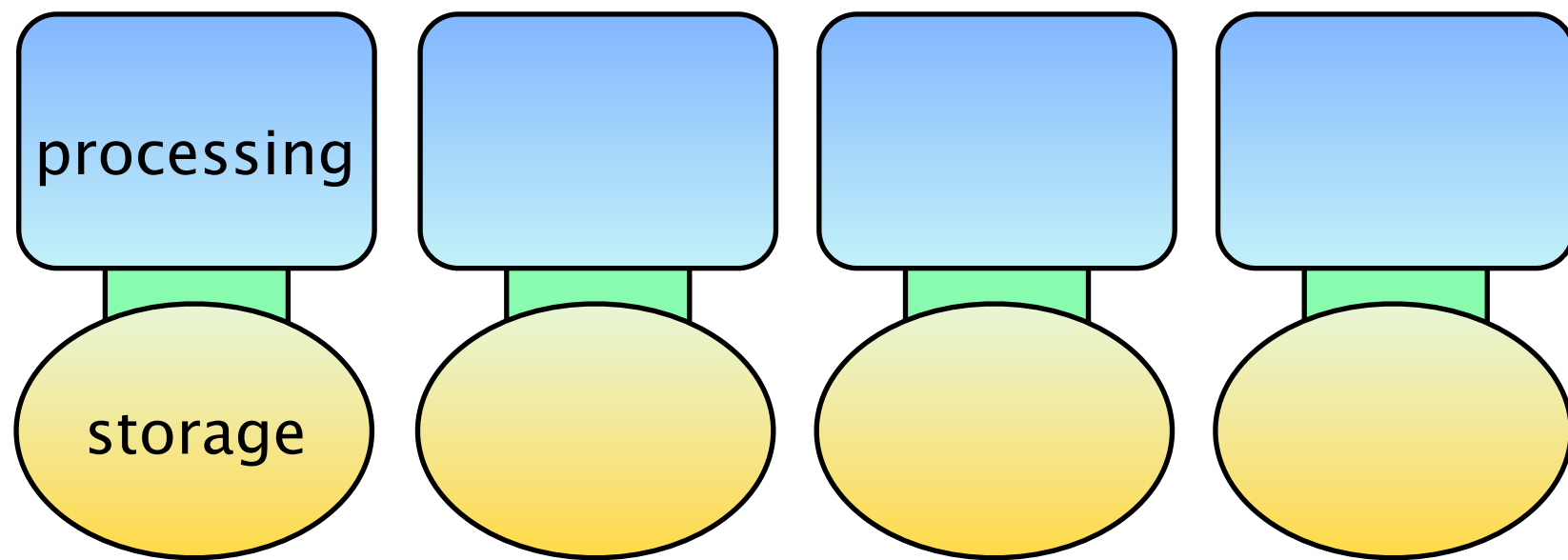
- Written by Owen O'Malley and Arun Murthy at Yahoo Inc.
- Won the annual general purpose terabyte sort benchmark in 2008 and 2009.
 - In 2008, 910 nodes x (4 dual-core processors, 4 disks, 8 GB memory). Sorting 1TB data in 3.48 minutes.
 - In 2009, 3452 nodes x (2 Quadcore Xeons, 8 GB memory, 4 SATA). Sorting 100TB data in 173 minutes.
- Terasort includes 3 MapReduce applications:
 - Teragen: generates the data.
 - Terasort: samples the input data and uses them with MapReduce to sort the data.
 - Teravaldiate: validates the output is sorted.

Le Duc Tung Weekly meeting

<http://www.slideshare.net/tungld/terasort>

Terra Sort

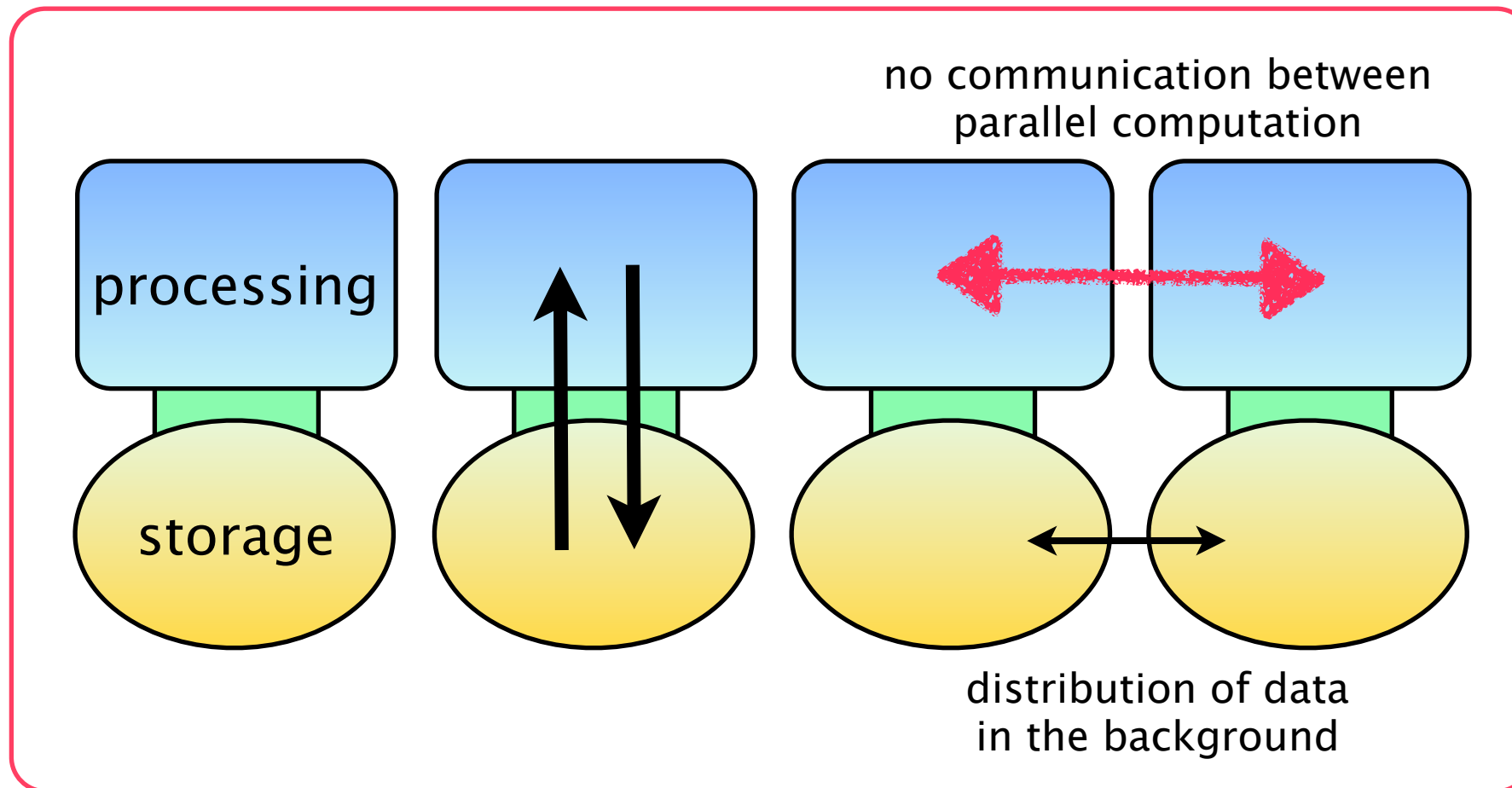
- easy benchmark for MR stream processing



Terra Sort

no general benchmark, but
a good starting point

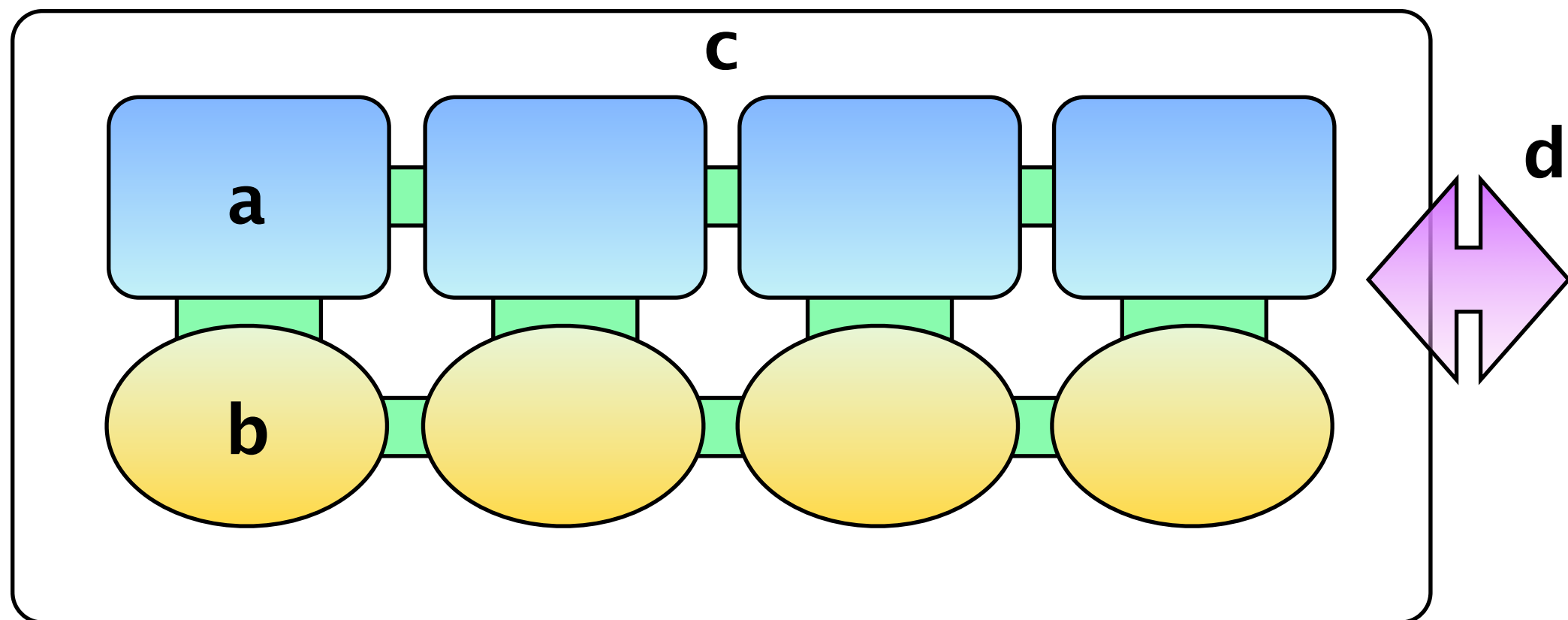
- just MapReduce stream processing



- no message passing between mappers
- limited to data processing within just one cluster ...
- **but** : we want to deal with large **linked data sets**!

What to measure?

- (a) **Processing capacity & speed**, depending on selected algorithms
- (b) **Storage efficiency**, depending on selected technology
- (c) **Intra-Cluster communication**, between processes
- (d) **Inter-Cluster communication**, between clusters





Crowd sourced data collection ...

- Define, create and run benchmarks
- Publish & share cluster profile data
- Publish & share job-log data (benchmark results???)
- Do statistical analysis on public data
- Build models for optimization

Apache Vidya, Apache Gridmix, and Starfish are established projects,
but: isolated!

Hadoop Benchmark Analytics ... is on the way to become a community platform
for collecting public benchmark data, sharing results and for collaborative research.

Collaboration ...

- Hadoop, is not (yet) a typical  RDF technology ... but benchmark related work is already done.
- Maybe the mentioned projects can be a contribution to **LDBC**  in the future.