Lambada: Interactive Data Analytics on Cold Data using Serverless Cloud Infrastructure

Is serverless attractive for data analytics?

Simulation of scanning 1 TB from cloud storage

~~ Only for interactive use!

We built Lambada to find out more!

Shared-storage database architecture with only serverless components

Challenge: low-latency burst invocation

- 2-level invocation solves driver bottleneck
- Invoke 4k serverless workers in < 3 s

Result: scan-heavy queries are interactive

TPC-H Q1 @ SF 10k, 1.5 TiB Parquet files, 3200 workers

~~ Outperforms commercial systems in speed (2 - 1000×) and price (10 - 100×)

Challenge: shuffle through cloud storage

- Workers can only communicate through cloud storage
- Prior work: “serverless shuffle unfeasable”

Result: novel serverless shuffle algorithm

- Shuffle in two levels needs $O(P\sqrt{P})$ IOs

~~ Purely serverless shuffle is competitive

I. Müller, R. Marroquín, G. Alonso