Data analytics on cold storage

Keywords: DBMS, Data Analytics, Cold Data, Rack-scale Storage

Problem:
Driven by the desire to extract insight out of data, and empowered by the cost-effectiveness of hosted data analytics services, businesses have started aggregating vast amounts of data. As these data lakes continue to grow in size, it is inevitable that a significant fraction of this data will be infrequently accessed. This data is referred to as cold data. Given the demand for increased storage capacity, Cloud Service Providers (CSP) have started designing and deploying custom-built, rack-scale Cold Storage Appliances (CSA) that provide extremely low cost data storage ($0.01/GB/month) by designing hardware explicitly to serve cold data workloads (with access latencies in tens of seconds) [1,2,3,4].

As such cold-storage-as-a-service offerings are quickly gaining popularity both in public and private clouds, it is inevitable that they will become the new data lake for data-analytics platforms. To this end, we have designed and implemented Skipper, an end-to-end Database-as-a-Service architecture that enables running cloud-hosted relational databases directly on CSA. The Skipper architecture proposes novel changes to database execution engine, buffer cache management, and CSA I/O scheduling algorithms to reduce the performance penalty associated with using CSA. Compared to traditional cloud-hosted databases, our Skipper implementation provides a 10x reduction in execution time, enabling a new class of cheap, cloud-hosted data analytics services possible.

We are now looking for motivated students to explore several interesting research directions based on Skipper. In particular, we are planning to investigate strategies that 1) minimize the number of CSA requests (caching, indexing), 2) reduce the amount of data transferred (columnar storage and compression), and 3) optimize data layout on CSA. If you are interested in helping out (as Master project or Semester project) and advancing this fascinating project, please do not hesitate to contact us.

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch
Responsible collaborator(s): Dr. Raja Appuswamy, raja.appuswamy@epfl.ch & Renata Borovica-Gajic, renata.borovica@epfl.ch
Duration: One semester
References


