Serverless Query Processing on a Budget

William Ma, RISE Lab, University of California, Berkeley

Goal: How can we use serverless techniques to improve query processing?

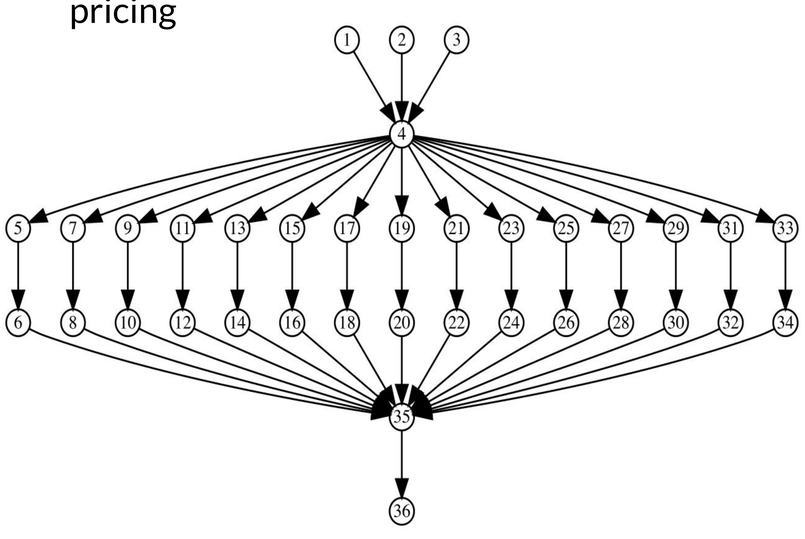
What is Serverless?

Benefits of serverless for query processing:

- Automatic cluster provisioning
- Infinitely-scalable elastic cluster
- Millisecond-based pricing

What is needed in serverless for query processing:

- New field without established systems



Spark query plan for TPC-DS Query 9

Current Serverless

Benefits of serverless for query processing in current offerings:

- Built on established systems (e.g., Presto)
- Automatic cluster provisioning

What is needed in serverless for query processing in current offerings:

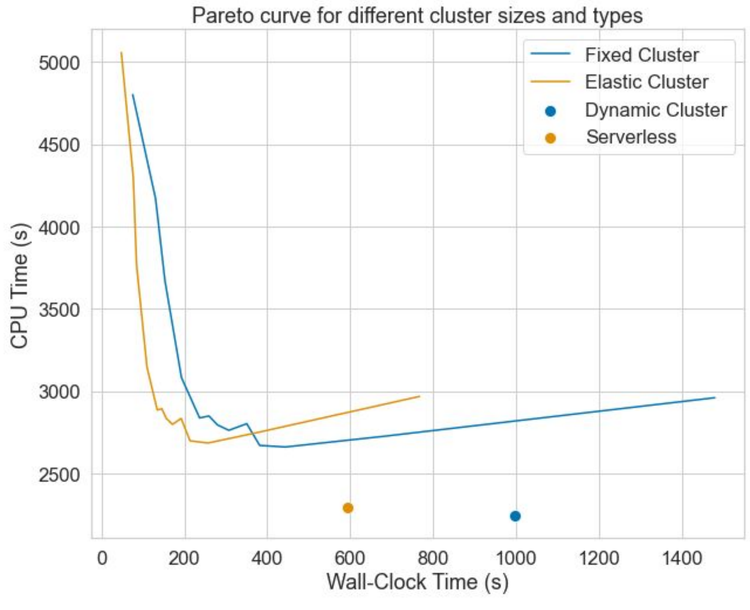
- Infinitely-scalable elastic cluster
- Millisecond-based pricing

Query	Wall-Clock Time	Cost
2 `SELECT` queries	2 minutes	114 GB x $5/TB = $0.72
1 `CROSS PRODUCT` query	>30 minutes	114 GB x $5/TB = $0.72

Table of results from 2 sample queries on GCP BigQuery

Performance

To demonstrate the performance of serverless computation in query processing, we use a representative **query from a Spark data science tutorial** on a **5 GB data set** containing the HTTP connections to a NASA website on clusters of AWS EC2 m5.large (**2 CPU & 8 GB RAM**) of sizes varying from **2 nodes to 64 nodes**.



With an elastic cluster (naively replicating the cluster to parallelize the query) leads to a 50% decrease in CPU time (proxy for cost). Dynamic cluster (serial execution but dynamic cluster size) significantly reduces the minimum CPU time to execute query. Serverless combines both for more performance.

Modeling