Serverless Scalable Data Engineering with Cylon and Amazon Web Services

Mills Staylor
Advised by:
Geoffrey Fox
Yue Cheng

Department of Computer Science
University of Virginia
Email: qad5gv@virginia.edu
Introduction

• Data Science domain has expanded monumentally over the past few decades
• Main drivers have been the BigData revolution, AI, ML

“Significant developer time is spent on data exploration, preprocessing, and prototyping”

-- The State of Data Science 2020, anaconda.com

• Improving performance & scalability is crucial
• Enables building efficient data engineering pipelines
Cylon: A High Performance Distributed Data Table

- **Cylon** is a high performance C++ kernel and a distributed runtime for big data processing supporting operator parallelism
  - Apache Parquet and Arrow based storage and in-memory data structure
    - Supports integration with Deep Learning workloads, Pandas and Numpy
    - Zero-Copy data transfer between heterogeneous systems and languages.
- **Table API**, an abstraction for ETL (extract, transform, load) for scientific computing and deep learning workloads including Pandas, HDF5
  - Join, Union, Intersect, Difference, Product, Project … ~40 operators
Redis Process Bootstrapping

- Information needed: world_size, rank, address of each communication endpoint
  - world_size: given by the environment
  - rank: generated with an atomic increment operation
  - Other endpoints’ addresses: emulated all_gather operation
**Novel High Performance Serverless Data Engineering on AWS**

Cylon Scheduler

- AWS Step Function
- Generate Lambda Payload
- Invoke AWS Cylon Lambda

Storage

- Redis Elasticache Key/Value Store
- Amazon S3

Rendezvous Server

- Initialization/Configure/Persist
- Final Results

User

- Cylon Scheduler
- Storage
- Rendezvous Server

Cylon Serverless Task Execution

- AWS Lambda/FaaS

Cylon Dependency Provided By

- AWS Step Function
- Amazon S3
- Redis Elasticache Key/Value Store

AWS Lambda/FaaS

- Initialization/Configure/Persist
- Final Results

ECS Task Registry

- Cylon Serverless Task Execution

Cylon

- Lambda/FaaS
- Initialize/Configure/Persist
- Final Results

Mills Staylor

May 29, 2024
Experiment Results

Strong Scaling of Join Operation

- EC2 - 4 CPU/26 GB Memory
- EC2 - 16 CPU/28 GB Memory
- Fargate - 4 CPU/26 GB Memory
- Fargate - 16 CPU/28 GB Memory
- Rivanna - 8 CPU
- Rivanna - 16 CPU

average time (s)

145m rows (1/2/4/8/16 Nodes)

parallelism