# 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"*

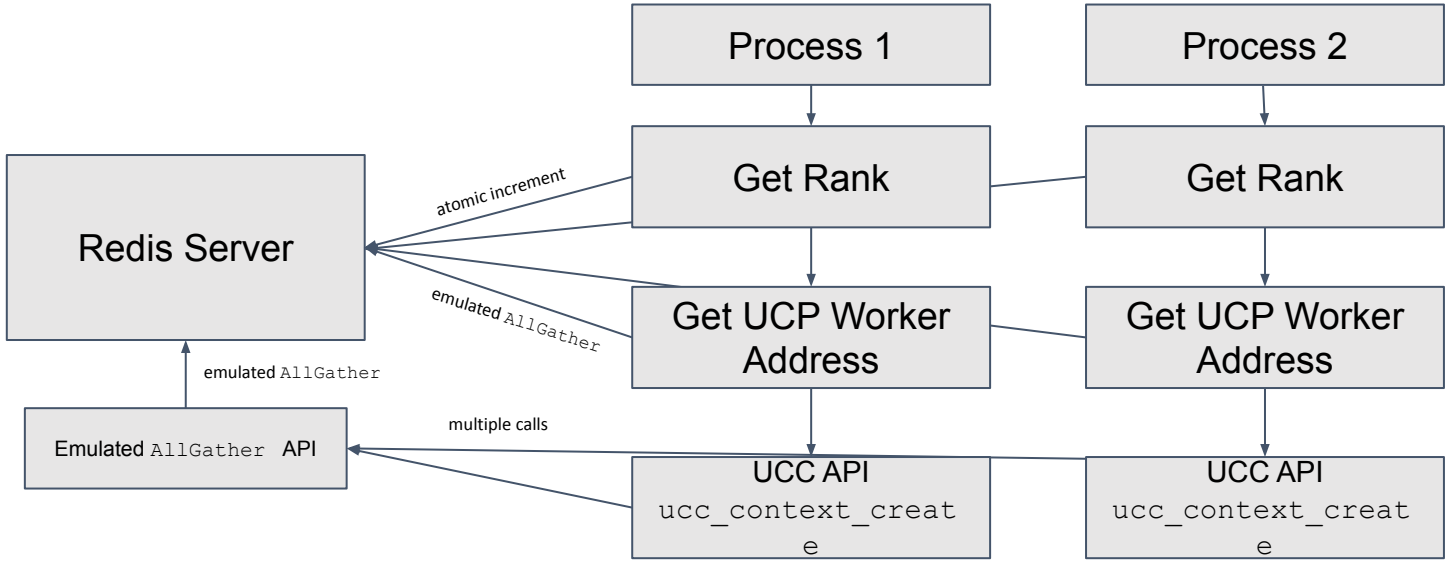
*-- The State of Data Science 2020, anaconda.com*

- Improving performance & scalability is crucial
- Enables building efficient data engineering pipelines

# Cylon: A High Performance Distributed Data Table

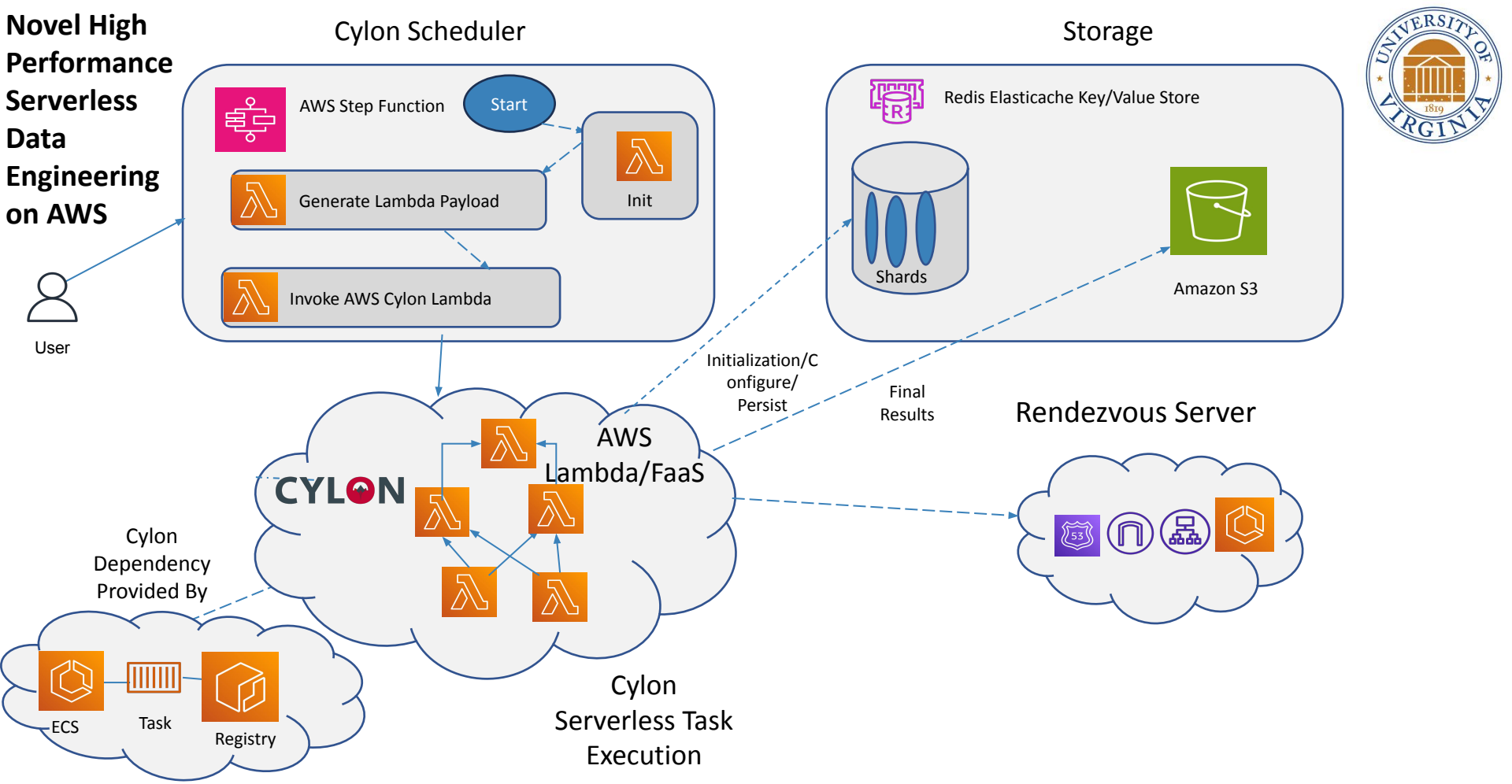

- **Cylon** is a high performance C++ kernel and a distributed runtime for **big data processing supporting operator parallelism**
  - Apache Parquet and Arrow based storage and in-memory data structure
    - Supports integration with Deep Learning workloads, Pandas and Numpy
    - Zero-Copy data transfer 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



# Experiment Results