CosmicAI: Scalable AI Redshift Inference

DS 5110: Data Engineering II – Big Data Systems

Group Members: Lionel Medal and Vicky Singh

Project Overview

Scientific Need

- Modern astronomy missions generate massive image datasets that require scalable, automated analysis

Proposed Solution

- Cloud-based Astronomy Inference (CAI) enables distributed redshift prediction using a pretrained model in a serverless architecture

System Design

- Processes partitioned data in parallel using Amazon Web Services Step Functions and Lambda Functions

Project Objective

- Build a reproducible, cost-efficient pipeline for high-throughput inference on large scientific workloads

Dataset and Preprocessing

Dataset Source

- Astronomy image data for redshift prediction, sourced from a shared Google Drive repository
- Total Dataset Size: ~12.6 GB

Data Format

- Stored as serialized PyTorch tensors (.pt files), each representing multi-channel astronomy images

Preprocessing Steps

- Resized images to 32x32 resolution
- Selected first 5 channels (out of 64) for input
- Partitioned into 25-100 MB chunks for parallel processing
- Uploaded to Amazon S3 for serverless access

Pipeline Architecture

Initialize

- Generates job configurations from the input payload, including batch size, file limits, and path

Distributed Inference

- Runs parallel Lambda containers to perform model inference on partitioned data

Synchronize

- Uses a rendezvous server to enable FML-based communication between Lambdas

Summarize

- Aggregates JSON output files and produces a combined results file in Amazon S3

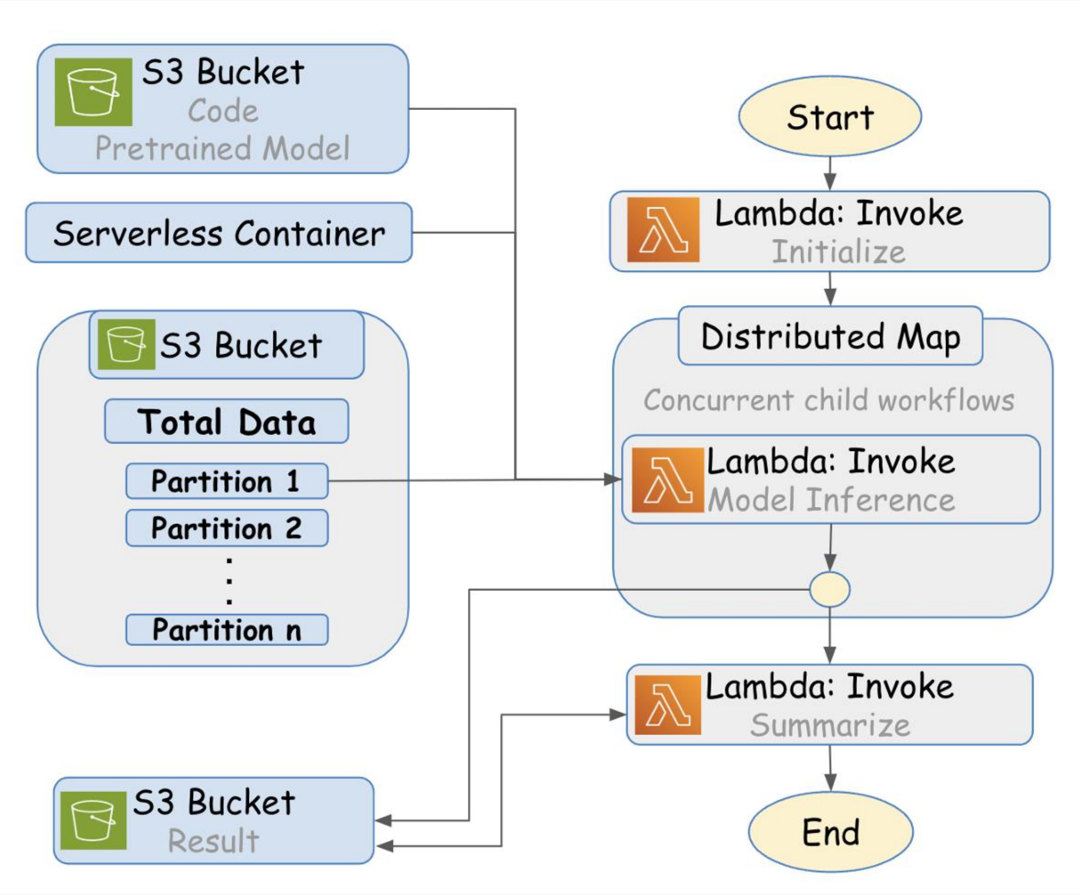


Figure 1. CAI Framework Design on AWS State Machine

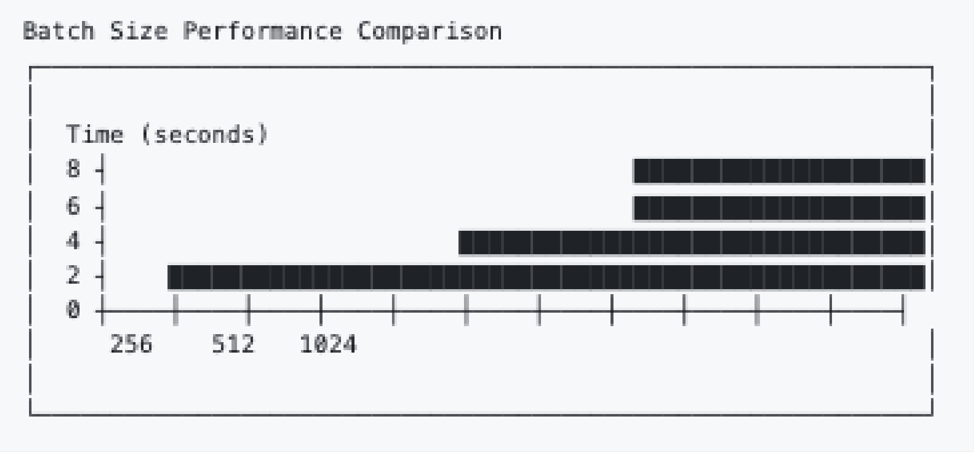
Benchmarking Results

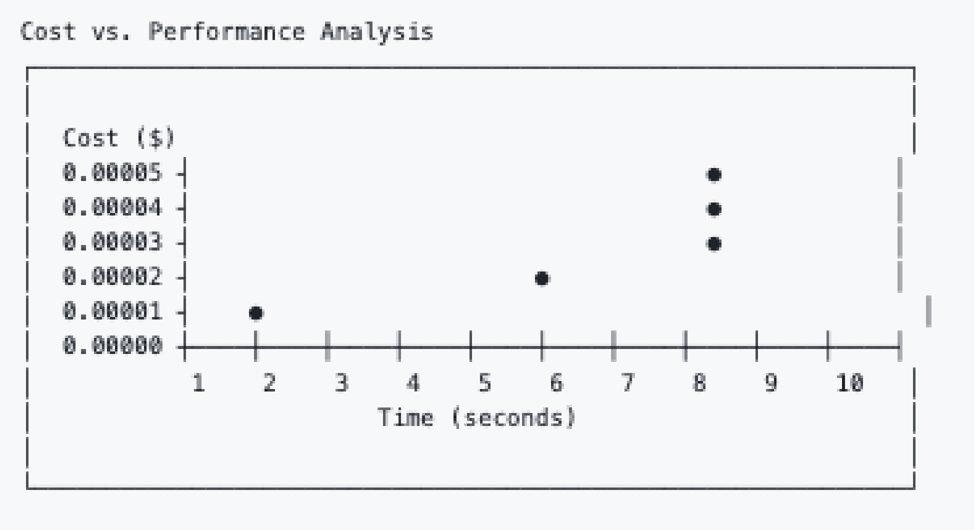
AWS Lambda Performance Metrics

- Execution Time: 1.23s - 7.56s per batch
- Memory Usage: 14,325-14,335 MB
- Throughput: 135-208 samples/second
- Cost: $0.000012 - $0.000045 per execution

Key Findings

- Batch size 256: Fastest execution, highest cost
- Batch size 512: Optimal balance
- Batch size 1024: Slowest, most expensive





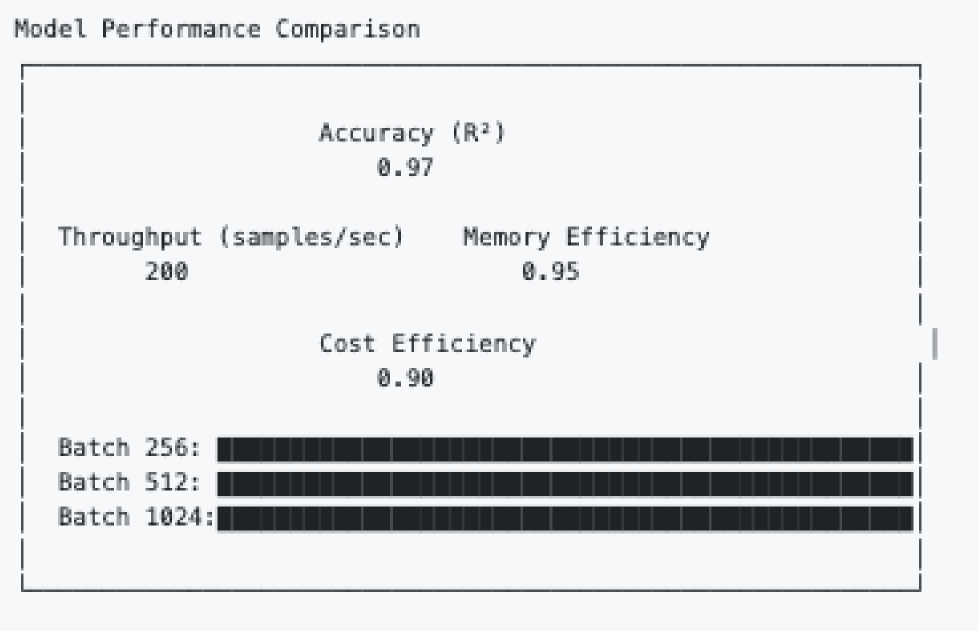
Performance Results

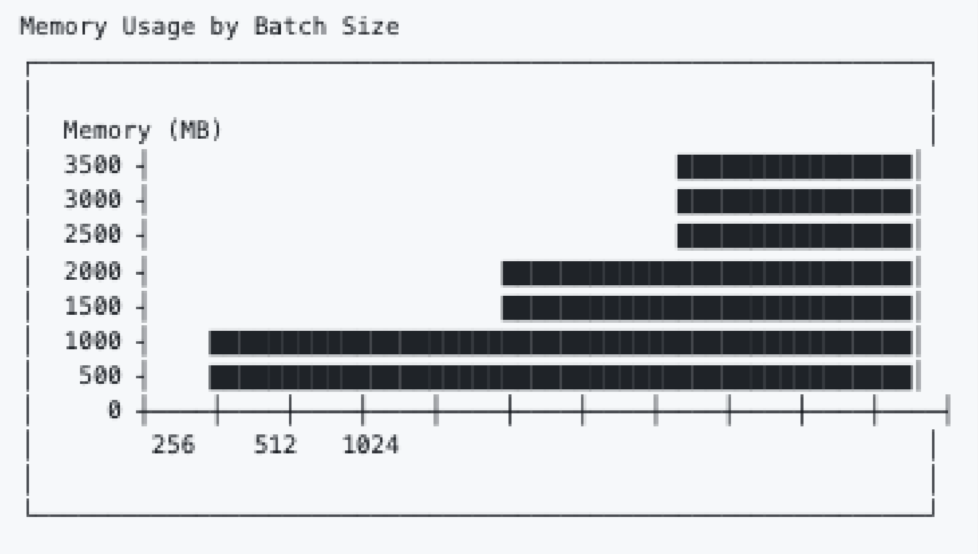
Analysis

- Batch 256: 208 samples/second (Fastest)
- Batch 512: 174 samples/second (Balanced)
- Batch 1024: 135 samples/second (Slowest)

Memory Efficiency

- Batch 512: Lowest memory utilization (678 MB max)
- Batch 256: Moderate memory usage (1,985 MB max)
- Batch 1024: Highest memory usage (3,420 MB max)





Conclusion and Impact

Relevance

- Validates serverless AI as a scalable solution for large scientific datasets
- Framework is reusable across domains requiring distributed inference

Improvements

- Integrate GPU-based inference for improved speed and accuracy
- Optimize Lambda configurations to reduce cold starts and runtime

Opportunities to Expand

- Adapt CAI for real-time or streaming data pipelines
- Apply framework to additional fields: medical imaging, climate science, geospatial analytics, etc.