Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Damien Jones (he/him)

AWS Consultant @ Steamhaus

2024-07-11 AWS UG Liverpool

MrDamienJones

MrDamienJones

amazonwebshark.com

damien@amazonwebshark.com

@amazonwebshark

@amazonwebshark

https://www.flaticon.com/

Here For
Data & Analytics?

Here For
Development & Operations?

Here For The Scran?

Damien Jones

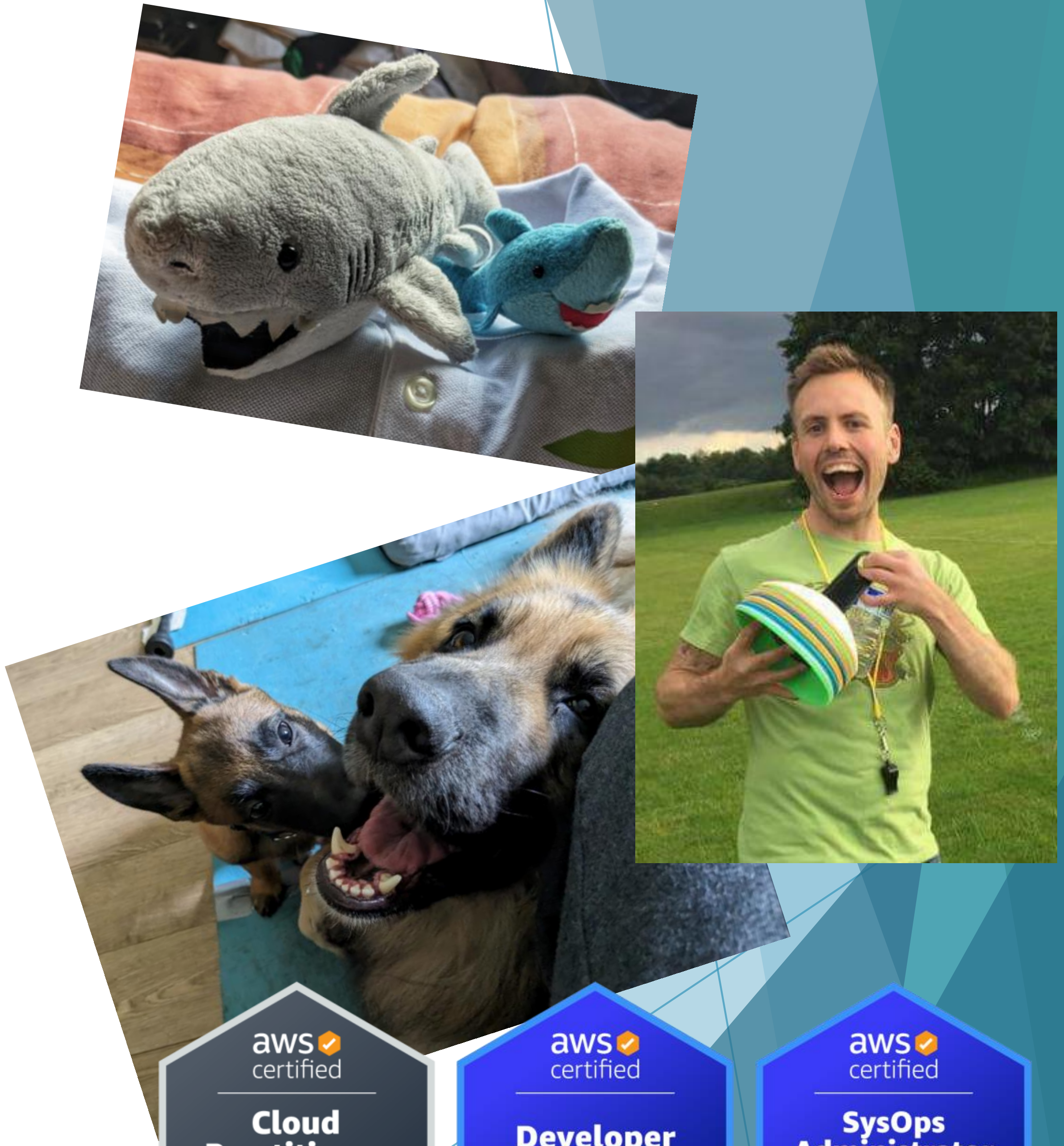
♂️ He/Him 🌍 Manchester UK 🦈 Fin Fan

Consultant @ Steamhaus

Using AWS since 2019

Creator @ amazonwebshark.com

Runner; Keen Gardener; Dog Dad



aws certified
Cloud Practitioner
FOUNDATIONAL

aws certified
Developer
ASSOCIATE

aws certified
SysOps Administrator
ASSOCIATE

Agenda

Problem Definition

Solution Architecture

Demo

Summary & Questions

github.com/MrDamienJones
/Community-Sessions

The 4 Vs Of Big Data

Characteristics of Big Data…

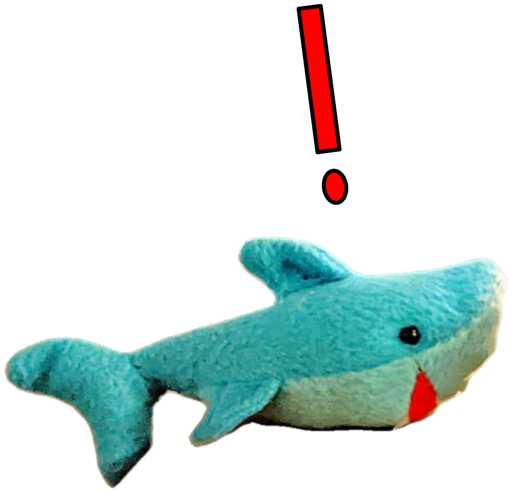
…and events…

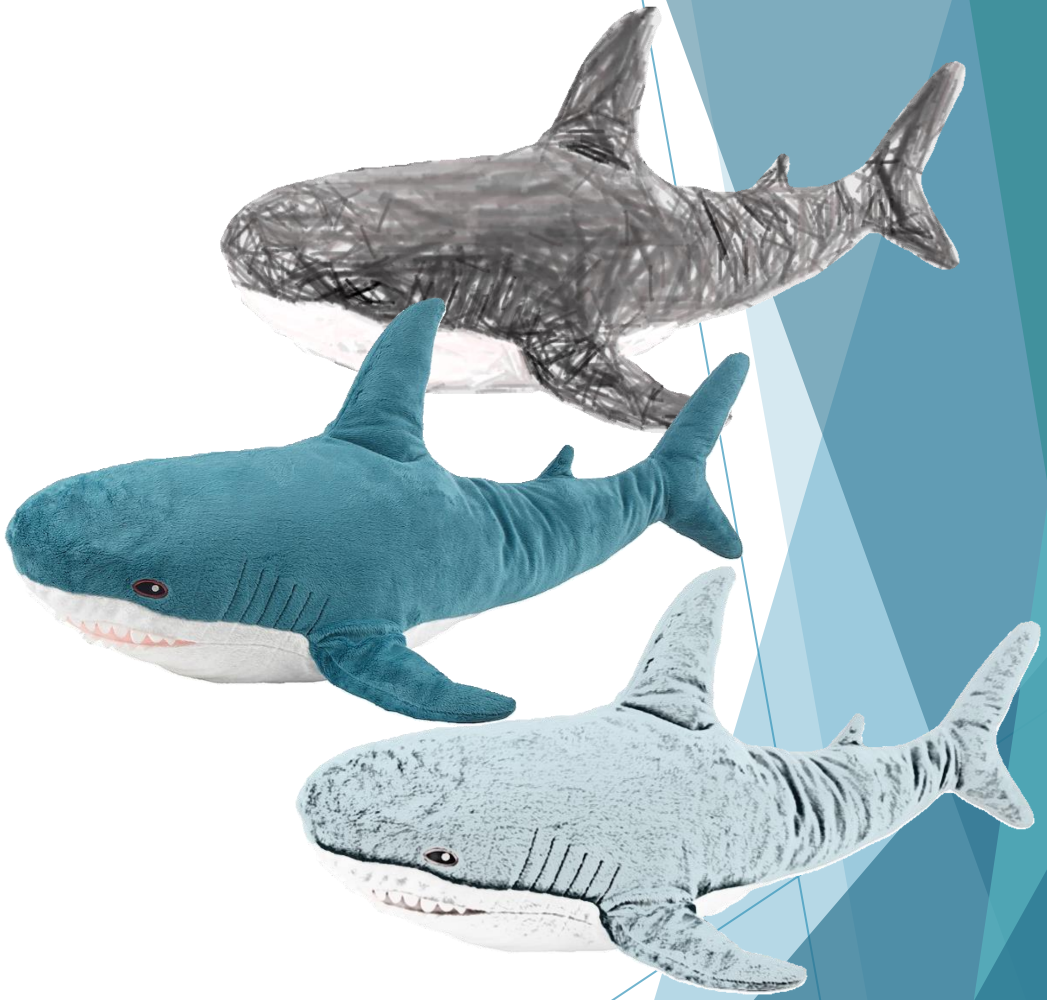
…and API requests…

…metrics …traces …logs . . .

Variety

"The state of being diverse or varied."





Variety

“The state of being diverse or varied.”

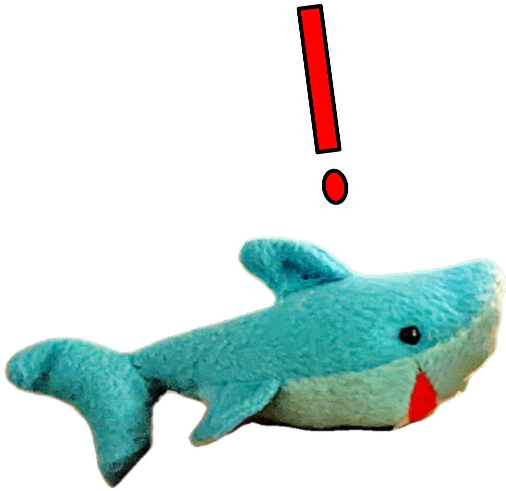
Structure

Purpose

Sensitivity

Velocity

“The speed at which something is moving in a given direction.”



Velocity

“The speed at which something is moving in a given direction.”

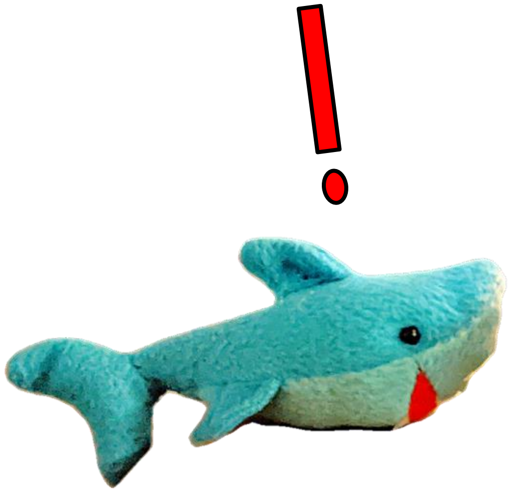
Streaming or Batch

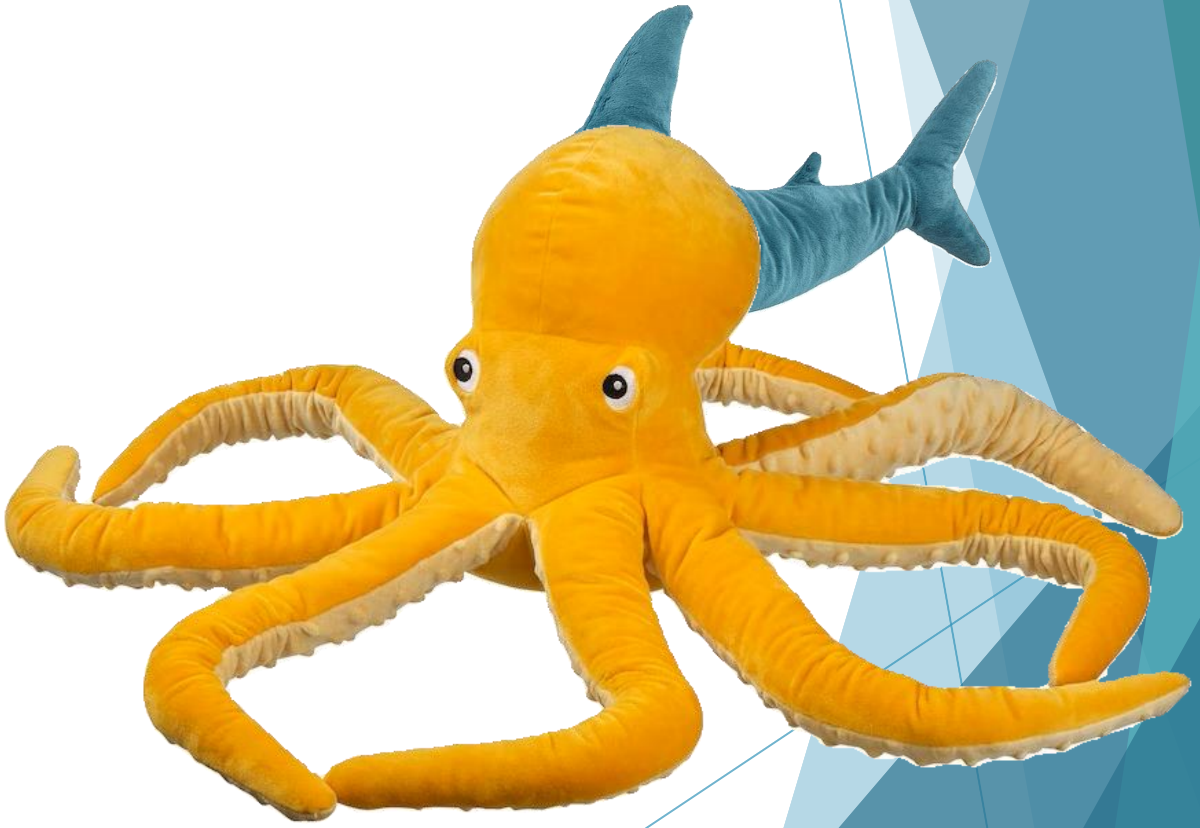
Synchronous or Asynchronous

Scheduling

Veracity

“The quality of being true or the habit of telling the truth.”





Veracity

“The quality of being true or the habit of telling the truth.”

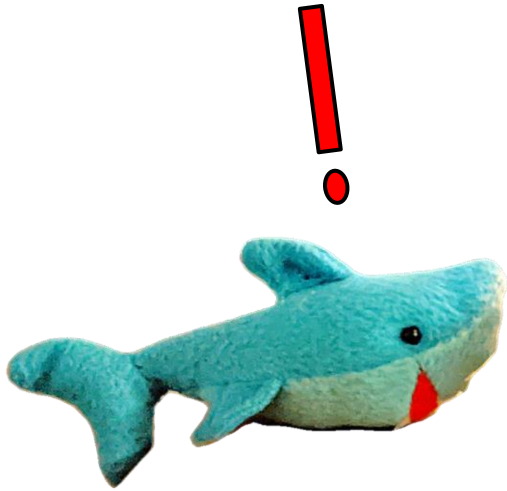
External Security

Validation & Health

Internal Security

Volume

“The amount of space occupied.”



Volume

“The amount of space occupied.”

Size or Amount

Access Patterns

Backups

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

AWS Lambda

Amazon S3

AWS Lambda

Serverless compute service

Supports multiple languages

Auto-scales on demand

Up to 1000 concurrent executions

Amazon S3

Serverless object storage

Store anything for any reason

1000s of requests per second

Object protection & integrity checks

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

AWS Glue

Amazon Athena

AWS Glue

Fully managed serverless ETL service

Crawlers discover data automatically

Up to 2000 concurrent ETL job runs

ML-backed data quality checks

Amazon Athena

Serverless interactive query service

Analyse Amazon S3 data with standard SQL

Source data is read-only

Create derived tables

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Amazon EventBridge Scheduler

AWS Step Functions

Amazon EventBridge Scheduler

Automate recurring & one-off tasks

Invoke over 220 AWS services

Set times or fixed-rate schedules

Checks target response

AWS Step Functions

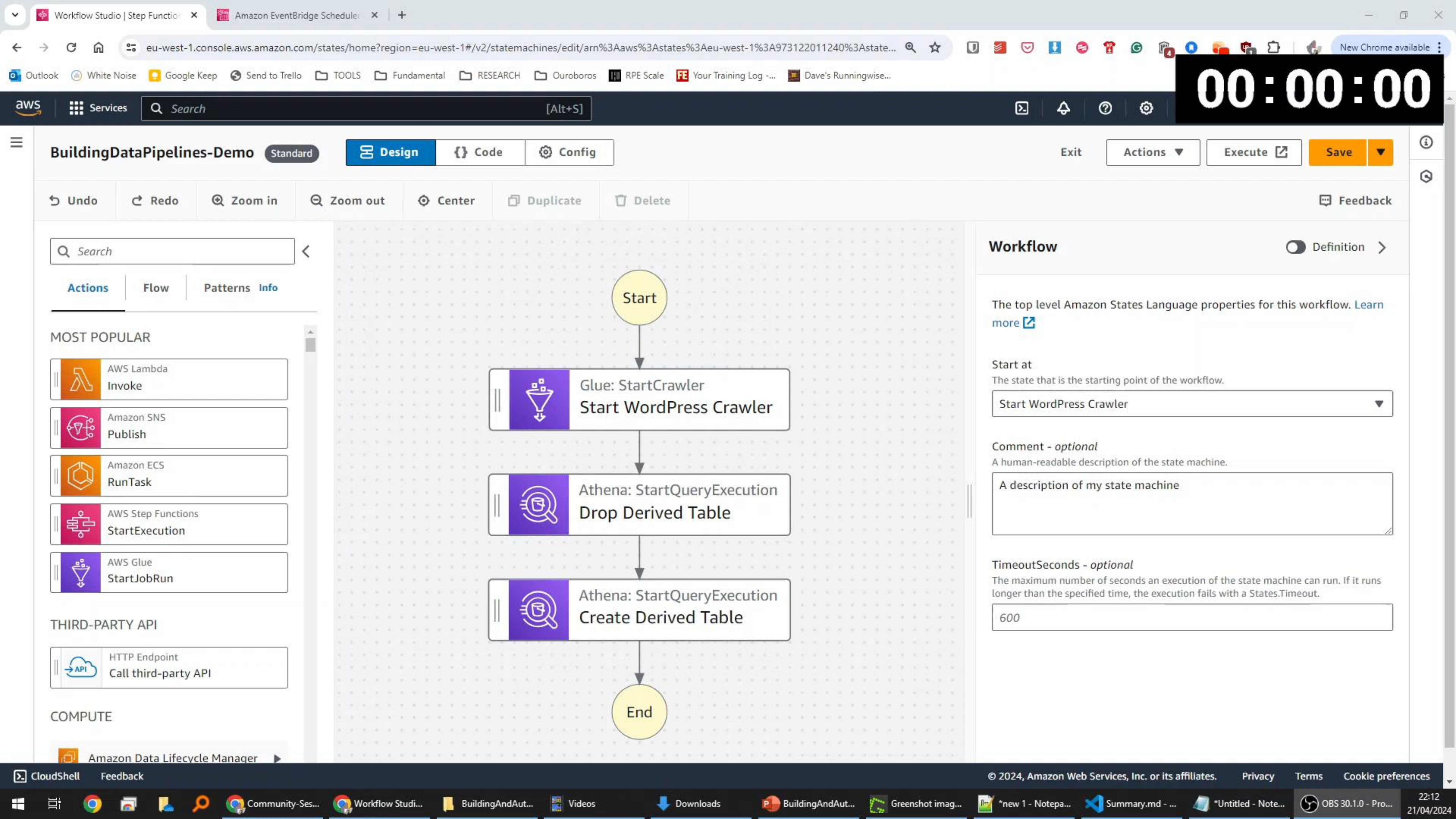
Serverless task orchestration

Invoke over 220 AWS services

Design workflows visually and as code

Standard & Express workflows

Demo



BuildingDataPipelines-Demo Standard

The top level Amazon States Language properties for this workflow. Learn more

Start at
The state that is the starting point of the workflow.
Start WordPress Crawler

Comment - *optional*
A human-readable description of the state machine.
A description of my state machine

TimeoutSeconds - *optional*
The maximum number of seconds an execution of the state machine can run. If it runs longer than the specified time, the execution fails with a States.Timeout.
600

Build an AWS CodeBuild Project

Description

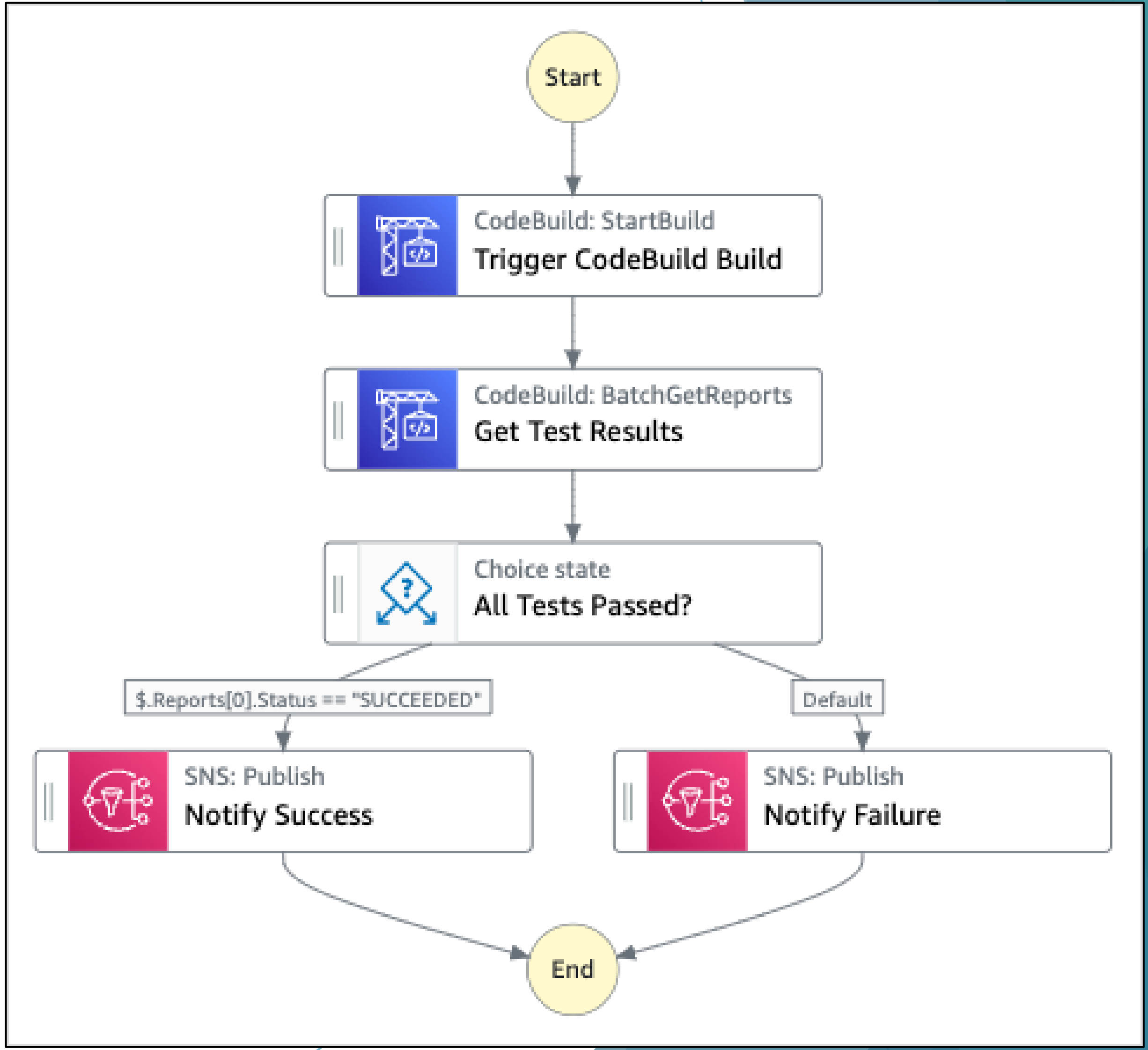
Build a CodeBuild project and send a notification based on the test results.

Documentation Link

https://docs.aws.amazon.com/step-functions/latest/dg/sample-project-codebuild.html

Services

CodeBuild, SNS



Tune a machine learning model

Description

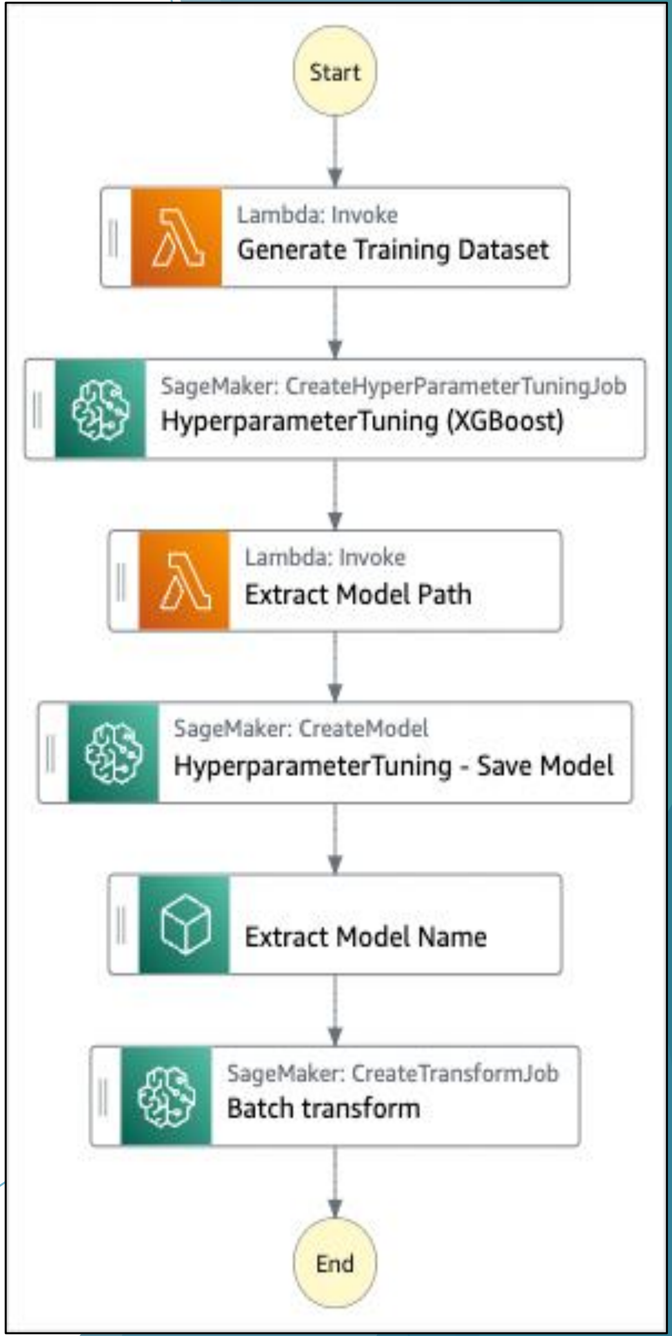
Tune hyperparameters of a machine learning model and batch transform a test dataset.

Documentation Link

https://docs.aws.amazon.com/step-functions/latest/dg/sample-hyper-tuning.html

Services

Lambda, S3, SageMaker



Start
Lambda: Invoke
Generate Training Dataset
SageMaker: CreateHyperParameterTuningJob
HyperparameterTuning (XGBoost)
Lambda: Invoke
Extract Model Path
SageMaker: CreateModel
HyperparameterTuning - Save Model
Extract Model Name
SageMaker: CreateTransformJob
Batch transform
End

Summary

Problem Definition

Solution Architecture

Demo

Summary & Questions

github.com/MrDamienJones
/Community-Sessions

Thanks!

MrDamienJones

MrDamienJones

amazonwebshark.com

damien@amazonwebshark.com

@amazonwebshark

@amazonwebshark

github.com/MrDamienJones
/Community-Sessions