

# Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Damien Jones (he/him)

AWS Consultant @ Steamhaus

*2024-11-27 AWS User Group Sheffield*



MrDamienJones



MrDamienJones



amazonwebshark.com



damien@amazonwebshark.com



@amazonwebshark



@amazonwebshark

# Agenda

Problem Definition

Solution Architecture

Demos

Summary & Questions



[github.com/MrDamienJones  
/Community-Sessions](https://github.com/MrDamienJones/Community-Sessions)

# AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

# Damien Jones

♂ He/Him 🌍 Manchester UK 🦈 Fin Fan

Consultant @ Steamhaus

Using AWS since 2019

Creator @ amazonwebshark.com

Runner; Keen Gardener; Dog Dad



# The 4 Vs Of Big Data

Characteristics of Big Data...

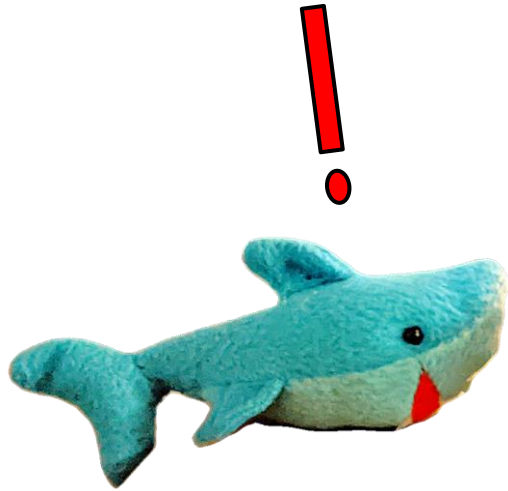
...and events...

...and API requests...

...metrics ...traces ...logs ...

# Variety

*“The state of being diverse or varied.”*



# Variety

*“The state of being diverse or varied.”*

Structure

Intent

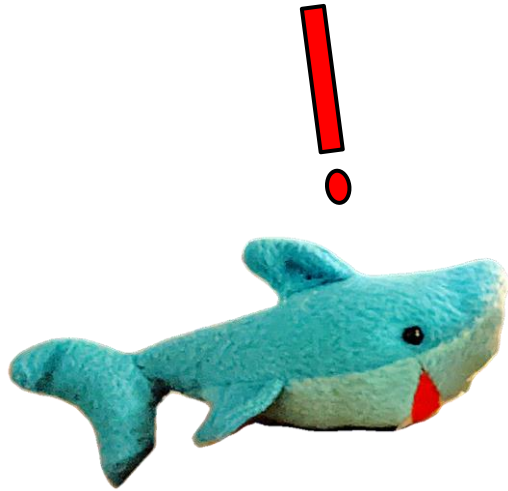
Sensitivity





# 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

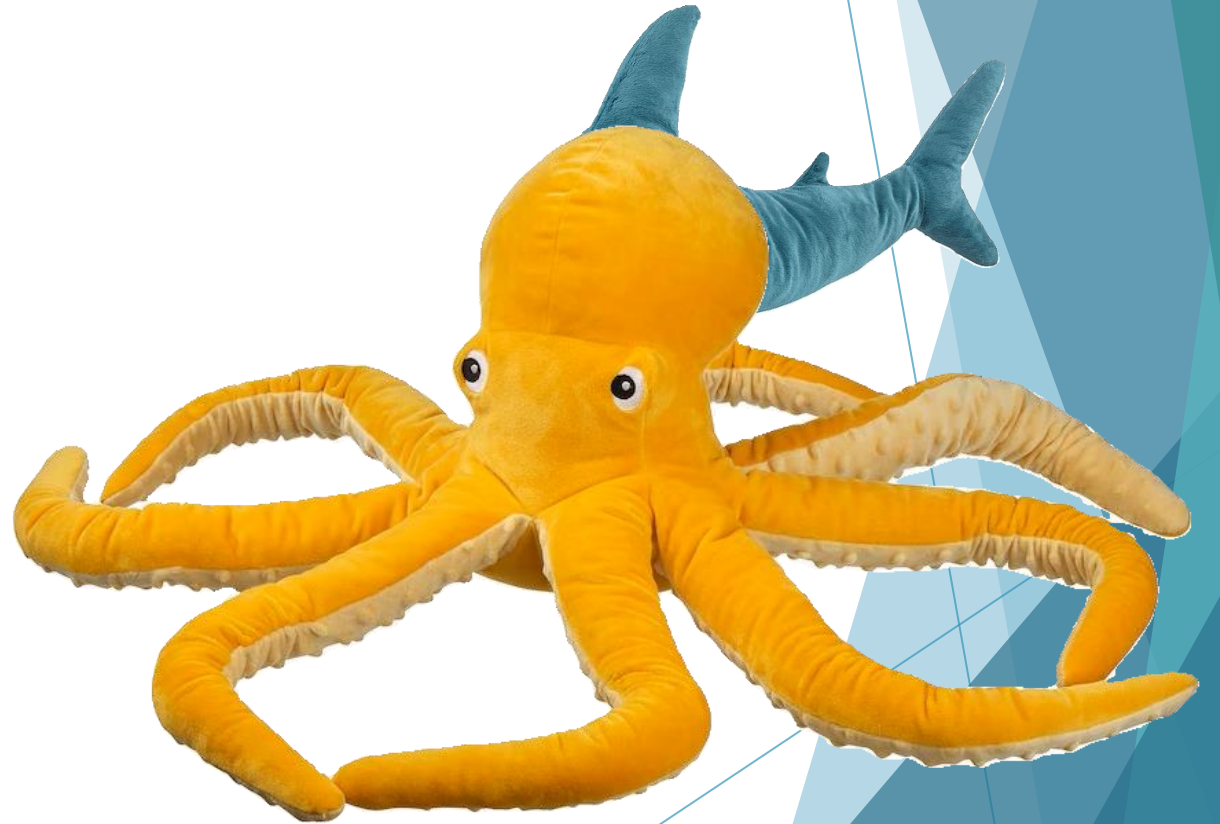
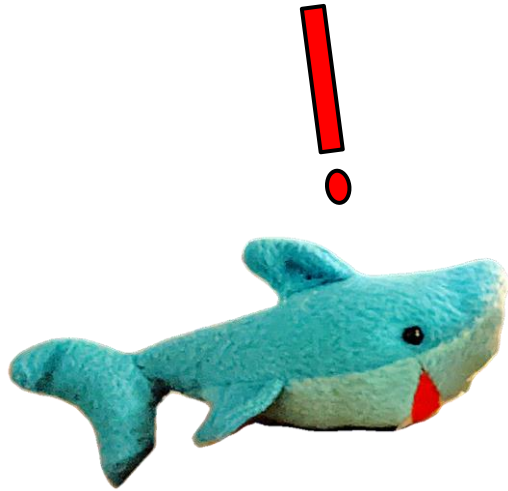
Synchronous or Asynchronous

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

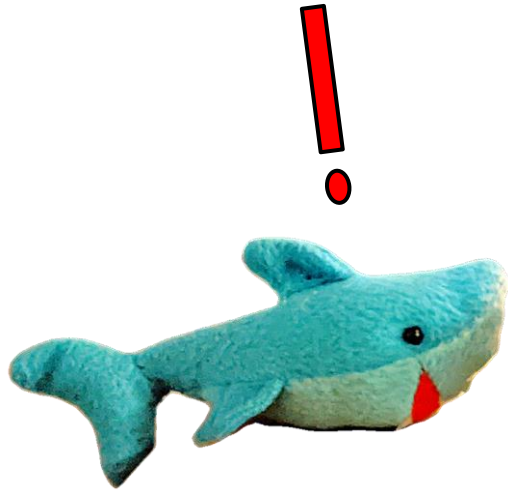
Validation & Health

Internal Security



# Volume

*“The amount of space occupied.”*



# Volume

*“The amount of space occupied.”*

Size or Amount

Storage Options

Backups





# Variety



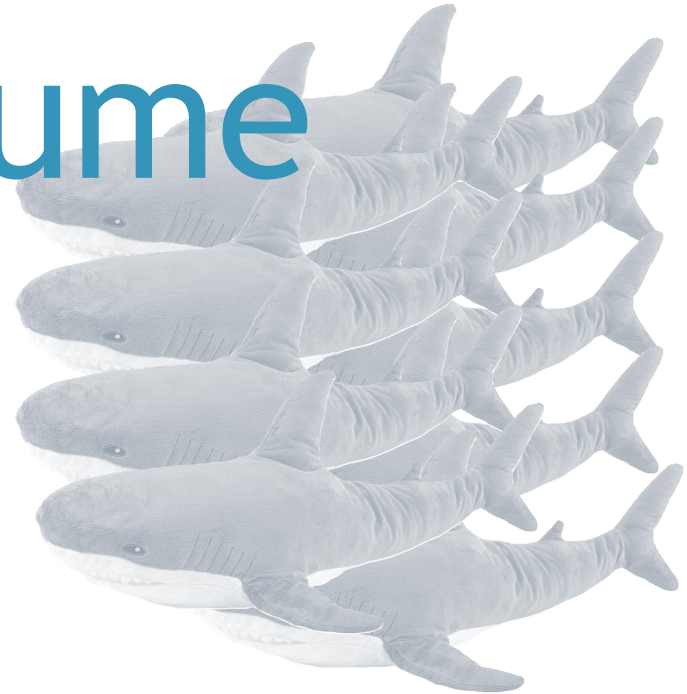
# Velocity



# Veracity

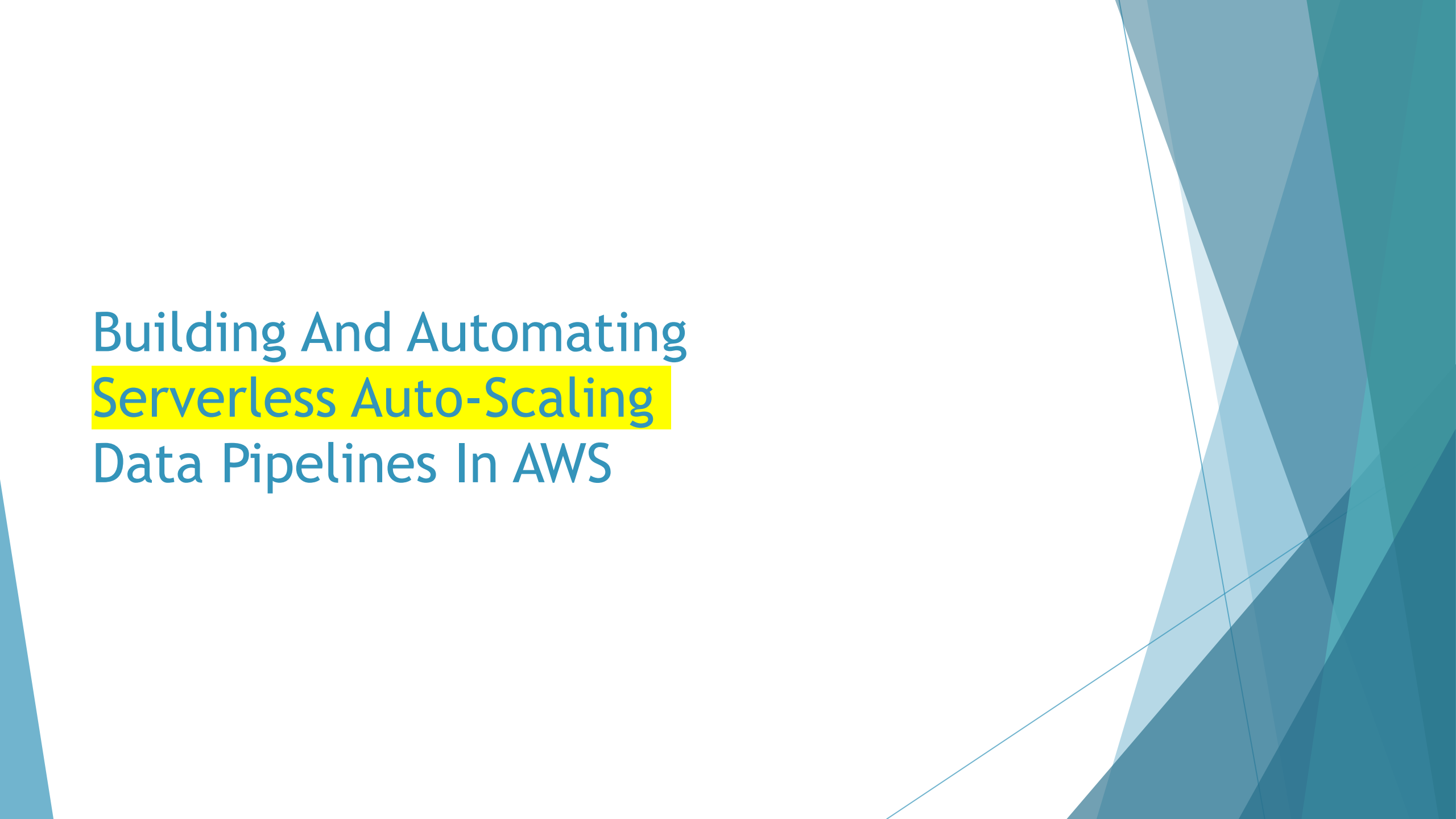


# Volume



# 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

[Code](#)[Test](#)[Monitor](#)[Configuration](#)[Aliases](#)[Versions](#)

## Code source [Info](#)

[Upload from](#) ▼

serverlesspresso-application-cor-GetQRcodeFunction-vOMa4YR8hdG5



EXPLORER



JS getCode.js X



JS getCode.js

JS localTest.js

() package.json

JS verifyCode.js



▼ DEPLOY

Deploy



▼ ENVIRONMENT VARIABLES

AWS\_NODEJS\_CONNECTION\_REUSE\_ENA...

BusName = Serverlesspresso

CodeLength = 10

ConfigTableName = serverlesspresso-config...

Source = awsserverlessda.serverlesspresso

TableArn = arn:aws:iam::123456789012:role/lambda-role

▼ TEST EVENTS

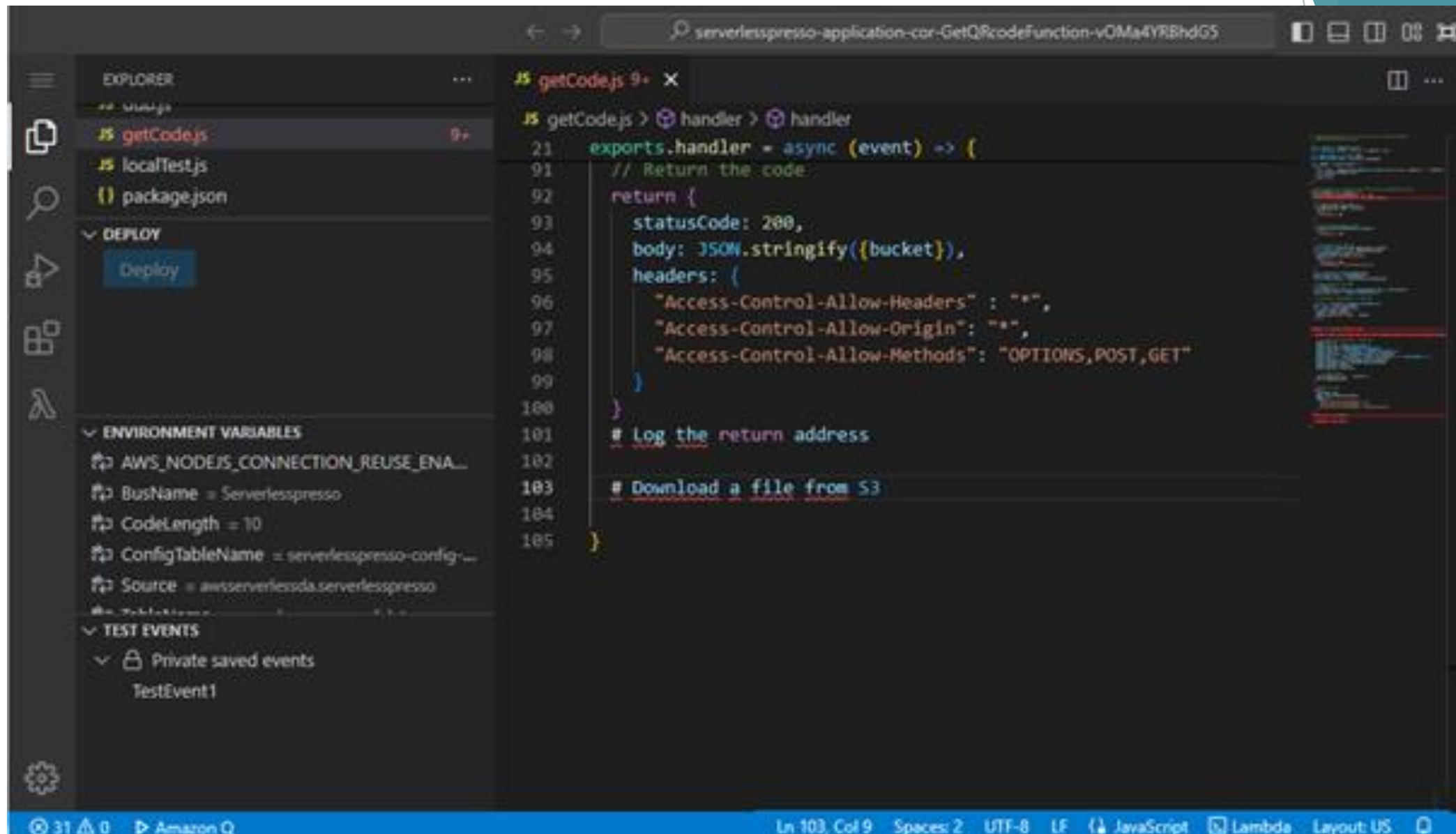
You haven't created any test events.

Create test event



```
JS getCode.js > ...
21 exports.handler = async (event) => {
83
84   // Return the code
85   return {
86     statusCode: 200,
87     body: JSON.stringify({bucket}),
88     headers: {
89       "Access-Control-Allow-Headers" : "*",
90       "Access-Control-Allow-Origin": "*",
91       "Access-Control-Allow-Methods": "OPTIONS,POST,GET"
92     }
93   }
94 }
```





## Create function [Info](#)

Choose one of the following options to create your function.

☒ **Author from scratch**  
Start with a simple Hello World example.

☐ **Use a blueprint**  
Build a Lambda application from sample code and configuration presets for common use cases.

☐ **Container image**  
Select a container image to deploy for your function.

### Basic information

#### Function name

Enter a name that describes the purpose of your function.

myFunctionName

Function name must be 1 to 64 characters, must be unique to the Region, and can't include spaces. Valid characters are a-z, A-Z, 0-9, hyphens (-), and underscores (\_).

#### Runtime [Info](#)

Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.

Python 3.13



#### Architecture [Info](#)

Choose the instruction set architecture you want for your function code.

☐ x86\_64

☒ arm64

# Amazon S3



Serverless object storage

Store anything for any reason

1000s of requests per second

Object protection & integrity checks

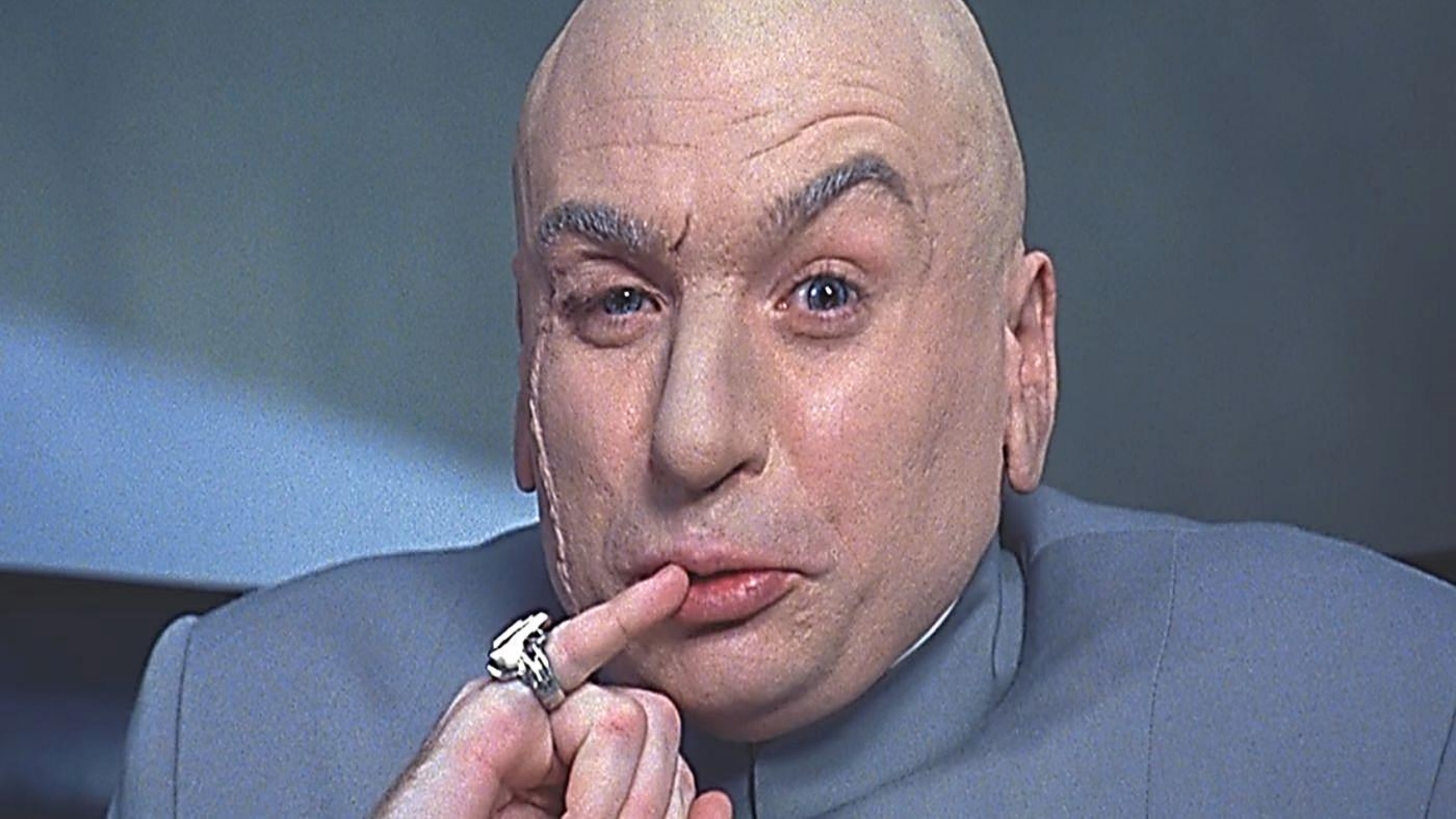


IT'S OVER

90000!!!!







# 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

Data Catalog indexes data assets

## Announcing generative AI upgrades for Apache Spark in AWS Glue (preview)

Posted on: Nov 22, 2024

## Announcing generative AI troubleshooting for Apache Spark in AWS Glue (Preview)

Posted on: Nov 22, 2024

# Amazon Athena



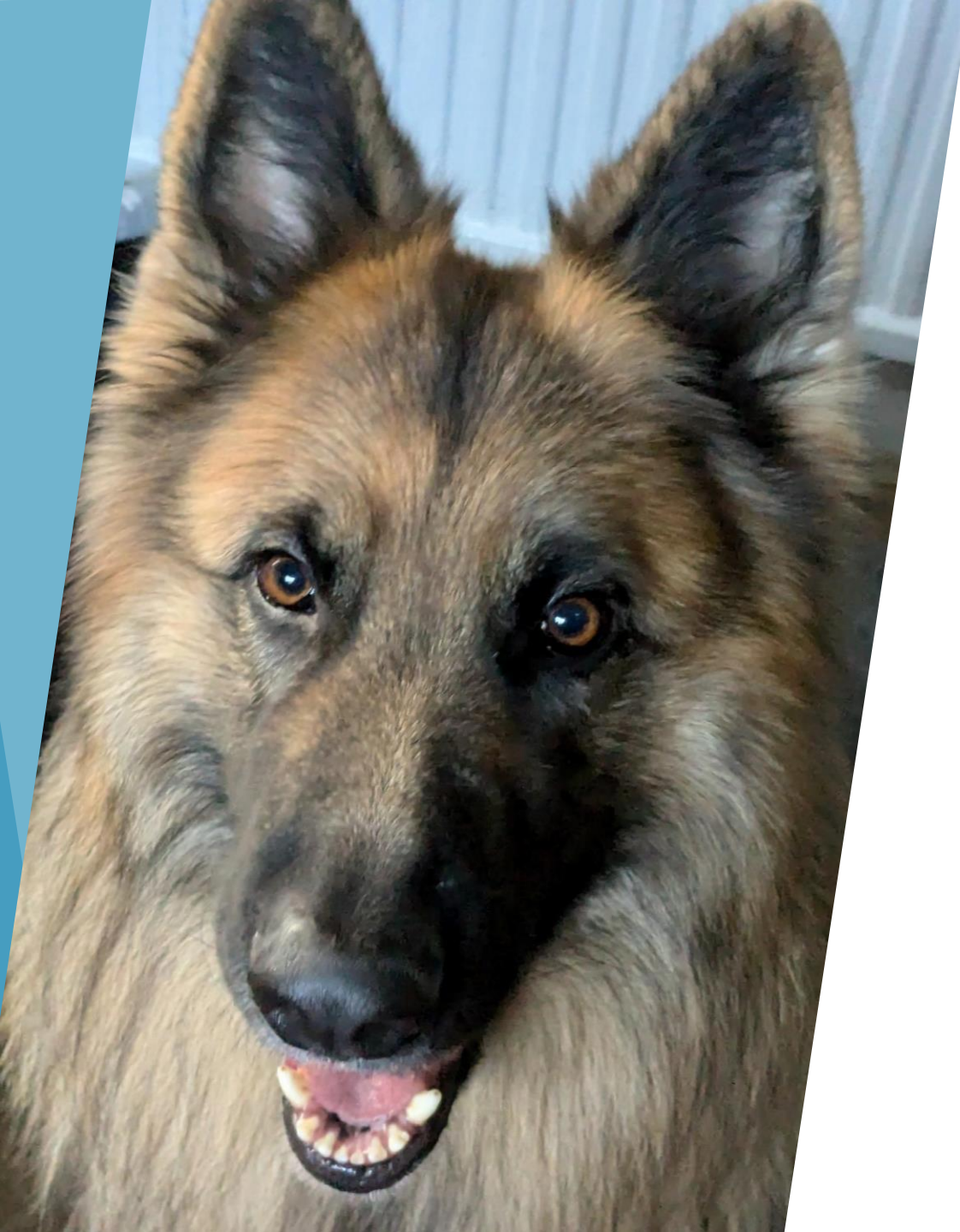
Serverless interactive query service

Query & access controls

Read-Only & Open Table support

Reads Variety Of Objects

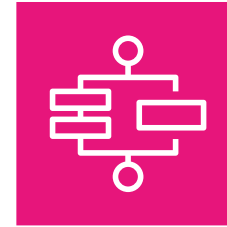




Wolfie Says  
Hi

# Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

# Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

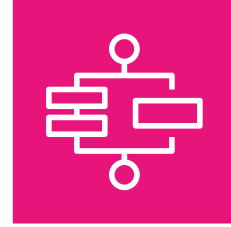


AWS Step Functions



Amazon EventBridge  
Scheduler

# AWS Step Functions



Serverless task orchestration

Invoke over 220 AWS services / 10k API calls

Standard & Express workflows

Design workflows visually and as code

## Export workflow to Infrastructure as Code (IaC) template ✕

Type [Learn more](#)

- ☒ **SAM**  
Export workflow to SAM template.
- ☐ **CloudFormation**  
Export workflow to CloudFormation template.

### ▼ Additional configurations

- ☒ Include IAM role created by console on your behalf.
- ☒ Replace resource references with DefinitionSubstitutions.
  - ☒ Include definitions for Activity resources.

[Cancel](#)

## Export workflow to Infrastructure as Code (IaC) template ✕

Type [Learn more](#)

- ☐ **SAM**  
Export workflow to SAM template.
- ☒ **CloudFormation**  
Export workflow to CloudFormation template.

### File format

- ☒ **JSON**
- ☐ **YAML**

### ▼ Additional configurations

- ☒ Include IAM role created by console on your behalf.
- ☒ Replace resource references with DefinitionSubstitutions.
  - ☒ Include definitions for Activity resources.

[Cancel](#)

[Download](#)

Search

Actions

Flow

Patterns

Info

MOST POPULAR

AWS Lambda

Invoke

Amazon SNS

Publish

Amazon ECS

RunTask

AWS Step Functions

StartExecution

AWS Glue

StartJobRun

THIRD-PARTY API

HTTP Endpoint

Call third-party API

```
graph TD; Start([Start]) --> LambdaRaw[Lambda: Invoke data_wordpressapi_raw]; LambdaRaw --> LambdaBronze[Lambda: Invoke data_wordpressapi_bronze]; LambdaBronze --> GlueCrawlerBronze[Glue: StartCrawler Start Bronze Crawler]; GlueCrawlerBronze --> GlueJobRun[Glue: StartJobRun Glue Start Silver Python Shell]; GlueJobRun --> Parallel[Parallel state Parallel]; Parallel --> RulesetEval[RulesetEvaluationRun ics_pages]; Parallel --> GlueDQRun[Glue: StartDataQualityRulesetEvaluationRun DQ Run: silver-posts]; RulesetEval --> GlueCrawlerSilver[Glue: StartCrawler Start Silver Crawler]; GlueDQRun --> SNSPublish[SNS: Publish PublishFailure]; GlueCrawlerSilver --> SNSPublish; SNSPublish --> End([End]);
```

The diagram illustrates an AWS Step Functions workflow for processing WordPress data. It begins with a 'Start' event, followed by two Lambda functions: 'data\_wordpressapi\_raw' and 'data\_wordpressapi\_bronze'. The workflow then uses AWS Glue to start a crawler for the bronze data and a job run for a silver Python shell. A parallel state follows, where one path evaluates rulesets for 'ics\_pages' and starts a silver crawler, while the other path starts a data quality evaluation run for 'silver-posts'. Both paths eventually lead to an SNS publish action, which handles both successful completion and failures. The workflow concludes with an 'End' event.

Workflow

Definition

The top level Amazon States Language properties for this workflow. [Learn more](#)

Start at

The state that is the starting point of the workflow.

data\_wordpressapi\_raw

Comment - optional

A human-readable description of the state machine.

Runs Lambda functions to ingest WordPress API data and transform to Parquet.

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

↶ Undo

↷ Redo

🔍 Zoom in

🔍 Zoom out

🔍 Search

<

Actions


Flow


Patterns


Info


MOST POPULAR


- ||

AWS Lambda  
Invoke
- ||

Amazon SNS  
Publish
- ||

Amazon ECS  
RunTask
- ||

AWS Step Functions  
StartExecution
- ||

AWS Glue  
StartJobRun

THIRD-PARTY API

- ||

HTTP Endpoint  
Call third-party API





Design

Code

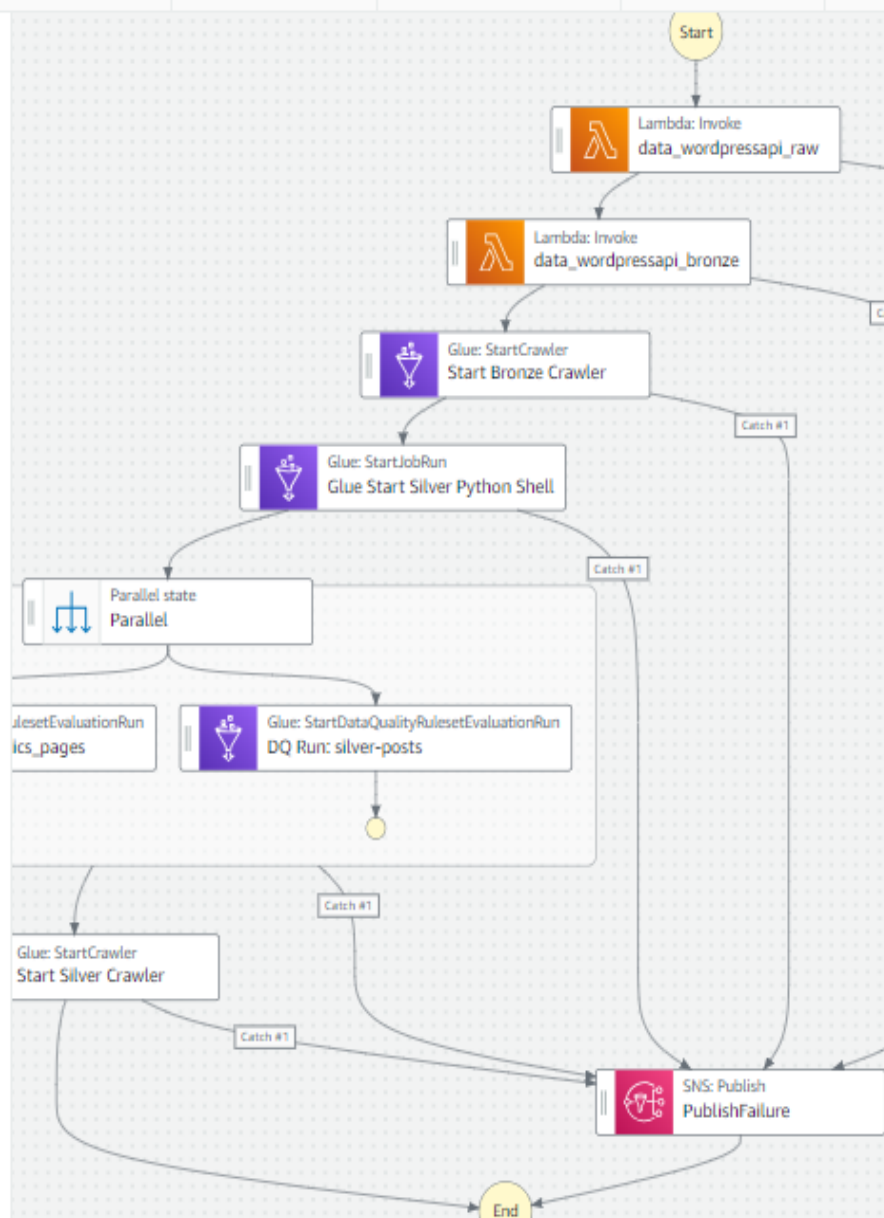
Config

Zoom out

Center

Duplicate

Delete



Exit

Actions ▼

Execute 


Save



 Feedback

## Workflow

☒ Definition >

The top level Amazon States Language properties for this workflow. [Learn more](#) 

### Start at

The state that is the starting point of the workflow.

data\_wordpressapi\_raw ▼

### Comment - optional

A human-readable description of the state machine.

Runs Lambda functions to ingest WordPress API data and transform to Parquet.

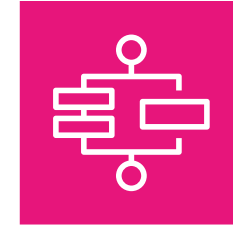
### TimeoutSeconds - optional

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



# AWS Step Functions Demo



Lambda Function: API Call

Glue Job: ETL

Athena Query: MSCK REPAIR TABLE

Step Functions | eu-west-1

eu-west-1.console.aws.amazon.com/states/home?region=eu-west-1#/v2/statemachines/...

aws Services Search [Alt+S]

00:00

MyStateMachine-xlcob549t

Design Code Config

Workflow not created Cancel Actions Create

Undo Redo Zoom in Zoom out Center Duplicate Delete Feedback

Search

Actions Flow Patterns Info

MOST POPULAR

AWS Lambda Invoke

Amazon SNS Publish

Amazon ECS RunTask

AWS Step Functions StartExecution

AWS Glue StartJobRun

THIRD-PARTY API

Start

Drag first state here

End

Workflow

Definition

The top level Amazon States Language properties for this workflow. [Learn more](#)

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

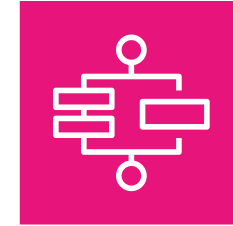
CloudShell Feedback

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Today - Todoist - G... Step Functions | eu... \*Untitled - Notepad OBS 30.2.3 - Profile...

16:53 03/09/2024

# AWS Step Functions Alternatives



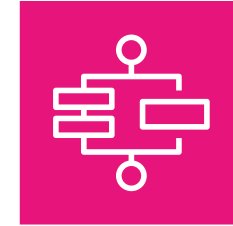
Lambda function:



+ Cheaper

- Less observability

# AWS Step Functions Alternatives



Glue Workflow:

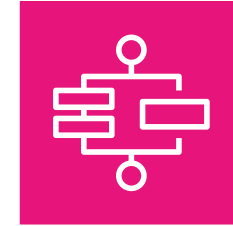


+ Free!

- Glue resources only



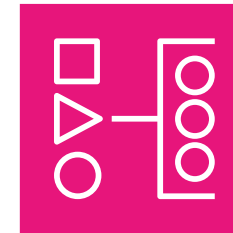
# AWS Step Functions Alternatives



Managed Airflow:

+ Customisable

- Complexity



# Amazon EventBridge Scheduler



Automate recurring & one-off tasks

Invoke over 220 AWS services

Set times or fixed-rate schedules

Checks target response

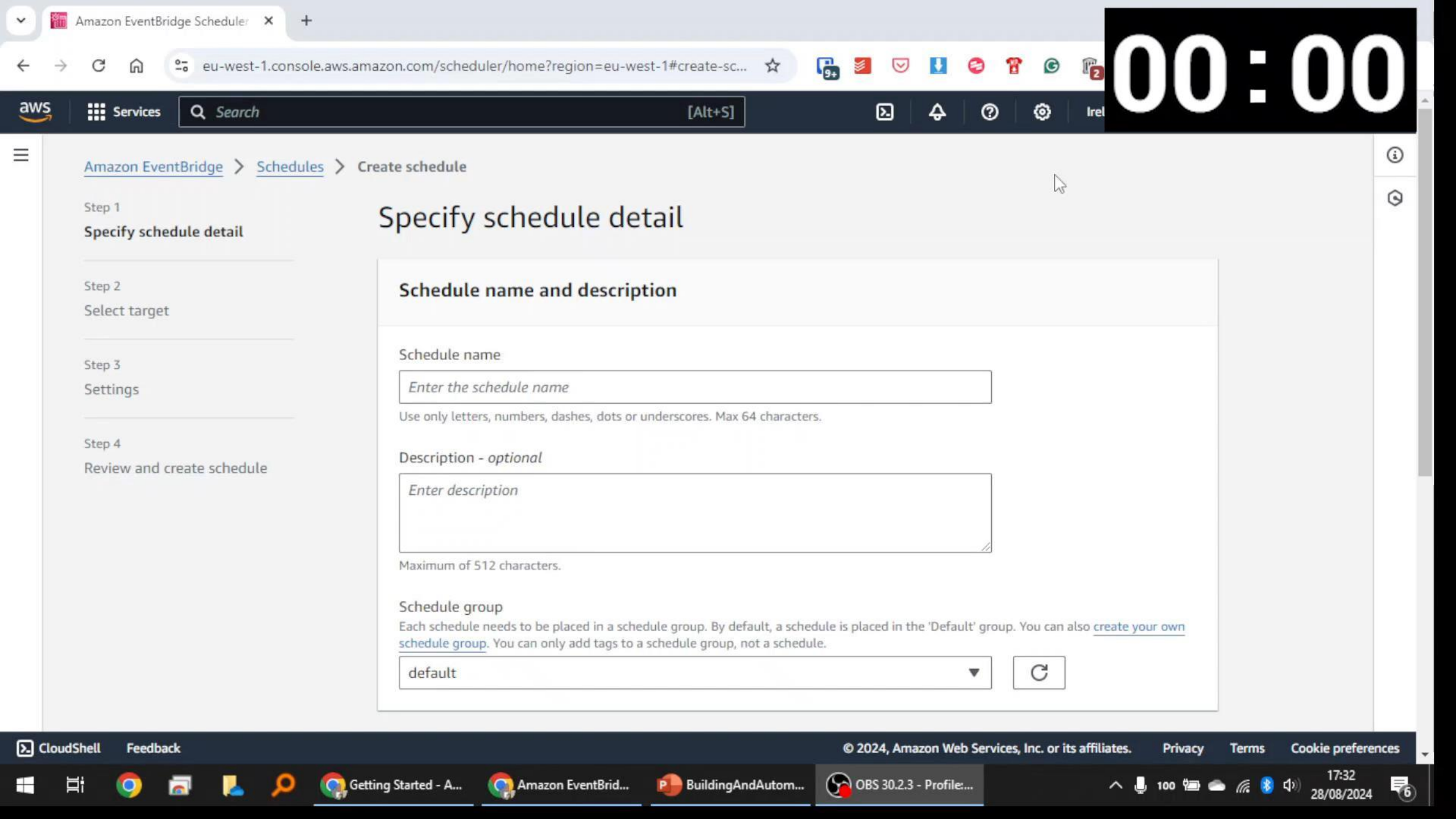
# Amazon EventBridge Scheduler Demo



Set schedule

Link Step Function workflow

Set configuration



00:00

Amazon EventBridge > Schedules > Create schedule

Step 1  
Specify schedule detail

Step 2  
Select target

Step 3  
Settings

Step 4  
Review and create schedule

## Specify schedule detail

### Schedule name and description

Schedule name

Enter the schedule name

Use only letters, numbers, dashes, dots or underscores. Max 64 characters.

Description - optional

Enter description

Maximum of 512 characters.

Schedule group

Each schedule needs to be placed in a schedule group. By default, a schedule is placed in the 'Default' group. You can also [create your own schedule group](#). You can only add tags to a schedule group, not a schedule.

default



# Summary

Problem Definition

Solution Architecture

Demos

Summary & Questions



[github.com/MrDamienJones  
/Community-Sessions](https://github.com/MrDamienJones/Community-Sessions)

# Thanks!



MrDamienJones



MrDamienJones



amazonwebshark.com



damien@amazonwebshark.com



@amazonwebshark



@amazonwebshark



github.com/MrDamienJones  
/Community-Sessions

