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

Damien Jones (he/him)

AWS Consultant @ Steamhaus

2024-09-26 AWS Community Summit Manchester

MrDamienJones

MrDamienJones

amazonwebshark.com

damien@amazonwebshark.com

@amazonwebshark

@amazonwebshark

https://www.flaticon.com/

Here For
Data & Analytics?

Here For
Development & Operations?

Here For
The Free Stuff?

Damien Jones

♂ He/Him 🌍 Manchester UK 🦈 Fin Fan

Consultant @ Steamhaus

Using AWS since 2019

Creator @ amazonwebshark.com

Runner; Keen Gardener; Dog Dad

aws certified
Data Engineer
ASSOCIATE

aws certified
SysOps Administrator
ASSOCIATE

aws certified
Cloud Practitioner
FOUNDATIONAL

aws certified
Developer
ASSOCIATE

Agenda

Problem Definition

Solution Architecture

Demos

Summary & Questions

github.com/MrDamienJones
/Community-Sessions

The 4 Vs Of Big Data

Characteristics of Big Data…

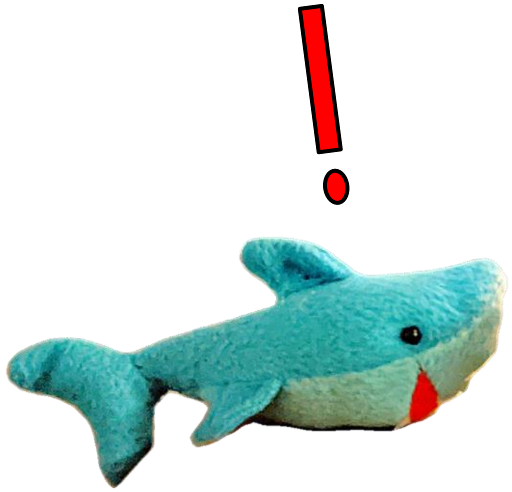
…and events…

…and API requests…

…metrics …traces …logs …

Variety

“The state of being diverse or varied.”



Variety

“The state of being diverse or varied.”

Structure

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."

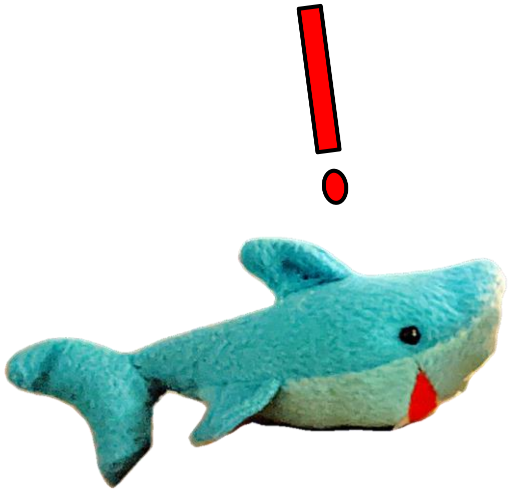
External Security

Validation & Health

Internal Security

Volume

“The amount of space occupied.”



Volume

“The amount of space occupied.”

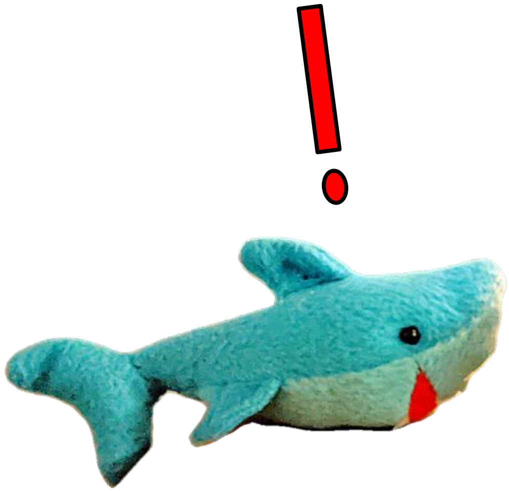
Size or Amount

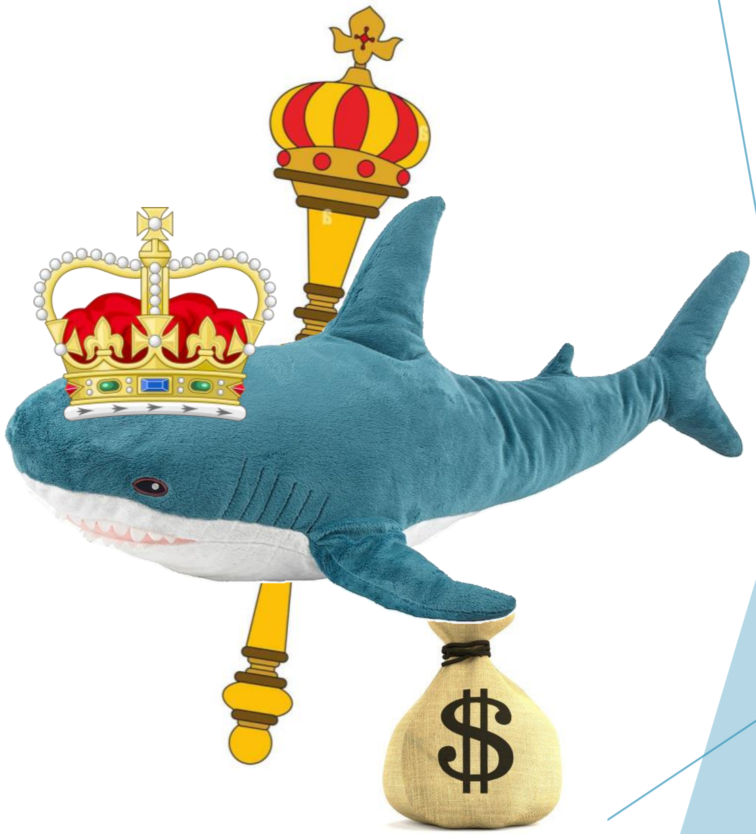
Access Patterns

Backups

Value

“The importance, worth or usefulness.”





Value

"The importance, worth or usefulness."

Return On Investment

Reconciliation

Liquidity

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

Amazon S3

Auditing with Inventories

Map value with Lifecycles

Monetise datasets with Data Exchange

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Building And Automating Serverless Auto-Scaling Data Pipelines In AWS

Amazon Athena

AWS Glue

Amazon Athena

Serverless interactive query service

Query & access controls

Read-Only & Open Table support

Create derived tables

AWS Glue

Fully managed serverless ETL service

Crawlers discover data automatically

Data Catalog indexes data assets

Up to 2000 concurrent ETL job runs

AWS Glue

Add value with ETL jobs

Add efficiencies with ETL jobs

Prevent downtime with Data Quality checks

Prove value with Data Quality scores

AWS Glue Data Quality Demo

Summary

Data Quality Chart

Recent Run

Run History

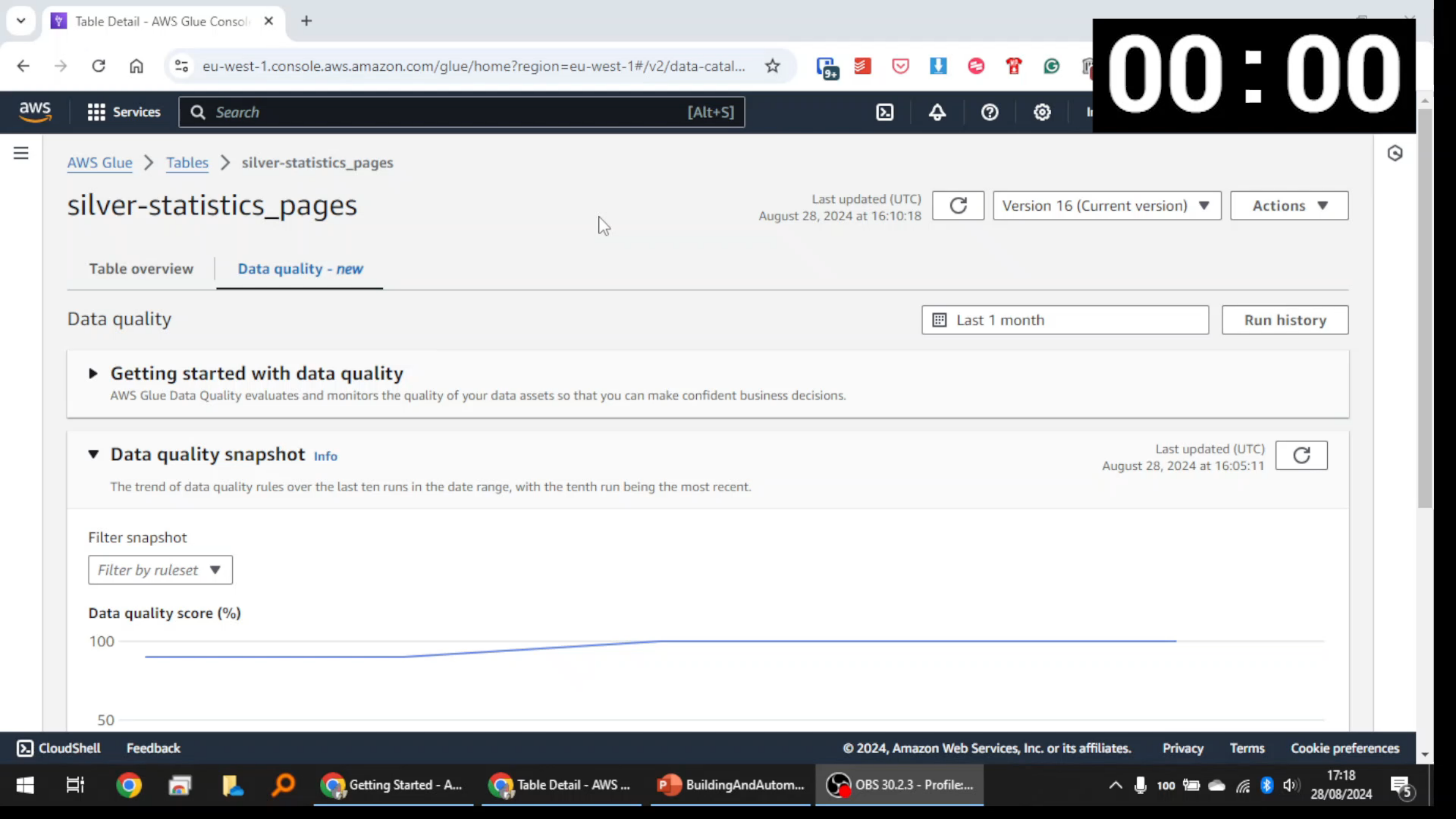


Table Detail - AWS Glue Console

eu-west-1.console.aws.amazon.com/glue/home?region=eu-west-1#/v2/data-catal...

00:00

AWS Glue > Tables > silver-statistics_pages

silver-statistics_pages

Last updated (UTC)
August 28, 2024 at 16:10:18

Version 16 (Current version) ▼

Actions ▼

Table overview | Data quality - *new*

Data quality

Last 1 month

Run history

▶ **Getting started with data quality**

AWS Glue Data Quality evaluates and monitors the quality of your data assets so that you can make confident business decisions.

▼ **Data quality snapshot** Info

Last updated (UTC)
August 28, 2024 at 16:05:11

The trend of data quality rules over the last ten runs in the date range, with the tenth run being the most recent.

Filter snapshot

Filter by ruleset ▼

Data quality score (%)

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

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

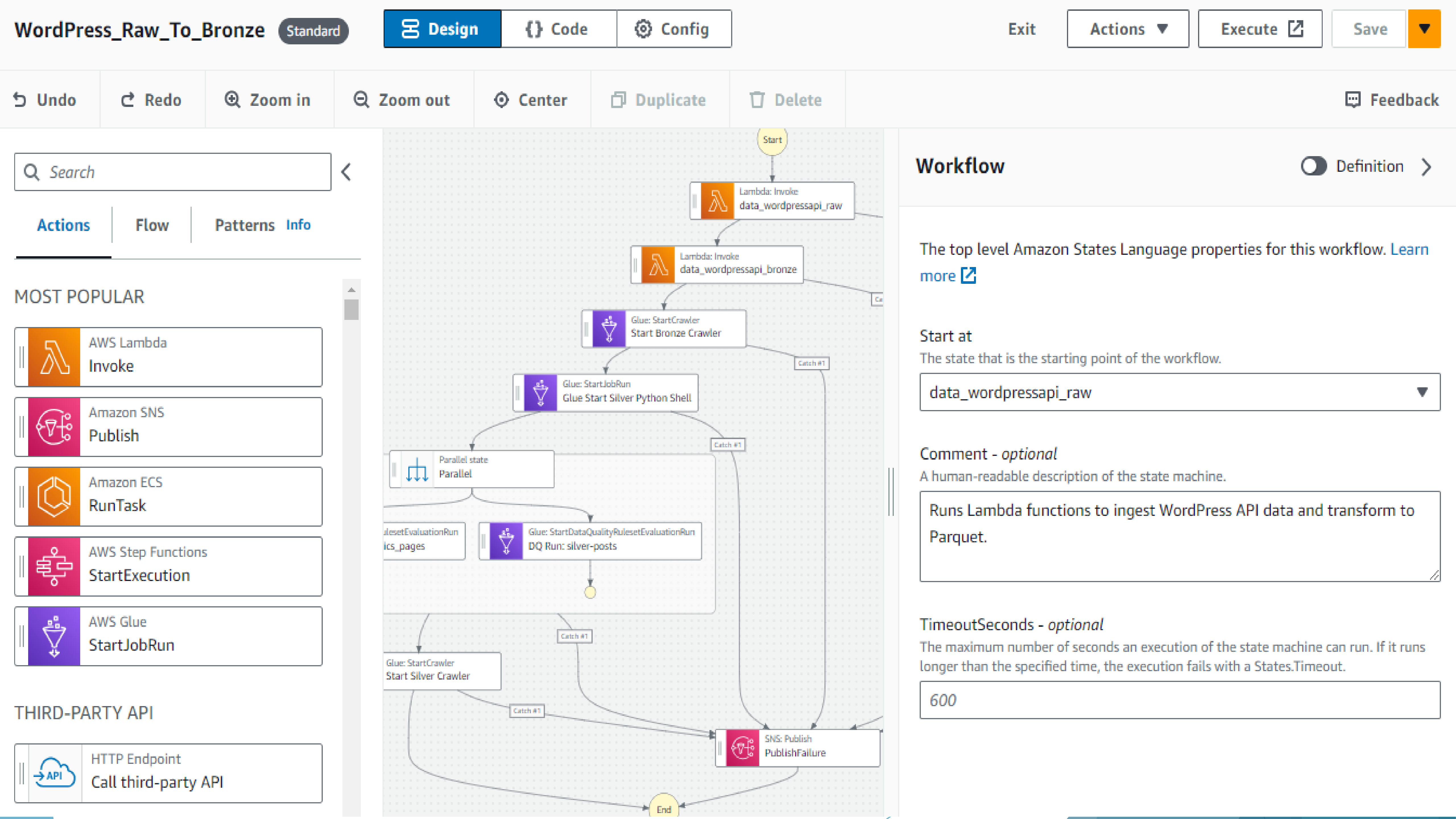
AWS Step Functions

Serverless task orchestration

Invoke over 220 AWS services / 10k API calls

Design workflows visually and as code

Standard & Express workflows



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*

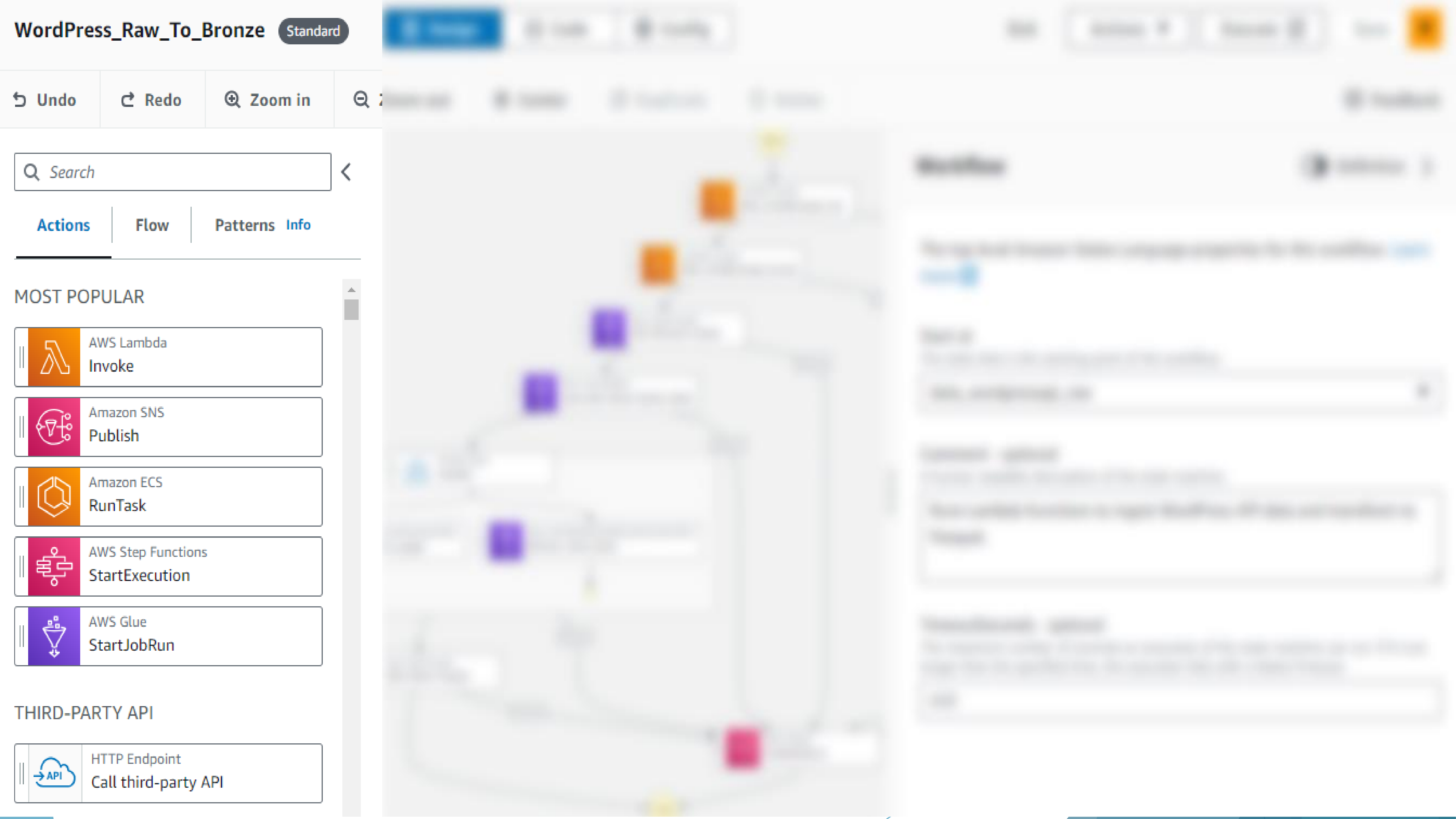
A human-readable description of the state machine.

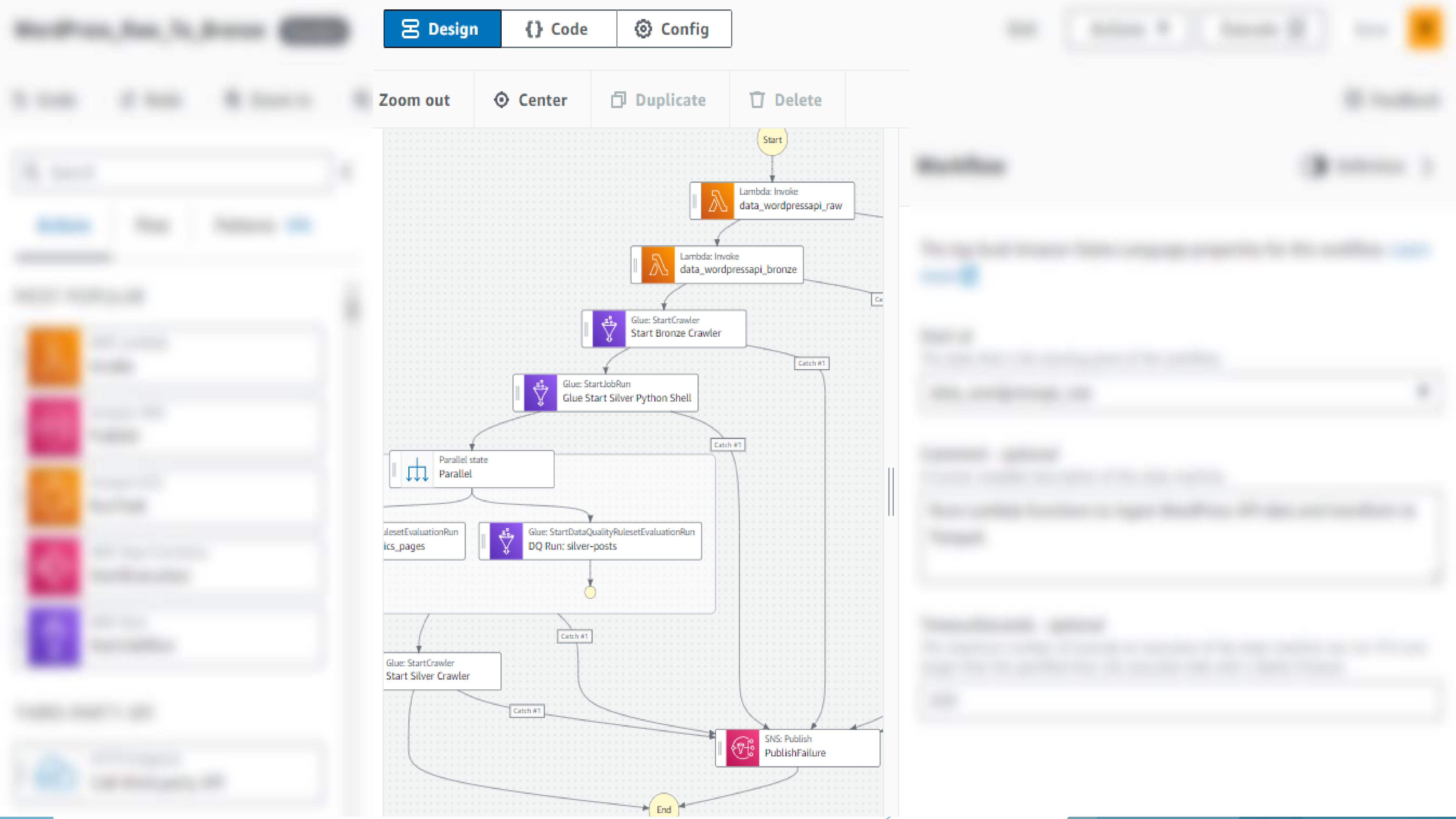
Runs Lambda functions to ingest WordPress API data and transform to Parquet.

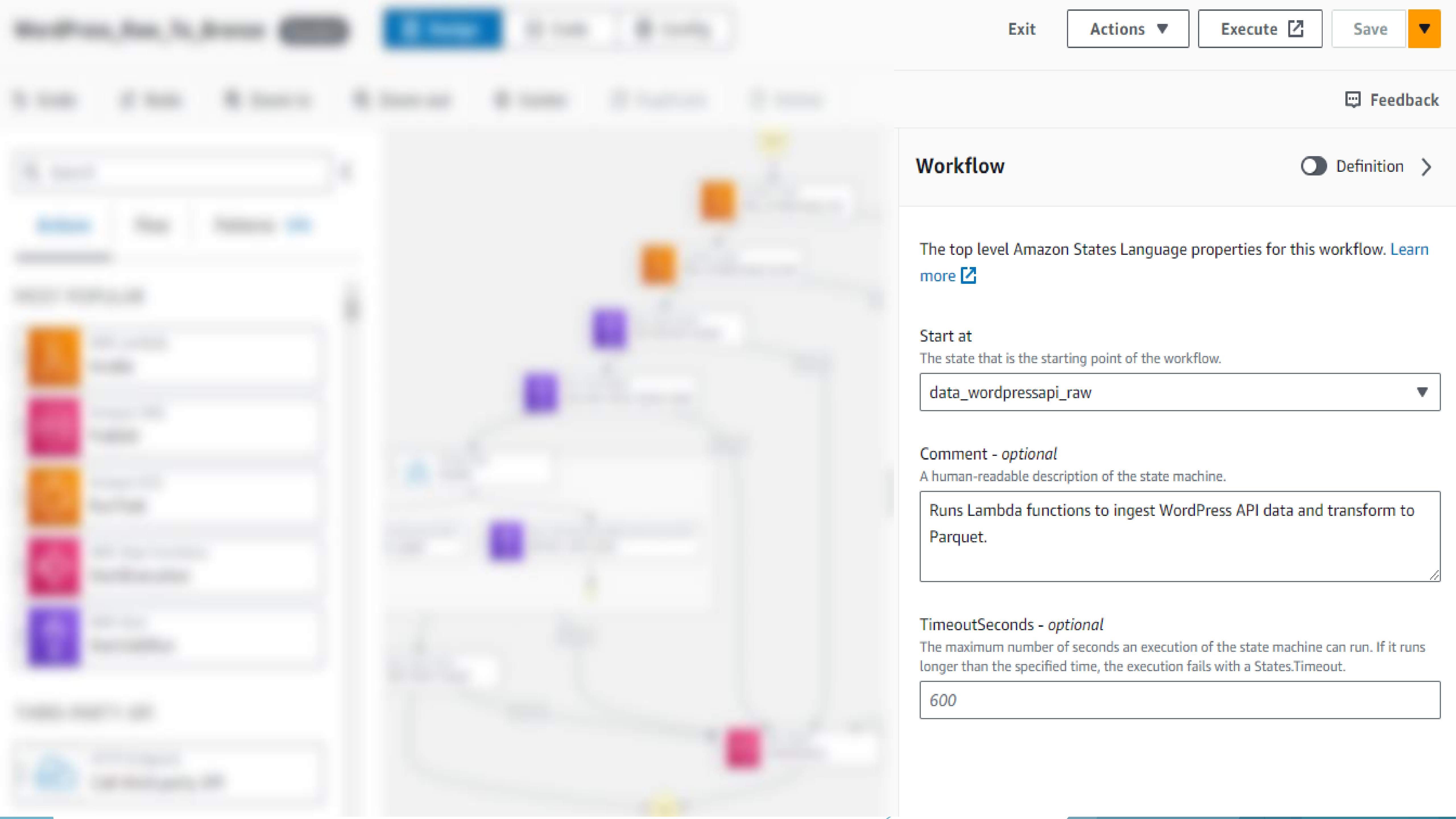
TimeoutSeconds - *optional*

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


WordPress_Raw_To_Bronze
Standard
Undo
Redo
Zoom in
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
API


Design
Code
Config
Zoom out
Center
Duplicate
Delete
Start
Lambda: Invoke
data_wordpressapi_raw
Lambda: Invoke
data_wordpressapi_bronze
Ca
Glue: StartCrawler
Start Bronze Crawler
Catch #1
Glue: StartJobRun
Glue Start Silver Python Shell
Catch #1
Parallel state
Parallel
lesetEvaluationRun
ics_pages
Glue: StartDataQualityRulesetEvaluationRun
DQ Run: silver-posts
Catch #1
Glue: StartCrawler
Start Silver Crawler
Catch #1
SNS: Publish
PublishFailure
End



Workflow

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.

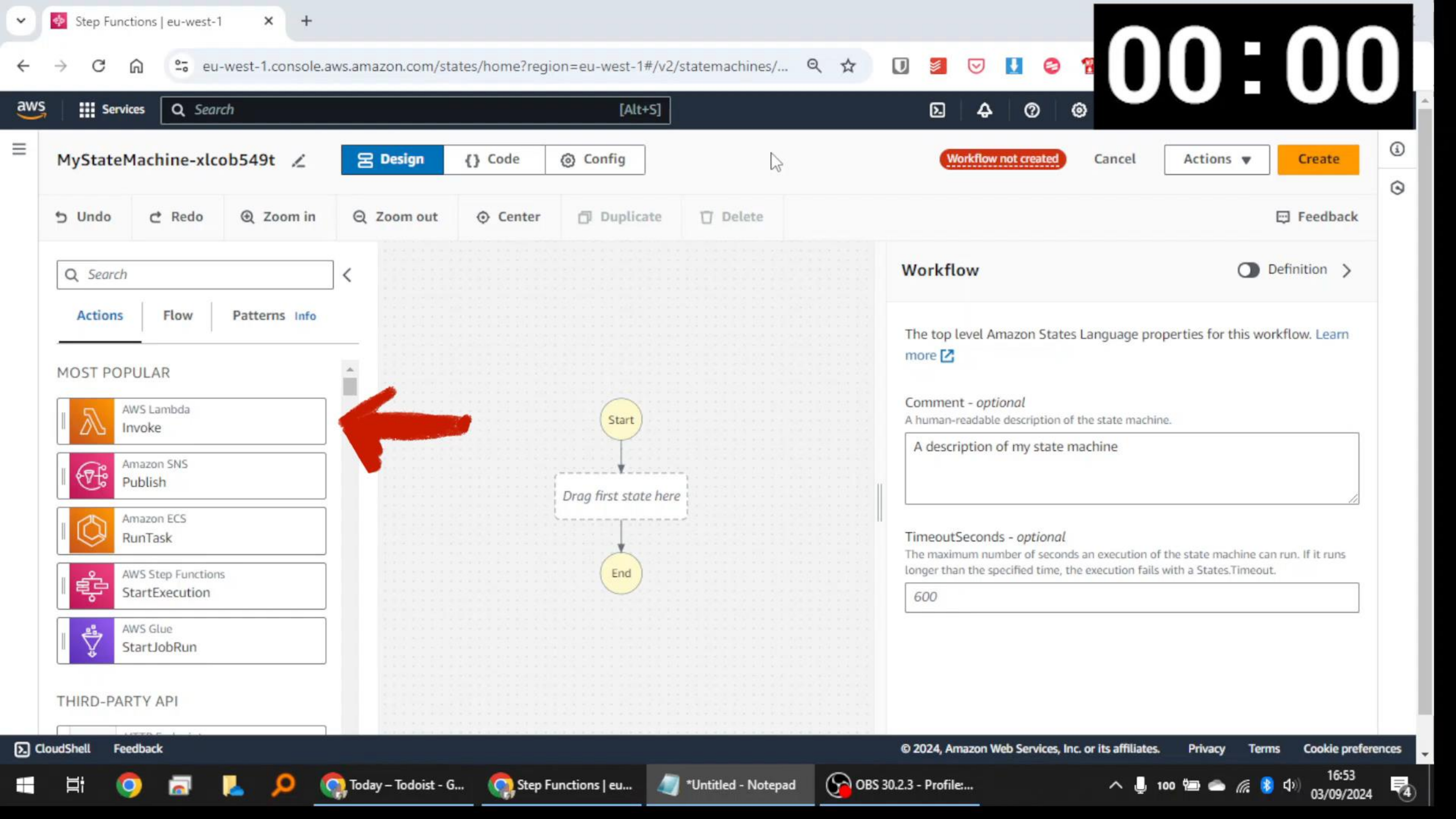
600

AWS Step Functions Demo

Lambda Function: API Call

Glue Job: ETL

Athena Query: MSCK REPAIR TABLE



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

Amazon EventBridge Scheduler

Automate recurring & one-off tasks

Invoke over 220 AWS services

Set times or fixed-rate schedules

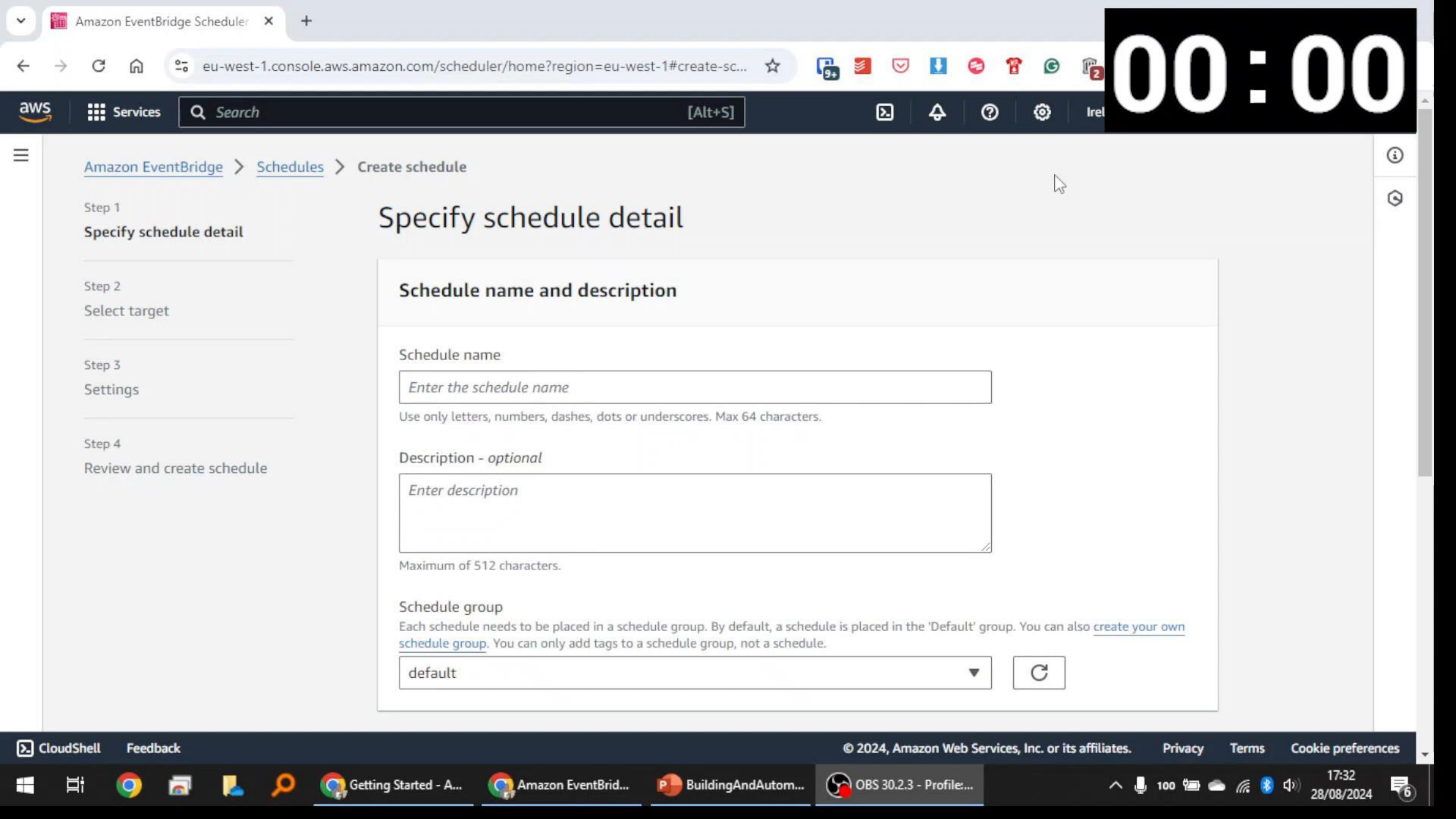
Checks target response

Amazon EventBridge Scheduler Demo

Set schedule

Link Step Function workflow

Set configuration



Amazon EventBridge Scheduler
eu-west-1.console.aws.amazon.com/scheduler/home?region=eu-west-1#create-sc...
00:00
Services
Search
[Alt+S]
Irel
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.

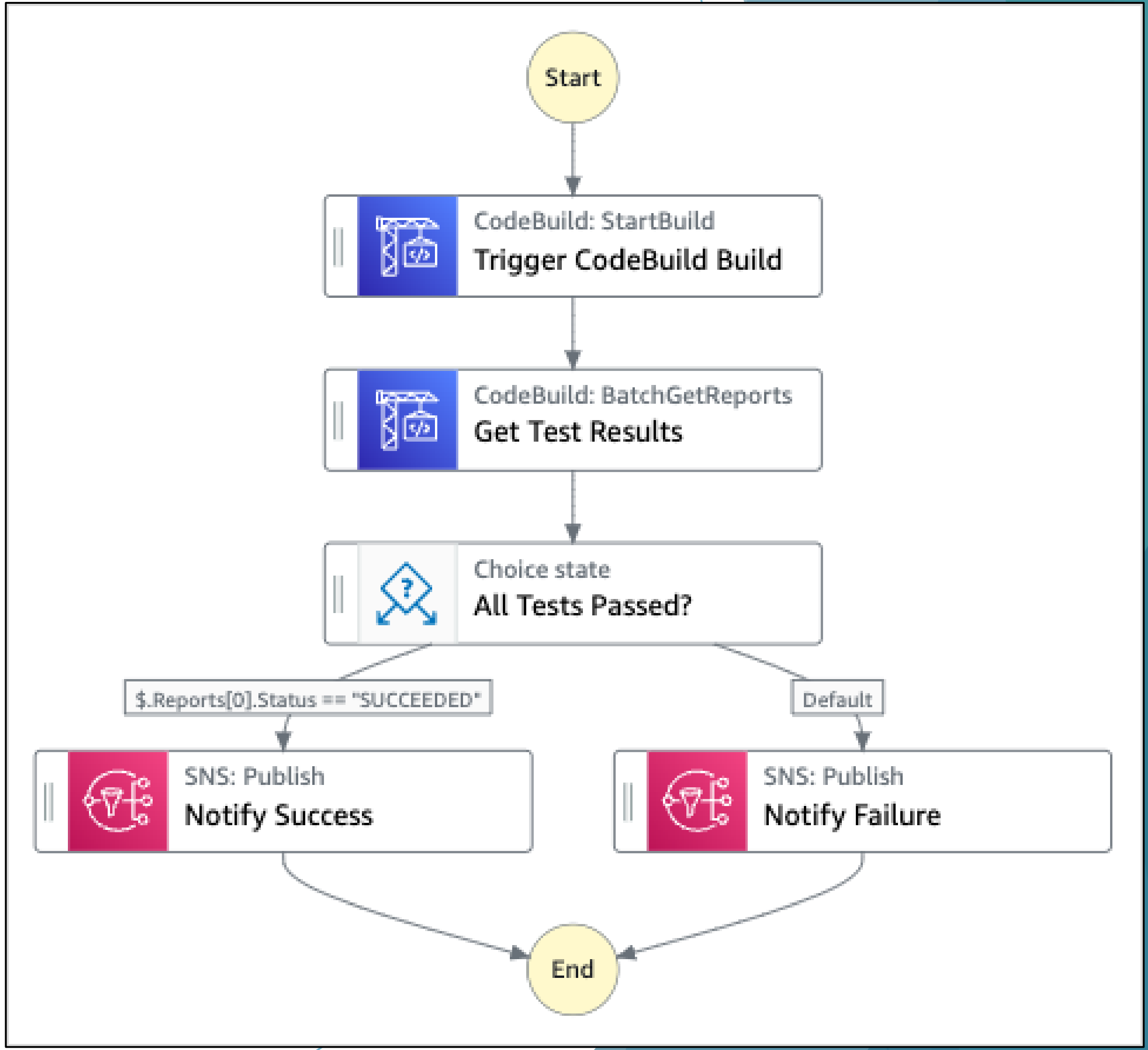
default

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

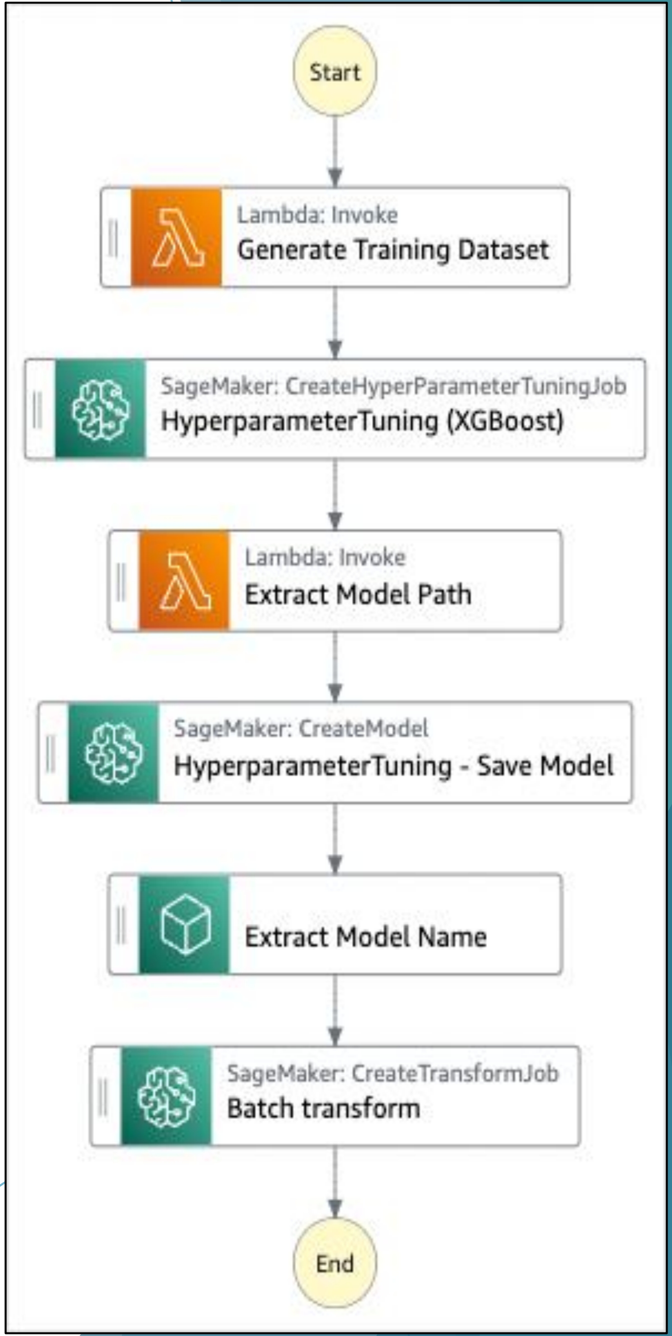
Tune hyperparameters of a machine learning model and batch transform a test dataset.

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

Demos

Summary & Questions

github.com/MrDamienJones
/Community-Sessions

Thanks!

MrDamienJones

MrDamienJones

amazonwebshark.com

damien@amazonwebshark.com

@amazonwebshark

@amazonwebshark

github.com/MrDamienJones
/Community-Sessions