Airflow 2.0 on Amazon MWAA

Airflow Summit 2021

Sam Dengler and John Jackson
July 15th, 2021
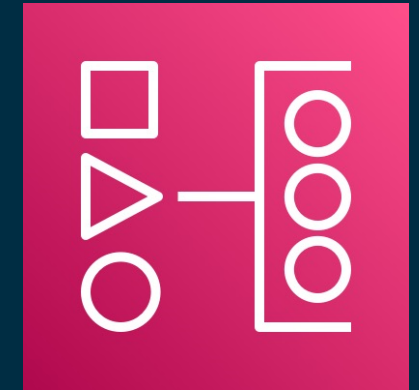
Table of contents

- How Amazon MWAA does Apache Airflow
- Staying current 2.x and beyond
    - Upgrades and new Apache Airflow version support
- Staying Open and contributing
    - AWS Contributions to Date
    - Support Plan for the Open Source Community

How Amazon MWAA does Apache Airflow

aws

Amazon Managed Workflows for Apache Airflow (MWAA)

- A managed service for Apache Airflow that makes it easy for data engineers and data scientists to execute data processing workflows on AWS

- Released November 24, 2020, added Airflow 2.0 support May 26, 2021
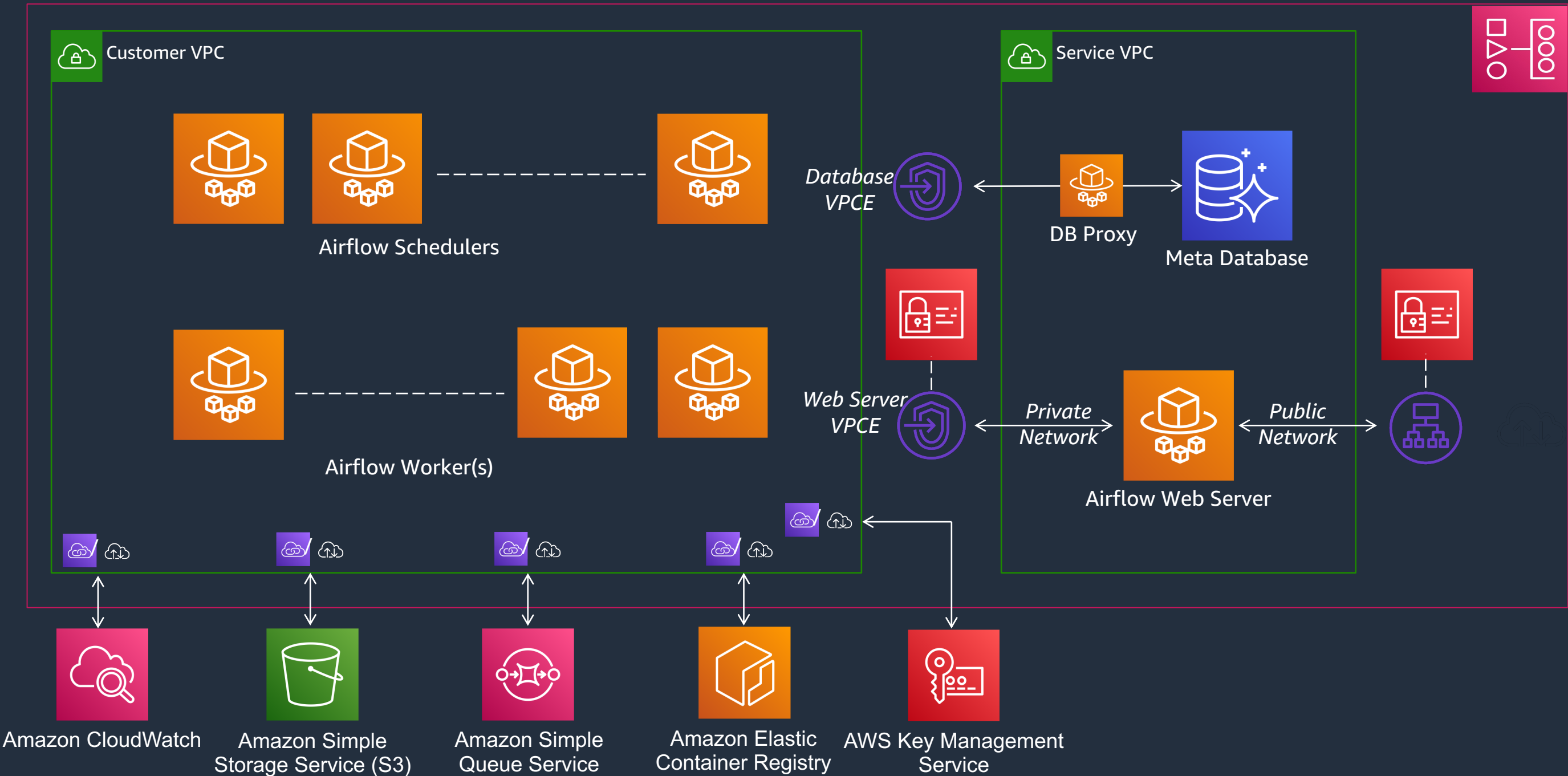
	Name		Status		Created date		Airflow version		Airflow UI	
○	MWAA-Demo-5		⊘ Available		May 25, 2021 17:26:01 (UTC-07:00)		2.0.2		Open Airflow UI ↗	
○	MWAA-Demo-4		⊘ Available		May 25, 2021 14:03:12 (UTC-07:00)		2.0.2		Open Airflow UI ↗	
○	MWAA-Demo-3		⊘ Available		May 24, 2021 18:03:09 (UTC-07:00)		2.0.2		Open Airflow UI ↗	
○	MWAA-Demo-2		⊙ Updating		Dec 14, 2020 08:28:22 (UTC-08:00)		1.10.12		Open Airflow UI ↗	
○	MWAA-Demo-1		⊘ Available		Dec 10, 2020 14:37:40 (UTC-08:00)		1.10.12		Open Airflow UI ↗	

Environments (5)

Edit   Delete   Actions ▼   **Create environment**

Find environments   ‹ 1 ›

aws

Amazon MWAA is Open Source Apache Airflow: No Forks

aws

Amazon MWAA Architecture

Customer VPC

Airflow Schedulers

Airflow Worker(s)

Service VPC

Database VPCE

DB Proxy

Meta Database

Web Server VPCE

Private Network

Public Network

Airflow Web Server

Amazon CloudWatch

Amazon Simple Storage Service (S3)

Amazon Simple Queue Service

Amazon Elastic Container Registry

AWS Key Management Service

© 2021, Amazon Web Services, Inc. or its Affiliates.

aws

How Amazon MWAA Helps

- Deployments and Operations
  - Easy to setup up and maintain
- Availability and Sizing
  - Multi AZ/HA with Airflow 2.0 on ECS Fargate
- Scaling
  - Auto scaling with Celery executor
- Security
  - IAM and VPC

Setup

Upgrades
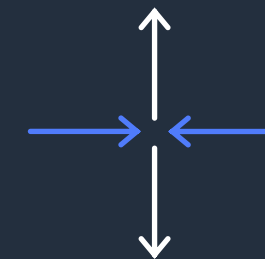
Scaling

Security

Maintenance

aws

Deployments and Operations

- Easy to set up multiple small environments
- Prod/Dev environments
- CloudWatch
  - Logging
  - Metrics/Alarms
  - Dashboards
- CloudFormation/Terraform
  - Configure User Access, Environment (Execution Role) access, and Airflow Configurations without accessing the UI

Scaling/Environment Sizes

- Using Airflow metrics
- (Tasks running+tasks queued)/tasks per container = number of containers required

- Downscaling occurs when (Tasks running+tasks queued)=0

Environment class **Info**

Class	Maximum worker count
mw1.small	10
Scheduler count	Minimum worker count
2	1

aws

Security with MWAA

- Auth (IAM/Federation)
- VPC/SG
- Secrets Manager
- Execution Role

aws

CI/CD

- Versioned S3 Bucket deployment target
- Flexible integration with existing CI/CD pipelines
- DAGs automatically synchronized every 30-60s
- Custom Plugin and Python dependency changes require Environment update

Create an MWAA Environment

Copy your DAGs and Plugins to S3

Access the Airflow UI

aws

Amazon MWAA Best Practices

Configuration and Migration

aws

Best Practices – Creating the Environment

- VPC Network
  - Connectivity to AWS Services
  - Private Web Server Option
- IAM Execution Role
  - Access to AWS Services
  - AOK vs CMK
- Python Requirements
  - Compatible Versions
  - Debugging
- Configuration Overrides
  - Performance Tuning
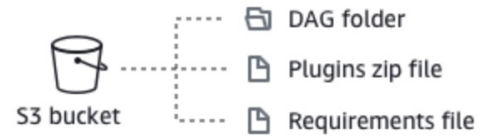 - Secrets Backend Support

Environment details Info

Name

MyAirflowEnvironment

Use only letters, numbers, dashes, or underscores. Max 80 characters.

Airflow version

2.0.2 (Latest) ▼

DAG code in Amazon S3 Info

Amazon MWAA uses your Amazon S3 bucket to load your DAGs and supporting files. Specify your S3 bucket, and the paths of your DAG folder, plugins.zip, and requirements.txt.

S3 bucket ---- 📁 DAG folder
        ---- 📄 Plugins zip file
        ---- 📄 Requirements file

ⓘ Create or specify an S3 bucket to store your DAG code. The bucket name must have versioning enabled. You can create a new bucket in the **Amazon S3 console** ⬈

S3 Bucket
The S3 bucket where your source code is stored. Enter an S3 URI or browse and select a bucket.

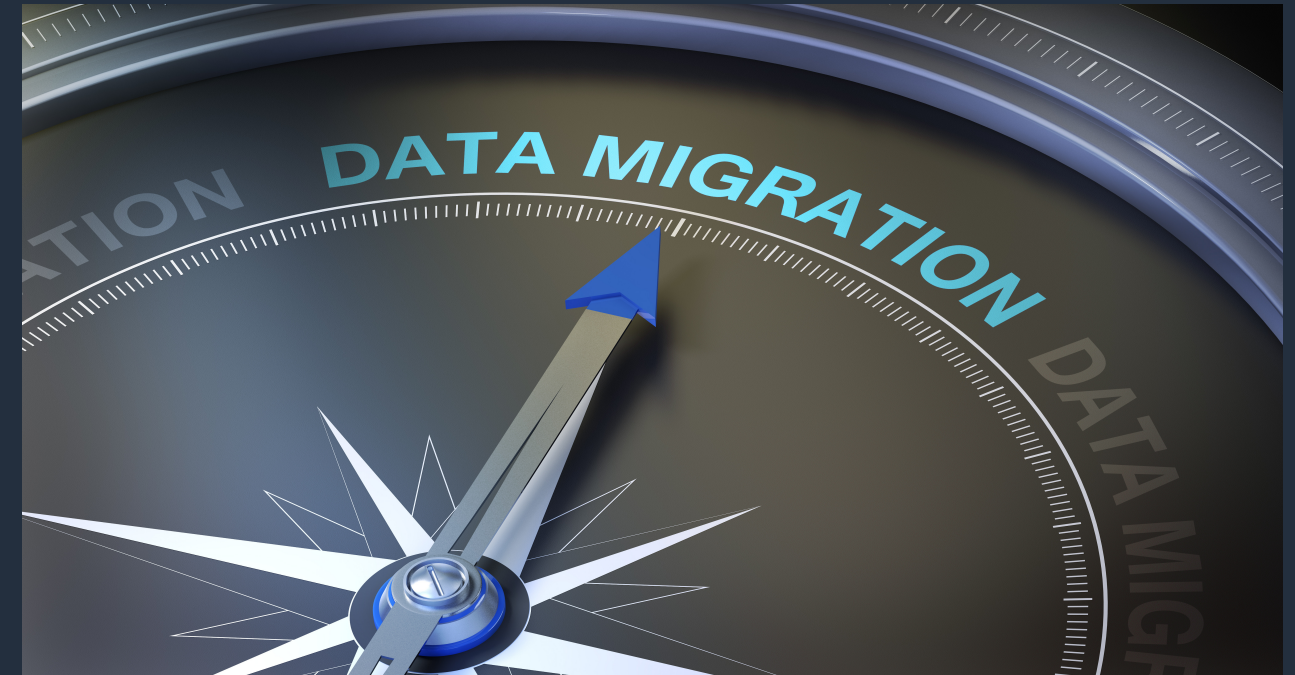🔍 s3://bucket   View ⬈   Browse S3

Format: s3://mybucketname

aws

Best Practices – Migration

- Moving tasks off the Worker that require
    - Local Docker
    - Custom runtimes
    - Large CPU/Memory
    - Sudo
- Assume Ephemeral Workers
    - Limited Bash Value
    - Local storage between tasks
- Web Server Restrictions
    - Not currently installing requirements.txt or plugins

aws

Contributions and Staying Current

Staying Current

- Target is to provide minor versions on Amazon MWAA within 30 days of Apache Airflow community release
    - We're not there yet, but we're working on it
- Major releases will take a bit longer
- We'll keep older versions (i.e. 1.10.x) available for as long as we can
    - Depends on security, stability, and customer demand
- Version status will be posted at https://docs.aws.amazon.com/mwaa/latest/userguide/airflow-versions.html

aws

Open Source Contributions

- EKS Native operator
- Improved Apache Airflow AWS integrations
- Performance and security improvements
- Native Serverless executor

Resources

- Amazon MWAA docs https://docs.aws.amazon.com/mwaa
- Amazon MWAA product page https://aws.amazon.com/mwaa
- #airflow-aws Slack Channel: https://apache-airflow.slack.com
- GitHub samples: https://github.com/aws-samples/amazon-mwaa-examples
- Local Runner https://github.com/aws/aws-mwaa-local-runner

aws

Q&A

Sam Dengler *@samdengler*

John Jackson *@JohnJacksonPM*

aws