INNOVATE

DATA EDITION

23 August, 2022

Deliver insights faster with AWS Glue interactive sessions and continuous delivery

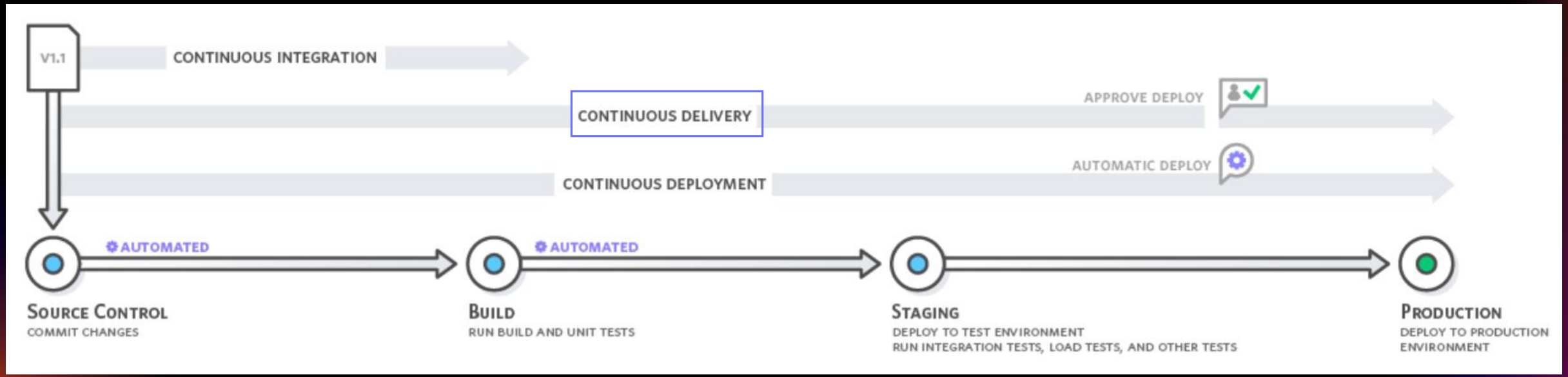
Niladri Bhattacharya
Senior Analytics Specialist Solutions Architect
Amazon Web Services

Agenda

- What is continuous delivery
- Continuous delivery to deliver data products
- Key steps in implementing CD for data products
- AWS Glue to deliver data products
- Introducing AWS Glue interactive sessions
- Continuous delivery architecture with AWS Glue interactive sessions
- Demo

What is continuous delivery?

Continuous delivery automates the entire software release process. Every revision that is committed triggers an automated flow that builds, tests, and then stages the update. The final decision to deploy to a live production environment is triggered by the developer.



Why continuous delivery

1. Automates the software release process
2. Improve developer productivity
3. Find and address bugs quicker
4. Deliver updates faster
5. Faster feedback and lean
6. Minimize risk

Continuously delivering data products

It's a cultural change

Data products

IT'S MORE THAN JUST DATA

Data + Metadata + Code + Tests + Policies

Data products are complex

- Integrity, constraints, relationships, dependencies
- Continuous schema evolution, complex lineage
- Continuously evolving frameworks
- Optimizing for various production and consumption patterns
- Testing data is hard, automating it is harder
- Security, governance, compliance policies

Continuously delivering data products – Step 1

DELIVER IN SMALLER SEGMENTS & FREQUENTLY

- Break a data product in to multiple features at the smallest possible granularity. I.N.V.E.S.T principle.
- Fast and frequent commits. Once a day at a minimum.
- Smaller synchronous code reviews.
- Less branches and more short-lived branches.

Continuously delivering data products – Step 2

END-TO-END AUTOMATION

- Infrastructure, pipeline logic, governance defined as code.
- Provide developers a blueprint to start from, incorporating best practices.
- Automated tests - integration tests, user acceptance test, regression tests, data reconciliations, performance test, security testing etc.
- Notifications and alerts.
- Able to provide FAST feedback on the success and failure of pipeline executions.

Continuously delivering data products – Step 3

MANAGE DEPLOYMENT HEALTH

- Builds on top of our foundation of automation.
- Purpose-built to verify that a service is working after a new deployment.
- Use feature toggles/tags to reduce blast radius of changes.
- Validate infrastructure health – serverless services help.
- Standardize and automate the rollback process.

Continuously delivering data products – Step 4

IMPLEMENT PIPELINE GOVERNANCE

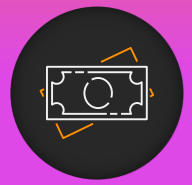
- Build pipeline which blocks production pushes on non-compliant pipelines.
  - Use approvals to pause production deployments.
  - AWS Lambda to automatically approve when pipeline is compliant.
- Build config rules.
  - These alert when pipelines are not configured up to company best practices.
- Resolve conflicting releases.
- Add common actions to all pipelines.

AWS Glue to deliver data products

AWS Glue
Serverless data integration for complex workloads

Serverless
No infrastructure to maintain. Allocate needed compute power and run jobs.

Data integration for every user
Development environments catered to different skillsets - visual ETL development for data engineers, notebook styled development for data scientists, and no code development for data analysts.



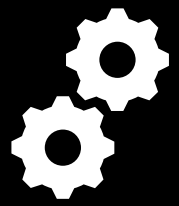
Cost-effective
All-in-one pricing model is 55% cheaper than other cloud data integration solutions.

Handles complex workloads
Connect to 65+ data sources, process petabytes of data in real-time, includes batch and event driven modes.

No lock-in
Develop data integration pipelines in open source SparkSQL, PySpark, Python, Scala.

AWS Glue

SERVERLESS DATA INTEGRATION IN THE CLOUD

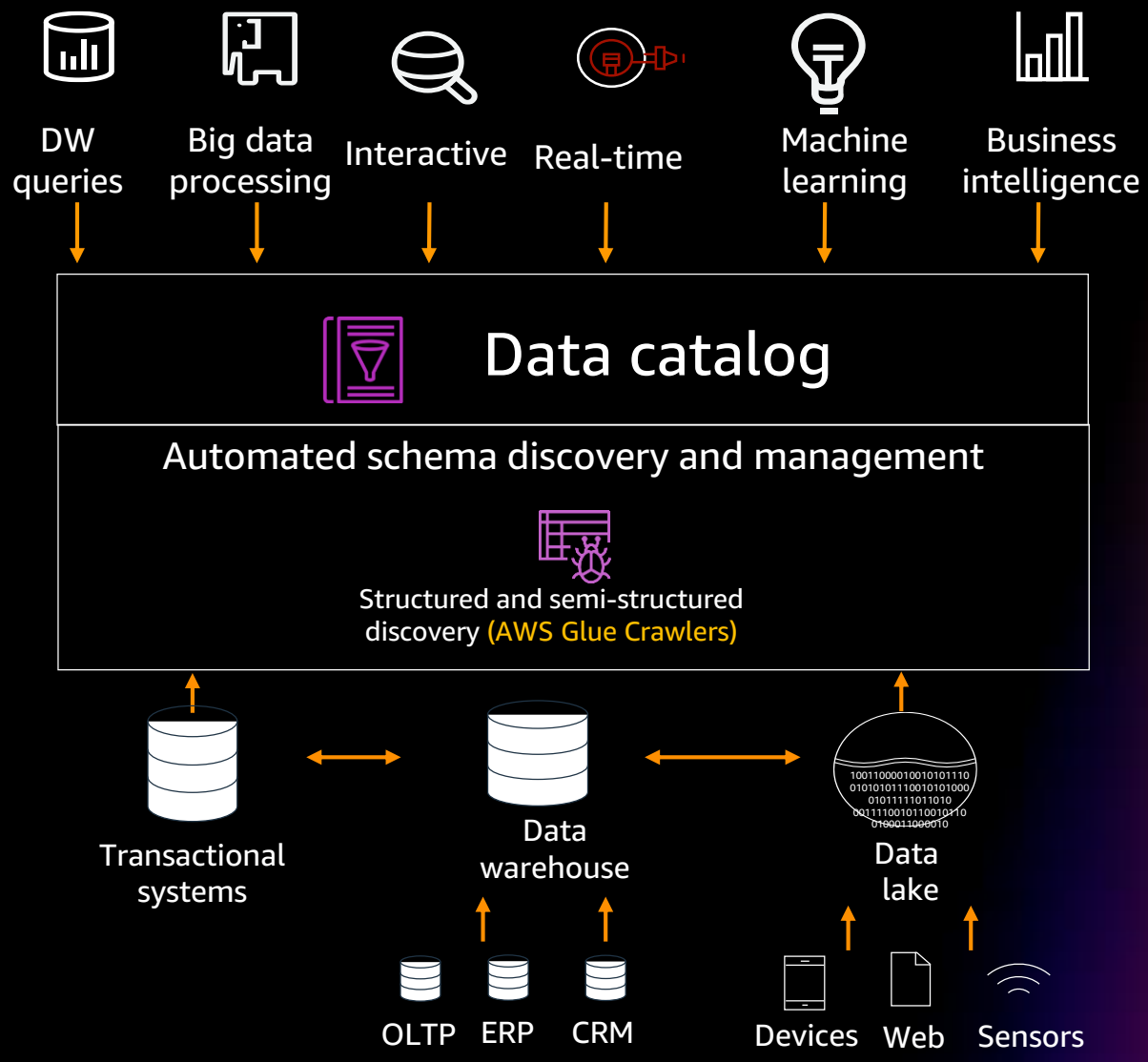


Transform

Deploy

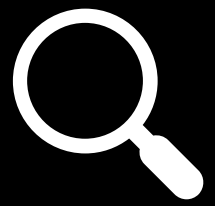
Unified data catalog with automated schema discovery

- No movement of data = Low costs/admin
- All metadata centrally available for search and query = Productivity
- Unify structured, semi-structured data = Speed to insight
- Automate data discovery = Productivity



AWS Glue

SERVERLESS DATA INTEGRATION IN THE CLOUD



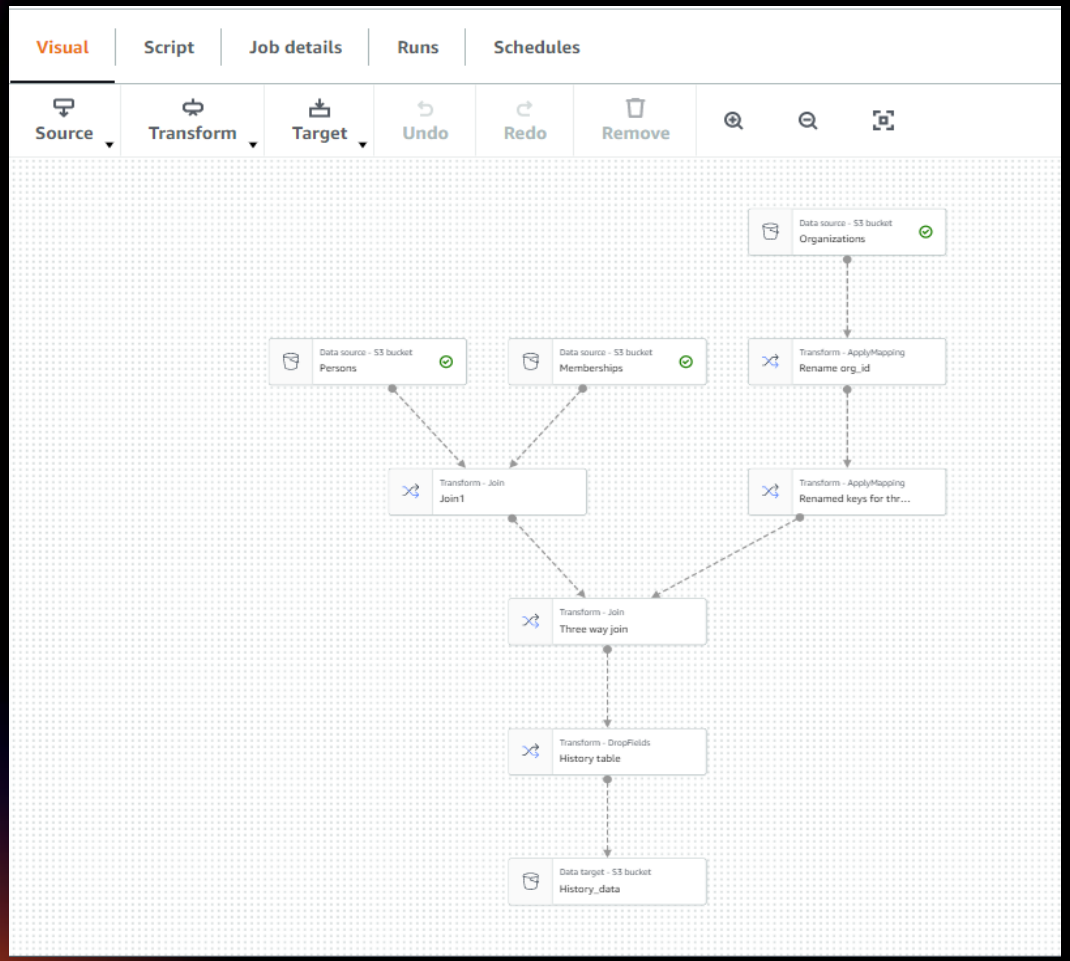
Catalog

Transform

Deploy

AWS Glue Studio

VISUAL JOB AUTHORING AND MONITORING



Monitor thousands of jobs through a single pane of glass

Advanced transforms though code snippets

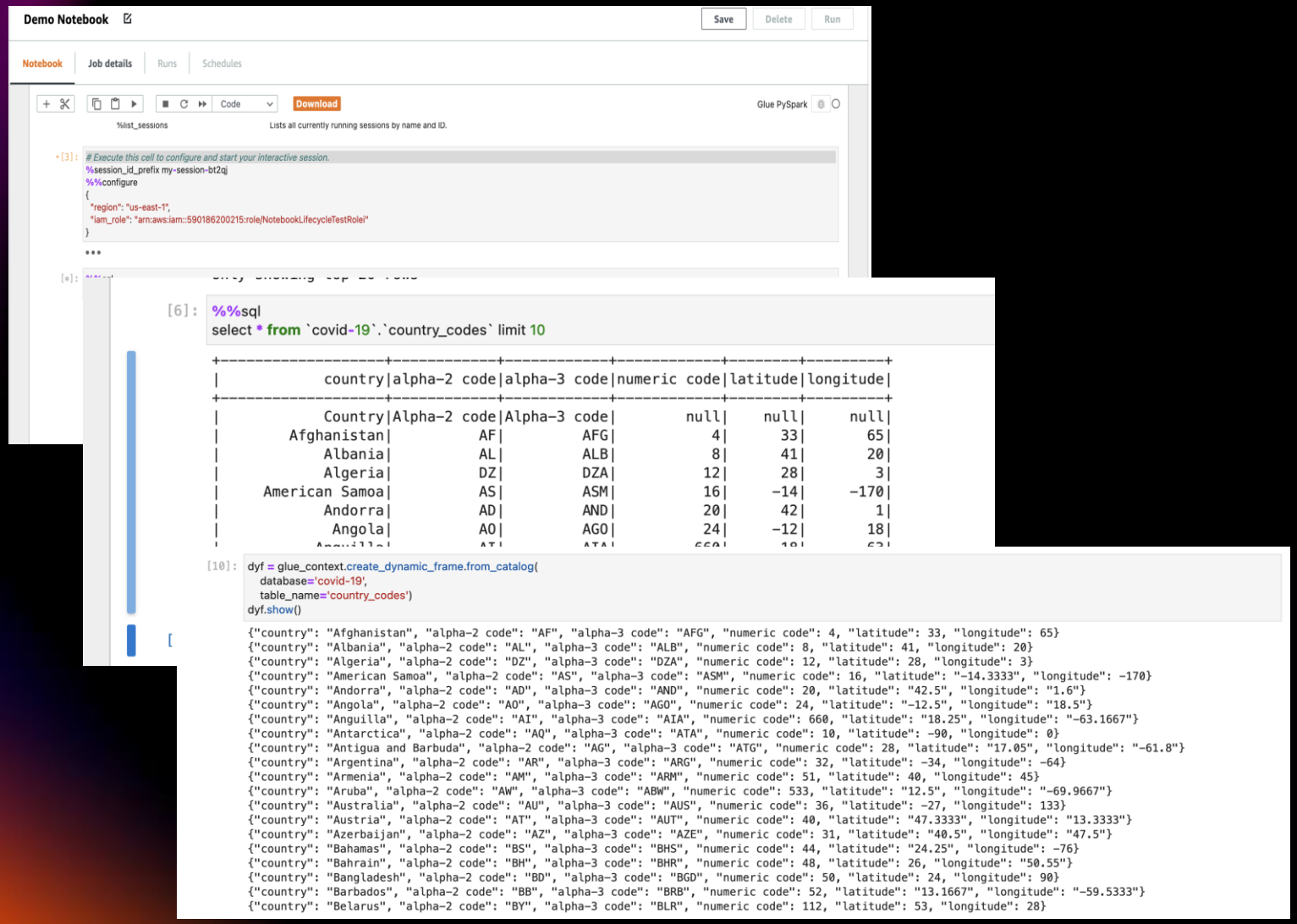
Support for AWS Marketplace and custom connectors

Preview your data at each step of the visual job authoring process

Real-time schema inference without having to catalog

AWS Glue Studio Notebook

New



Interactive AWS Glue jobs development

Submit AWS Glue jobs from the AWS Glue Studio notebook

Use notebook magic to define transforms in SQL and control cost

Built-in monitoring support

AWS Glue interactive sessions

New

Existing options

Time to first query = 10-15 minutes

High cost of a long-running cluster

"Noisy neighbor" problem

AWS Glue interactive sessions

Steps	Task	Time required
1	Connect notebook to sessions API	In seconds
	Time to first query	**~ 1 min**

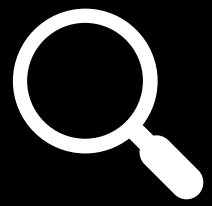
Development tool of your choice

Rapid development

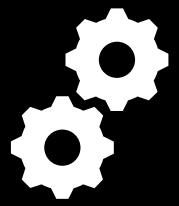
Built-in cost control

AWS Glue

SERVERLESS DATA INTEGRATION IN THE CLOUD



Ingestion



Transform

Deploy

Glue APIs to build CI/CD pipeline

- `create_classifier()`
- `create_connection()`
- `create_crawler()`
- `create_database()`
- `create_dev_endpoint()`
- `create_job()`
- `create_ml_transform()`
- `create_partition()`
- `get_table()`
- `get_table_version()`
- `get_table_versions()`
- `get_tables()`
- `list_crawlers()`
- `list_dev_endpoints()`
- `list_jobs()`
- `list_ml_transforms()`
- `list_triggers()`
- `list_workflows()`

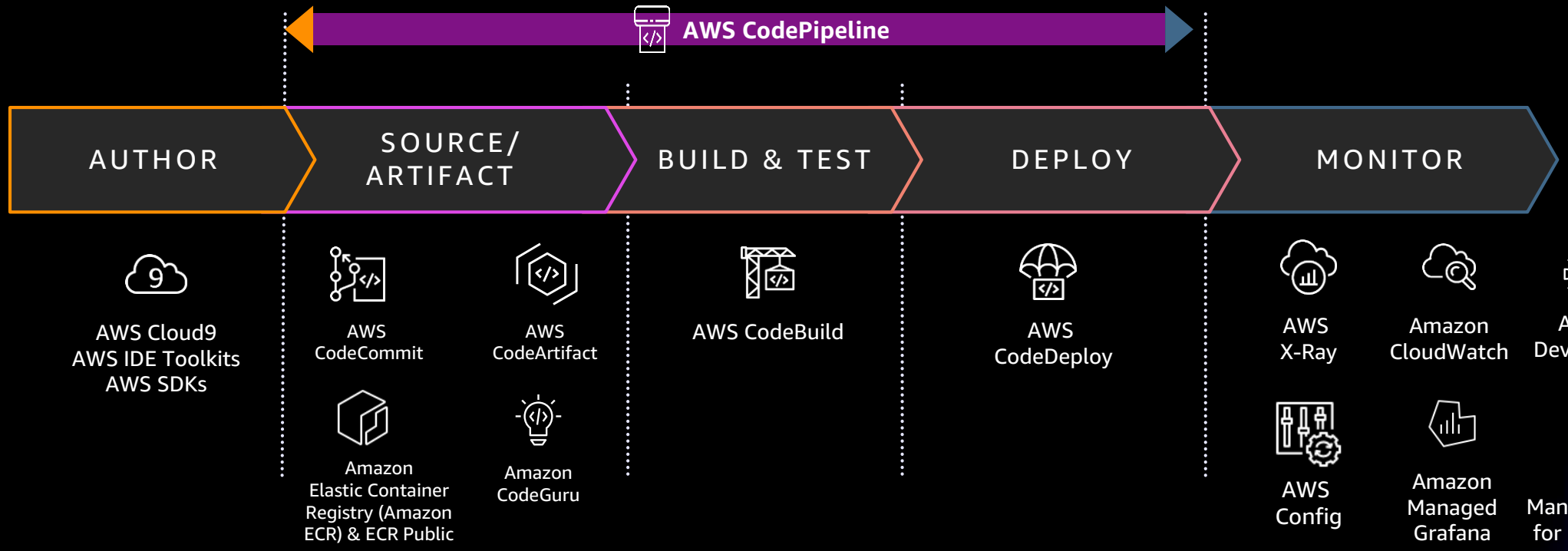
boto3 endpoints to automate CI/CD pipeline

Automate to save development hours

Deploy jobs faster without any manual intervention

Manage data catalog though code snippets

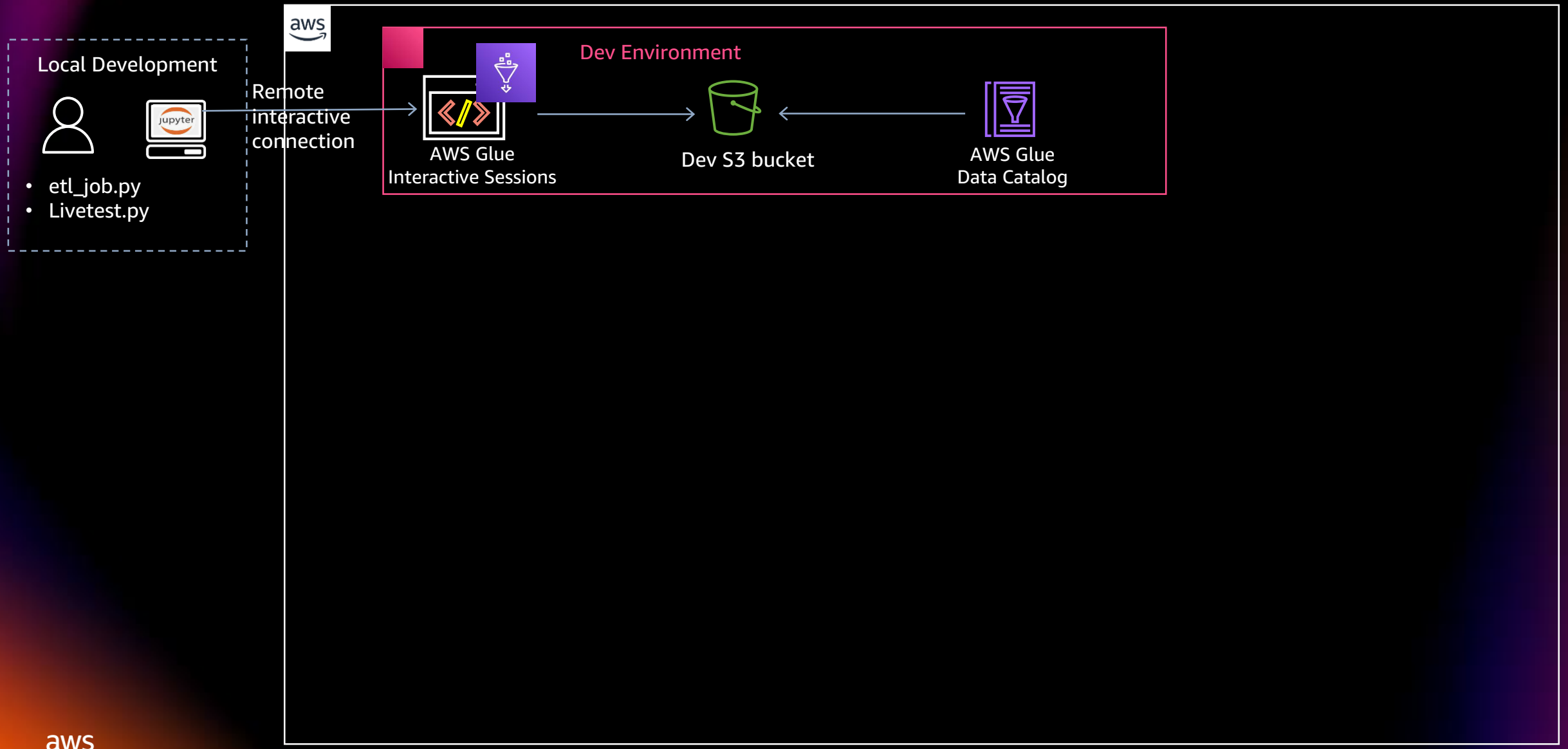
End-to-end solution helps development organizations go faster



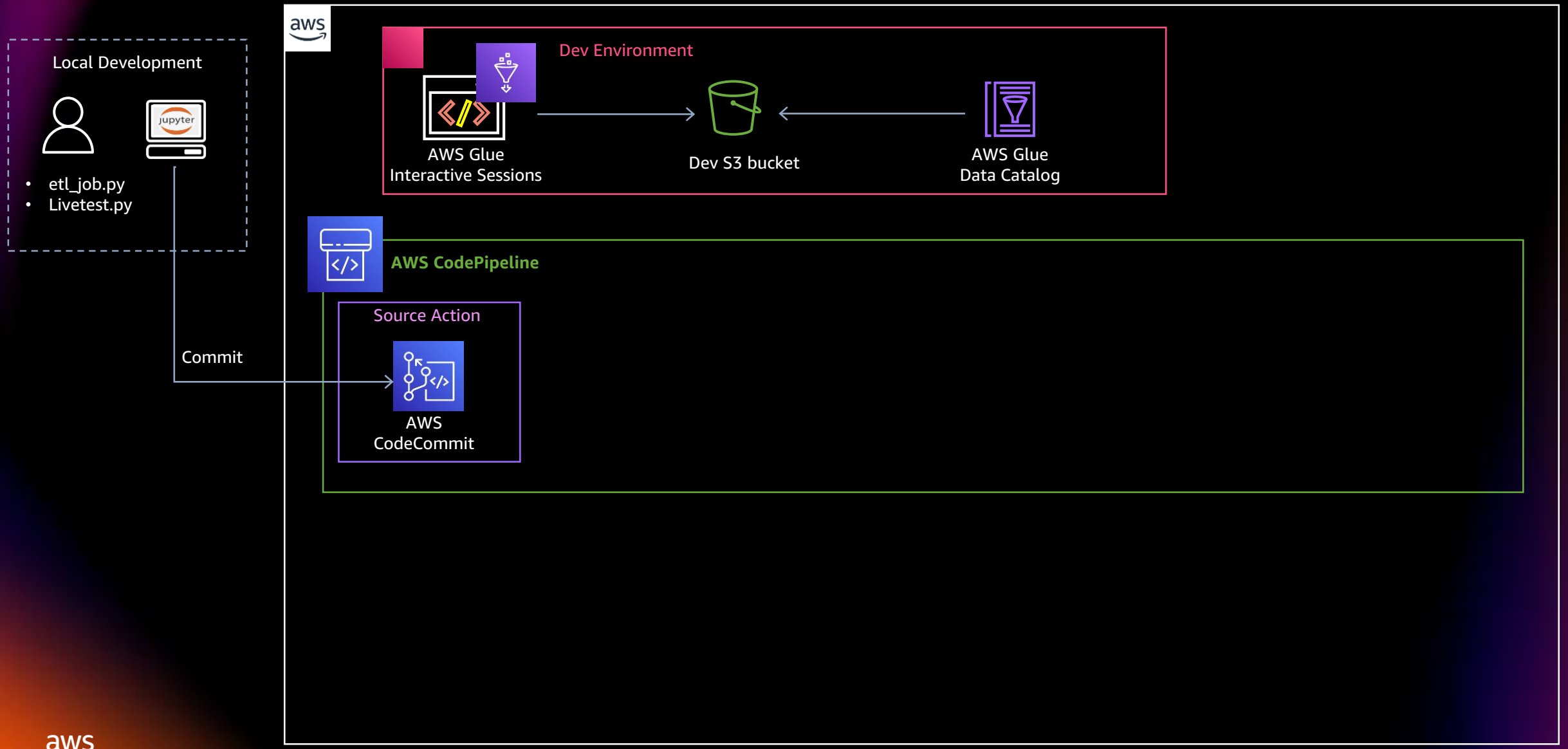
Continuous delivery architecture with AWS Glue interactive sessions

Architecture

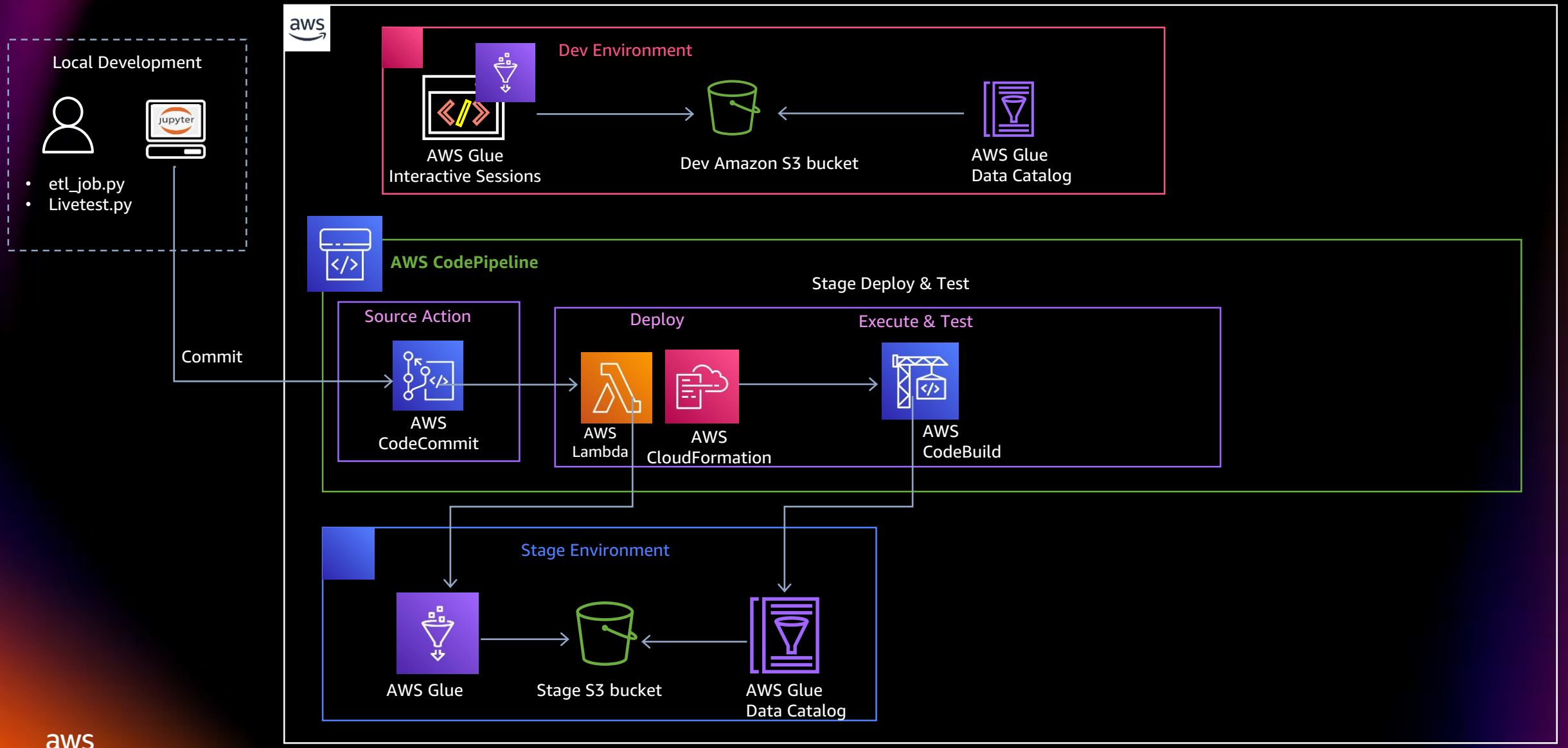
Architecture



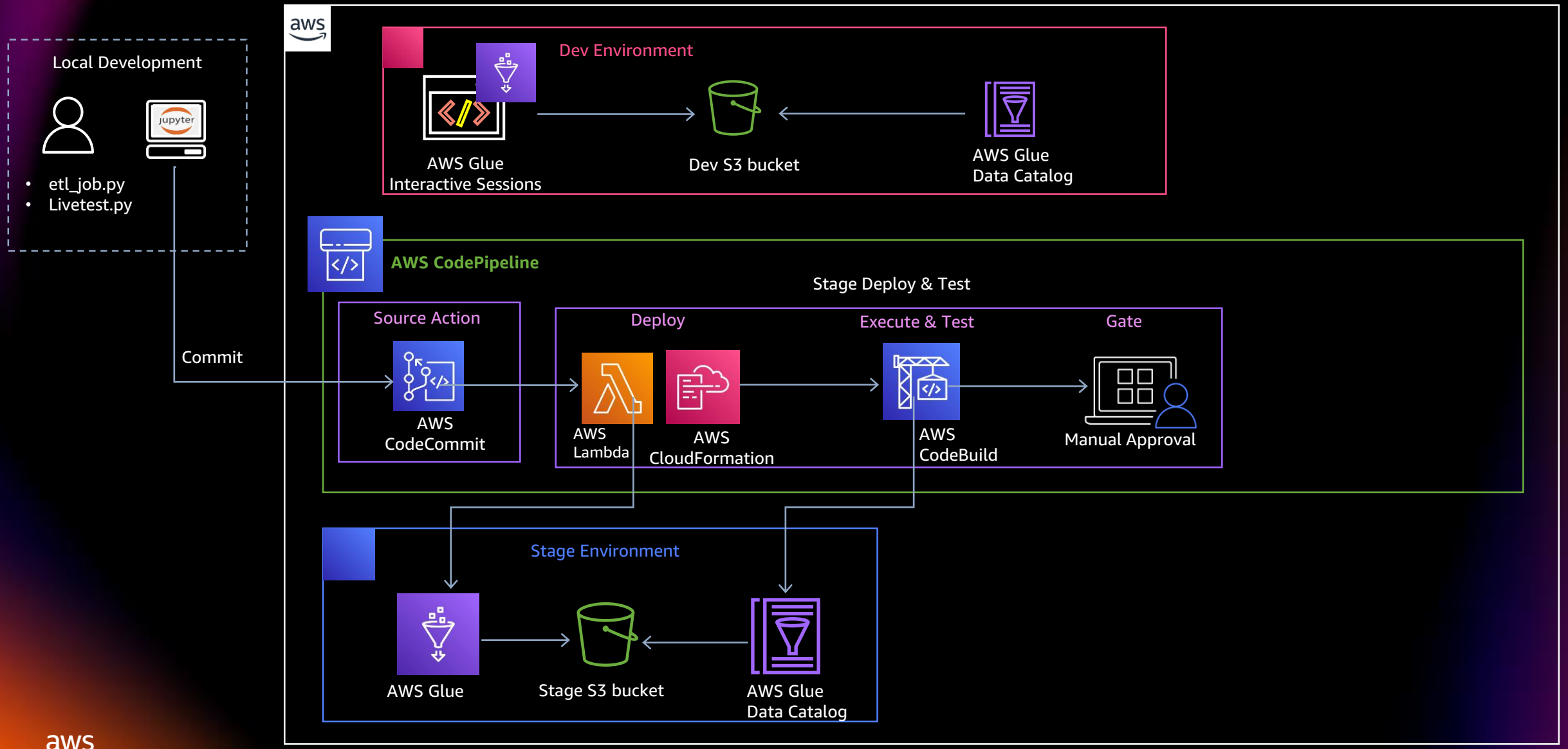
Architecture



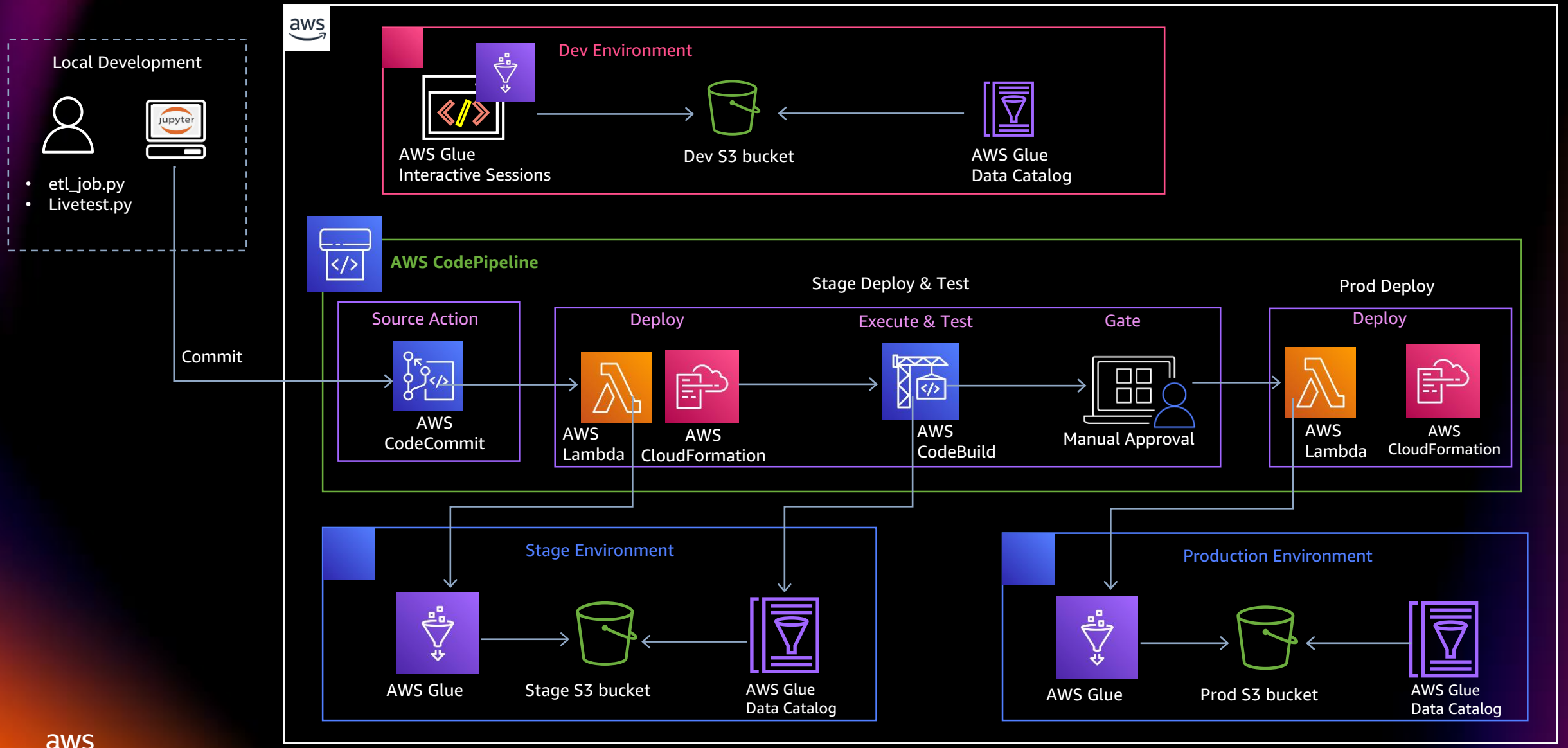
Architecture



Architecture

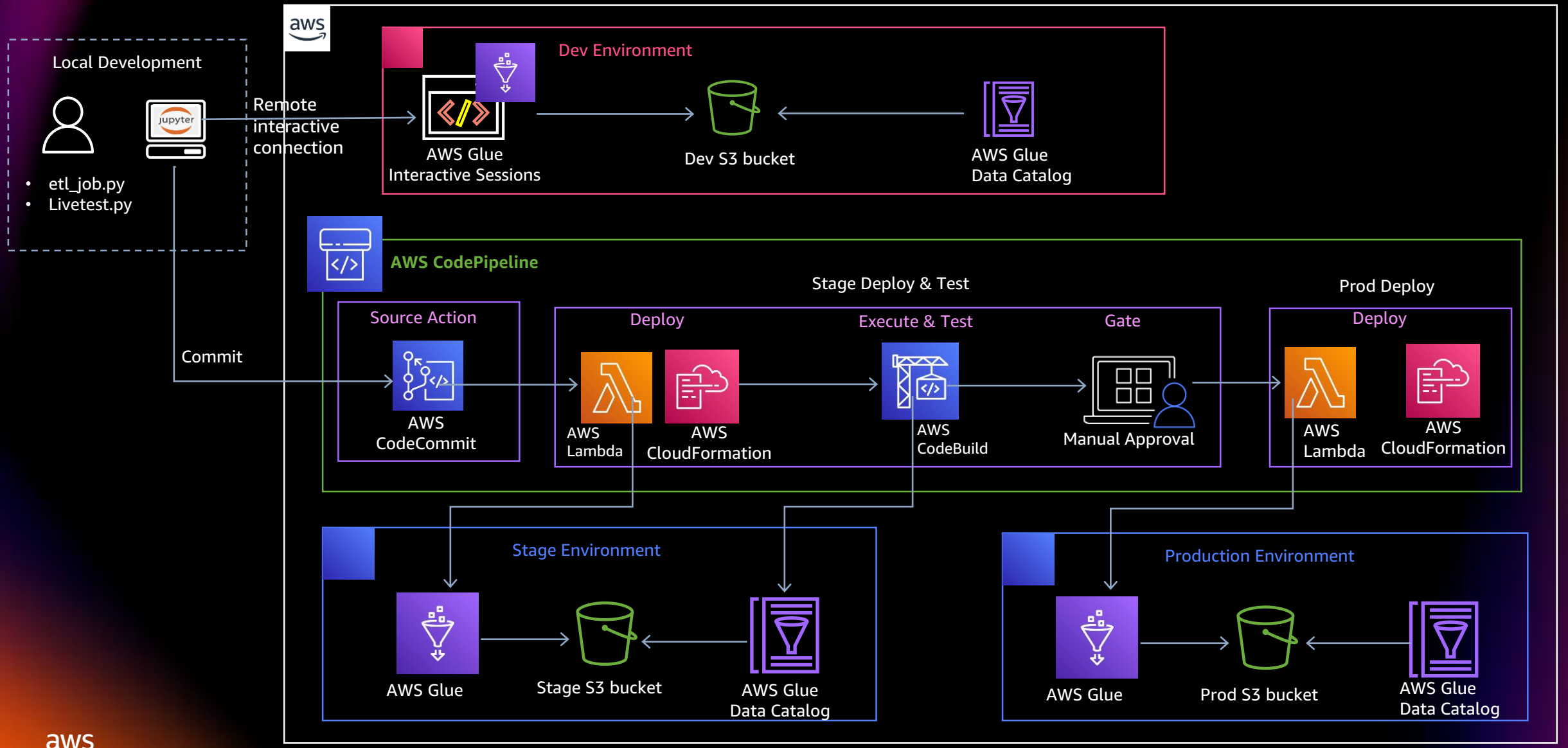


Architecture



Demo

Architecture recap



Other resources

- AWS Continuous Delivery: https://aws.amazon.com/devops/continuous-delivery/
- Deploy an AWS Glue job with an AWS CodePipeline CI/CD pipeline: https://docs.aws.amazon.com/prescriptive-guidance/latest/patterns/deploy-an-aws-glue-job-with-an-aws-codepipeline-ci-cd-pipeline.html
- Getting started with AWS Glue interactive sessions: https://docs.aws.amazon.com/glue/latest/dg/interactive-sessions.html
- AWS Glue Developer Tools: https://docs.aws.amazon.com/whitepapers/latest/aws-overview/developer-tools.html

Visit the AWS Data resource hub

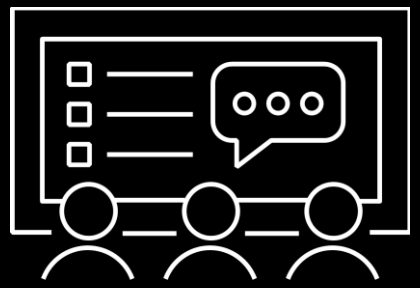
A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
- Put your data to work with a modern analytics approach
- … and more!

https://tinyurl.com/data-hub-aws

Visit resource hub

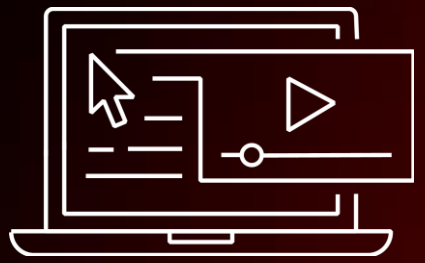
AWS Training and Certification for Data and Analytics



AWS Data & Analytics FREE Training Resources

Discover how to harness data, one of the world's most valuable resources, and innovate at scale.

https://bit.ly/3Ntlhy7



AWS Data Analytics Learning Plan

This learning plan expose you to the fastest way to get answers from all your data to all your users. It can also help prepare you for the AWS Certified Data Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

AWS Certified Data Analytics - Specialty

Earning AWS Certified Data Analytics – Specialty validates expertise in using AWS data lakes and analytics services.

https://go.aws/3lwF0RR

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apj-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

Thank you!