



aws INNOVATE

DATA EDITION

23 August, 2022

Simplify data integration using AWS Glue

Nico Anandito

Analytics Specialist Solutions Architect

Amazon Web Services



Agenda

1. AWS Glue introduction
2. AWS Glue to simplify data integration
 - Ingest
 - Transform
 - Operationalize
3. Demo
4. Summary

Data integration is hard

Data



GROWING
EXPONENTIALLY



FROM NEW
SOURCES



INCREASINGLY
DIVERSE

Data integration is hard

Data



GROWING
EXPONENTIALLY



FROM NEW
SOURCES



INCREASINGLY
DIVERSE

Personas



NO OR LOW CODE



DEVELOPERS



DATA ANALYSTS AND
DATA SCIENTISTS

Data integration is hard

Data



GROWING
EXPONENTIALLY



FROM NEW
SOURCES



INCREASINGLY
DIVERSE

Personas



NO OR LOW CODE

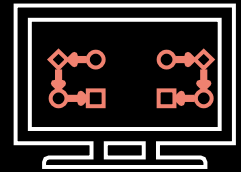


DEVELOPERS



DATA ANALYSTS AND
DATA SCIENTISTS

Applications



REAL-TIME/SLA
SENSITIVE



HIGHLY SCALABLE



PRICE PERFORMANCE

Data integration platform trends



Scalable infrastructure



Open standards

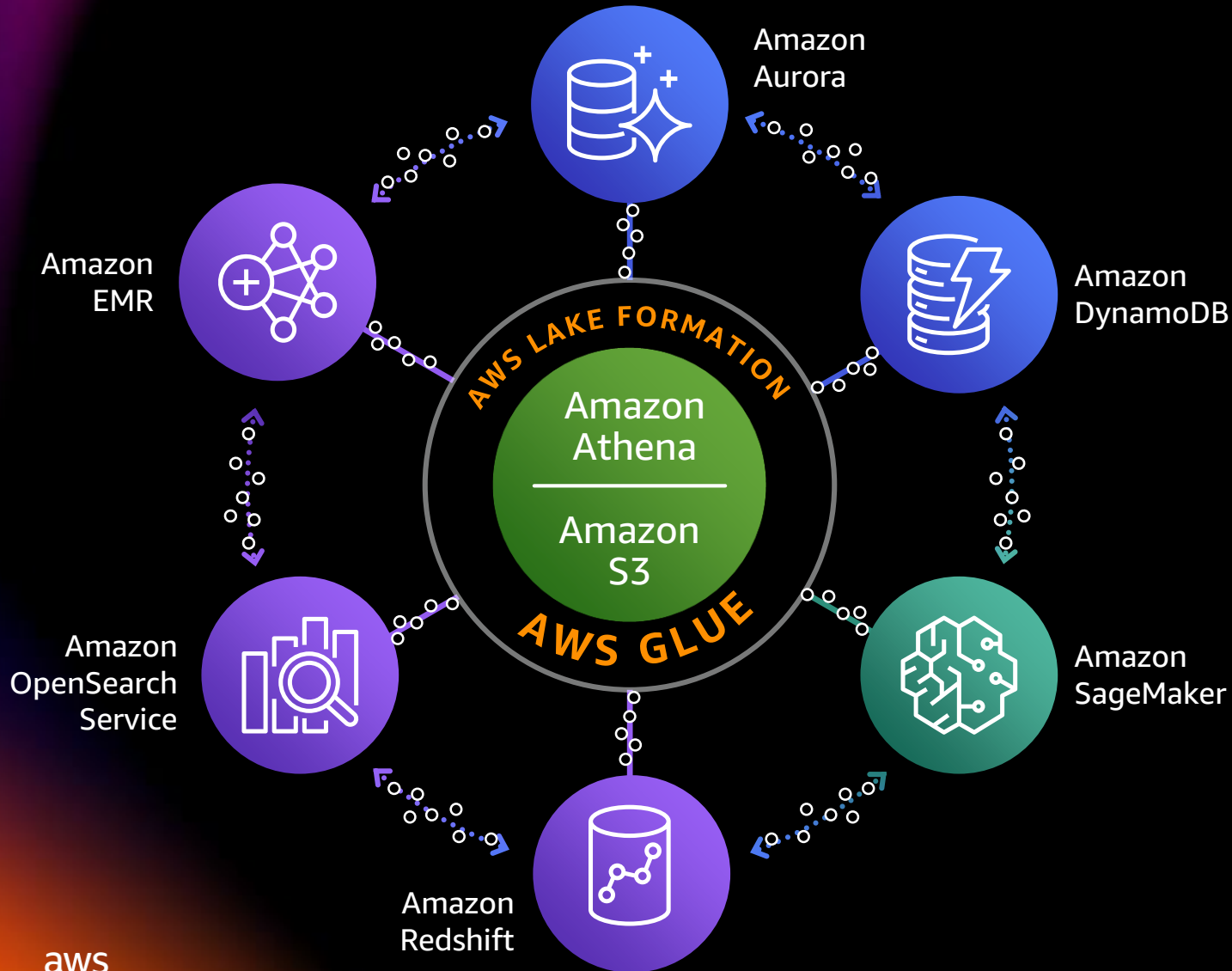


Low cost



Tools for all personas

Modern data architecture with AWS Glue



**DATA INTEGRATION IN BATCH
AND REALTIME**

**PERFORMANT AND
COST-EFFECTIVE**

**CENTRALIZED CATALOG AND
GOVERNANCE**

TOOLS FOR DIVERSE SKILLSETS

AWS Glue Serverless data integration for complex workloads



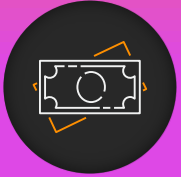
Serverless

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

Globe Telecom develops a 360-Degree customer view on AWS

Building a robust subscriber profile for more than 90 millions customers using AWS Glue

Can onboard 40 times more user attributes a month

High platform availability

Integrates easily with downstream applications



“Now, more than ever, multiple downstream applications and analytical functions have access to real-time behavioral data, placing us in a stronger position to deliver more relevant and meaningful interactions with each of our customers. We can personalize engagements from messaging, real-time offers, to product bundles and more”

Derick Adil

Director, Asset Delivery and Domain Integration, Globe Telecom

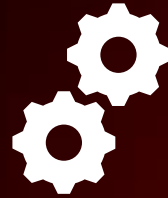
Read more: <https://aws.amazon.com/solutions/case-studies/globe-telecom-cadenz/>

AWS Glue

Serverless Data Integration in the Cloud



Ingestion



Transform



Deploy

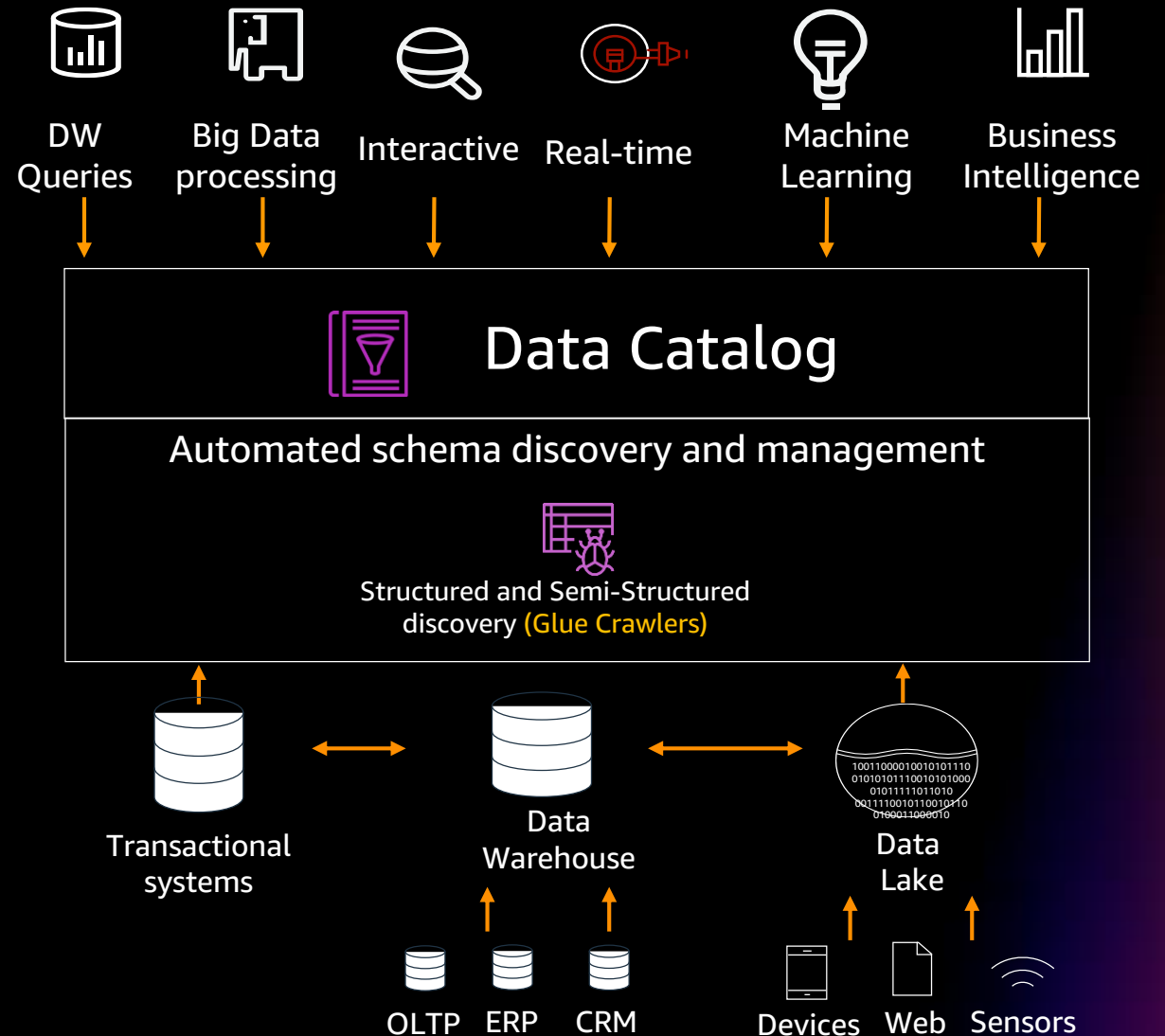
Unified Data Catalog with automated schema discovery

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**



Custom Connectors with AWS Glue

Data Sources

On-premises DBs

Proprietary stores

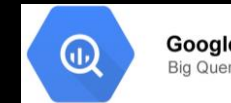
SaaS applications



CUSTOM CONNECTOR

- No additional cost for connecting to sources
- Flexible and easy to build connectors

AWS Glue Connectors Marketplace



+ Many more...

[Learn more](#)



AWS Glue

Serverless Data Integration in the Cloud



Ingestion



Transform



Deploy

AWS Glue Execution Engine



Diverse workloads

Serverless **Apache Spark** and
Python environment



Fast and predictable

job starts in **seconds**
Reduced job latencies
enabling micro-batching

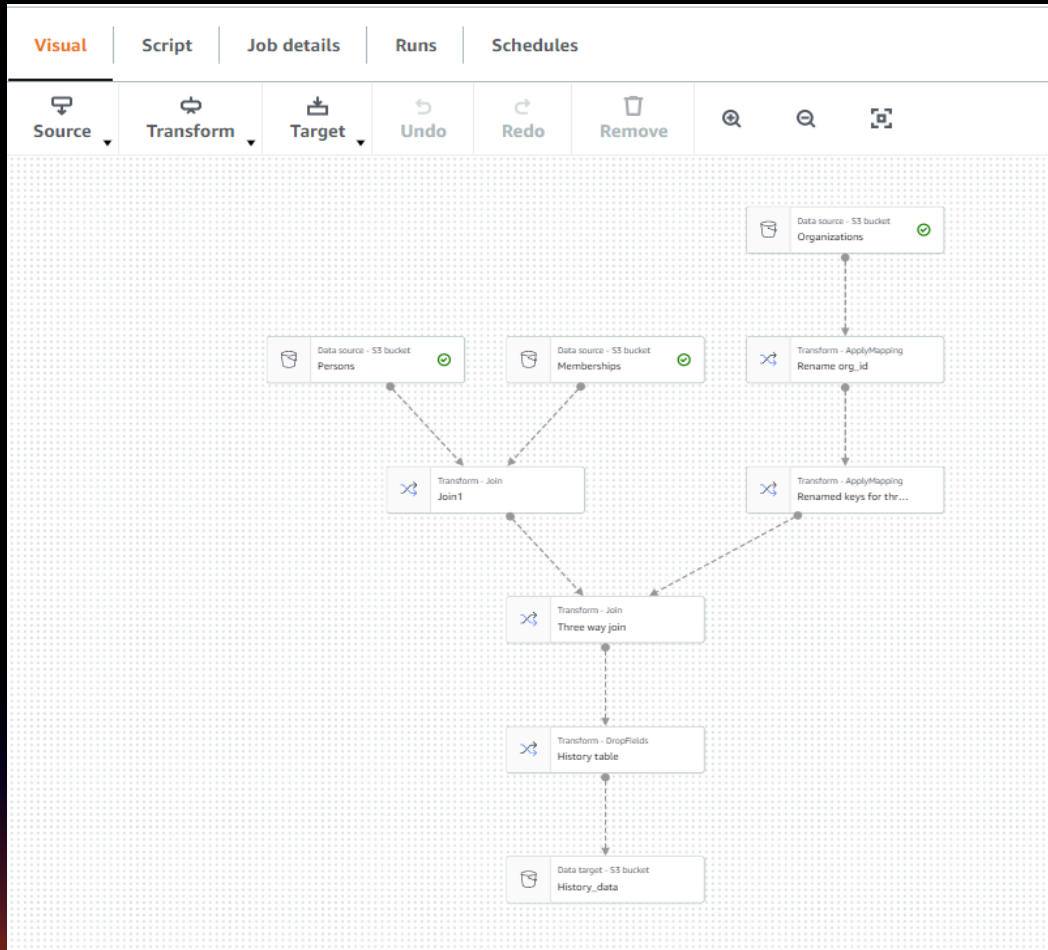


Cost effective

**Per second billing with a
1-minute** minimum billing

AWS Glue Studio

VISUAL JOB AUTHORIZING AND MONITORING



Monitor **thousands of jobs** through a **single pane of glass**

Advanced transforms **through code snippets**

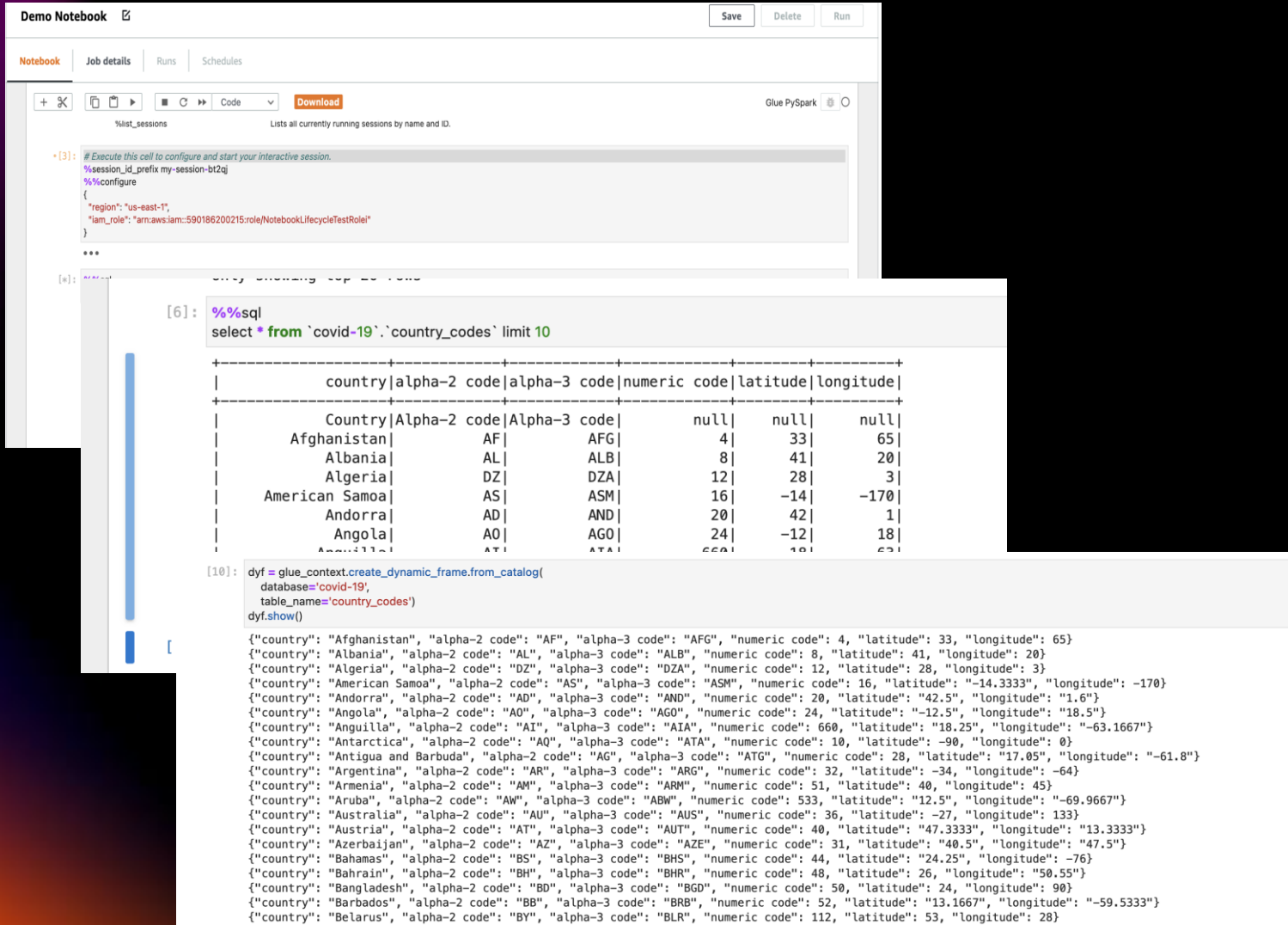
Support for **AWS Marketplace** and custom **connectors**

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



The screenshot displays the AWS Glue Studio Notebook interface. At the top, there's a 'Demo Notebook' header with 'Save', 'Delete', and 'Run' buttons. Below this, a toolbar includes icons for adding, deleting, and running code, along with a 'Download' button. The main area shows a notebook with two code cells. The first cell, labeled '[3]:', contains configuration code for an interactive session. The second cell, labeled '[6]:', executes a SQL query: `select * from `covid-19`.`country_codes` limit 10`. The result of this query is displayed as a table with columns: country, alpha-2 code, alpha-3 code, numeric code, latitude, and longitude. Below the table, a third code cell, labeled '[10]:', shows the creation of a dynamic frame from the catalog and its display.

```
[3]: # Execute this cell to configure and start your interactive session.
%%session_id_prefix my-session-bt2qj
%%configure
{
  "region": "us-east-1",
  "iam_role": "arn:aws:iam::590186200215:role/NotebookLifecycleTestRole"
}
***

[6]: %%sql
select * from `covid-19`.`country_codes` limit 10

+-----+-----+-----+-----+-----+-----+
| country | alpha-2 code | alpha-3 code | numeric code | latitude | longitude |
+-----+-----+-----+-----+-----+-----+
| Afghanistan | AF | AFG | 4 | 33 | 65 |
| Albania | AL | ALB | 8 | 41 | 20 |
| Algeria | DZ | DZA | 12 | 28 | 3 |
| American Samoa | AS | ASM | 16 | -14 | -170 |
| Andorra | AD | AND | 20 | 42 | 1 |
| Angola | AO | AGO | 24 | -12 | 18 |
| Antigua and Barbuda | AG | ATG | 10 | 17.05 | -61.8 |
| Argentina | AR | ARG | 32 | -34 | -64 |
| Armenia | AM | ARM | 51 | 40 | 45 |
| Aruba | AW | ABW | 533 | 12.5 | -69.9667 |
| Australia | AU | AUS | 36 | -27 | 133 |
| Austria | AT | AUT | 40 | 47.3333 | 13.3333 |
| Azerbaijan | AZ | AZE | 31 | 40.5 | 47.5 |
| Bahamas | BS | BHS | 44 | 24.25 | -76 |
| Bahrain | BH | BHR | 48 | 26 | 50.55 |
| Bangladesh | BD | BGD | 50 | 24 | 90 |
| Barbados | BB | BRB | 52 | 13.1667 | -59.5333 |
| Belarus | BY | BLR | 112 | 53 | 28 |

[10]: dyf = glue_context.create_dynamic_frame.from_catalog(
      database='covid-19',
      table_name='country_codes')
dyf.show()

{"country": "Afghanistan", "alpha-2 code": "AF", "alpha-3 code": "AFG", "numeric code": 4, "latitude": 33, "longitude": 65}
{"country": "Albania", "alpha-2 code": "AL", "alpha-3 code": "ALB", "numeric code": 8, "latitude": 41, "longitude": 20}
{"country": "Algeria", "alpha-2 code": "DZ", "alpha-3 code": "DZA", "numeric code": 12, "latitude": 28, "longitude": 3}
{"country": "American Samoa", "alpha-2 code": "AS", "alpha-3 code": "ASM", "numeric code": 16, "latitude": "-14.3333", "longitude": -170}
{"country": "Andorra", "alpha-2 code": "AD", "alpha-3 code": "AND", "numeric code": 20, "latitude": "42.5", "longitude": "1.6"}
{"country": "Angola", "alpha-2 code": "AO", "alpha-3 code": "AGO", "numeric code": 24, "latitude": "-12.5", "longitude": "18.5"}
{"country": "Anguilla", "alpha-2 code": "AI", "alpha-3 code": "AIA", "numeric code": 660, "latitude": "18.25", "longitude": "-63.1667"}
{"country": "Antarctica", "alpha-2 code": "AQ", "alpha-3 code": "ATA", "numeric code": 10, "latitude": -90, "longitude": 0}
{"country": "Antigua and Barbuda", "alpha-2 code": "AG", "alpha-3 code": "ATG", "numeric code": 28, "latitude": "17.05", "longitude": "-61.8"}
{"country": "Argentina", "alpha-2 code": "AR", "alpha-3 code": "ARG", "numeric code": 32, "latitude": -34, "longitude": -64}
{"country": "Armenia", "alpha-2 code": "AM", "alpha-3 code": "ARM", "numeric code": 51, "latitude": 40, "longitude": 45}
{"country": "Aruba", "alpha-2 code": "AW", "alpha-3 code": "ABW", "numeric code": 533, "latitude": "12.5", "longitude": "-69.9667"}
{"country": "Australia", "alpha-2 code": "AU", "alpha-3 code": "AUS", "numeric code": 36, "latitude": -27, "longitude": 133}
{"country": "Austria", "alpha-2 code": "AT", "alpha-3 code": "AUT", "numeric code": 40, "latitude": "47.3333", "longitude": "13.3333"}
{"country": "Azerbaijan", "alpha-2 code": "AZ", "alpha-3 code": "AZE", "numeric code": 31, "latitude": "40.5", "longitude": "47.5"}
{"country": "Bahamas", "alpha-2 code": "BS", "alpha-3 code": "BHS", "numeric code": 44, "latitude": "24.25", "longitude": "-76"}
{"country": "Bahrain", "alpha-2 code": "BH", "alpha-3 code": "BHR", "numeric code": 48, "latitude": 26, "longitude": "50.55"}
{"country": "Bangladesh", "alpha-2 code": "BD", "alpha-3 code": "BGD", "numeric code": 50, "latitude": 24, "longitude": 90}
{"country": "Barbados", "alpha-2 code": "BB", "alpha-3 code": "BRB", "numeric code": 52, "latitude": "13.1667", "longitude": "-59.5333"}
{"country": "Belarus", "alpha-2 code": "BY", "alpha-3 code": "BLR", "numeric code": 112, "latitude": 53, "longitude": 28}
```

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

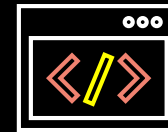
Time to first query = 10-15 minutes

High cost of a long-running cluster

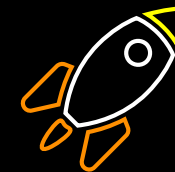
"Noisy Neighbor" problem

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

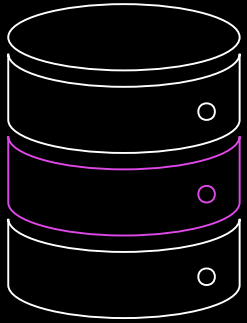


Rapid
development



Built-in cost
control

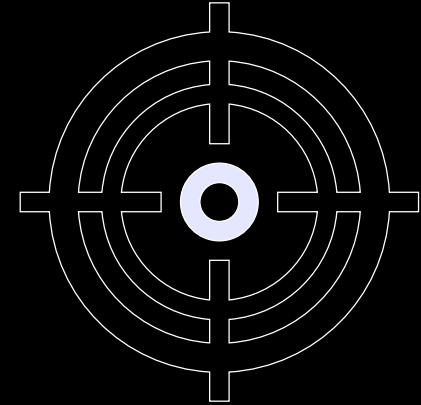
Challenges with PII detection and remediation



Size of the dataset



Location of PII



Accuracy

AWS Glue PII Detection and remediation

New

THREE SIMPLE STEPS

1

Type of Scan



Full scan

Sample scan

2

Entities to detect



Built-in entities
(e.g. SSN, passport)

Custom entities

3

Remediation

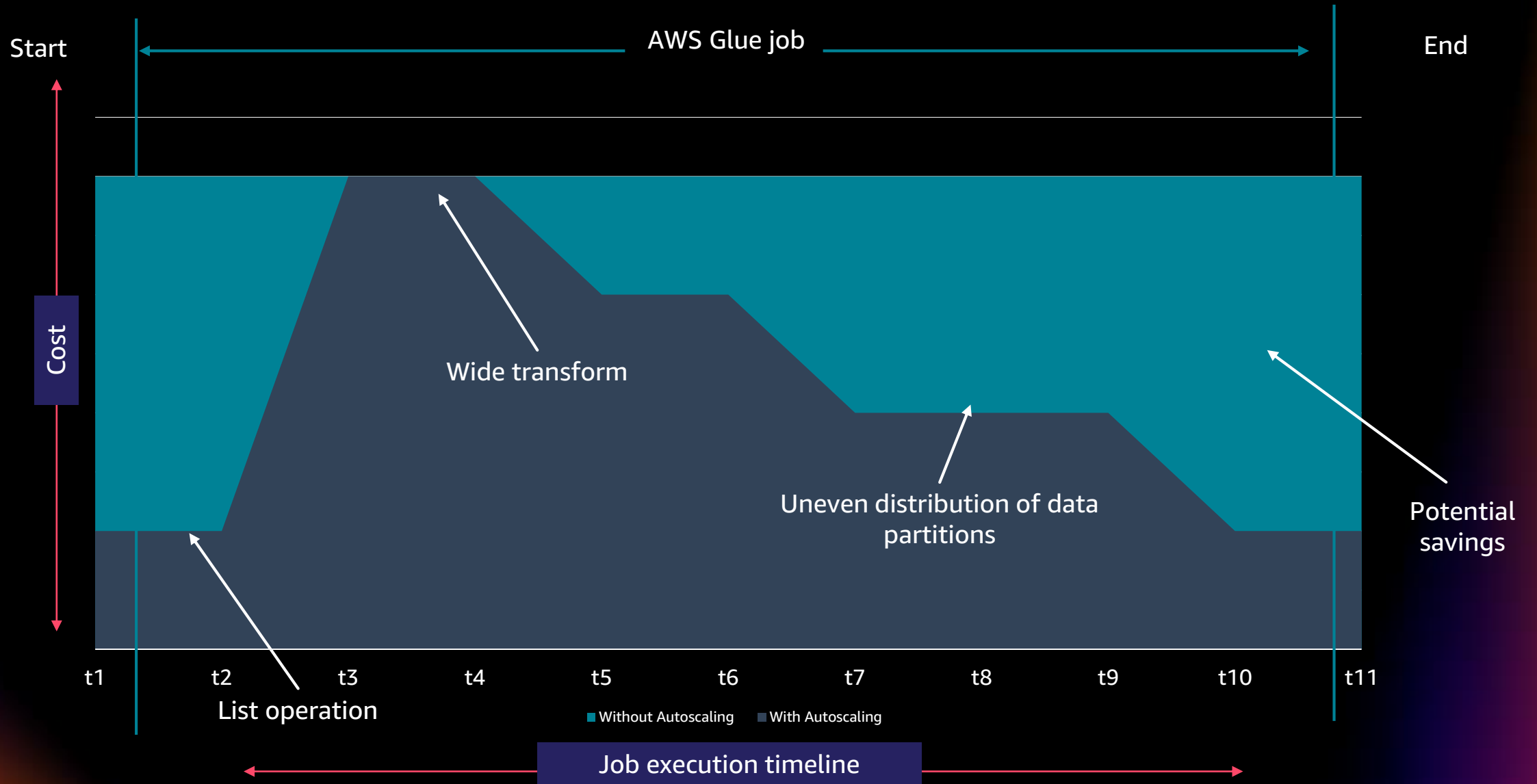


Store results

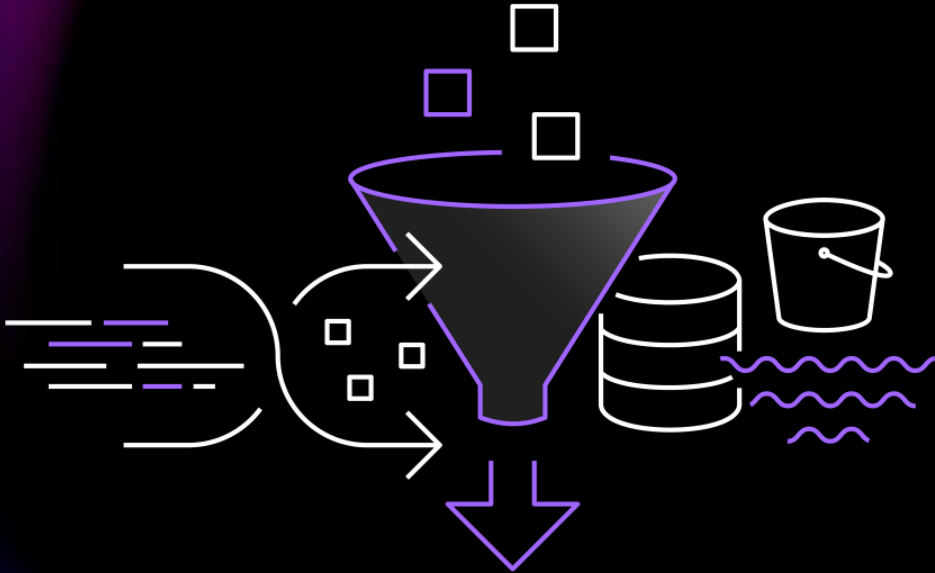
Redact/mask
results

AWS Glue Auto-scaling

New



AWS Glue Streaming ETL



Process stream data & it can be queried in seconds

Join streams against each other or static data

Automatic updates to the AWS Glue Data Catalog

Dozens of supported data **targets**

Simplify your architecture with **one service** for streaming and batch data integration

AWS Glue

Serverless Data Integration in the Cloud



Ingestion



Transform



Deploy

Monitoring dashboard to check job status

Monitor Job Runs [Info](#)

7 Day ▼

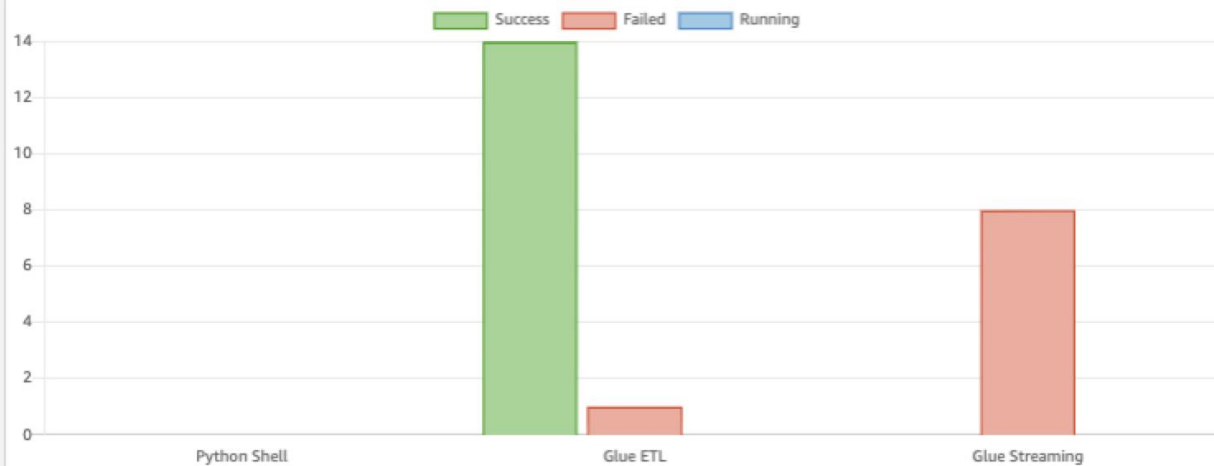
Job Runs Summary

Total Runs	Running	Canceled	Success	Failed
347	23	0	246	78

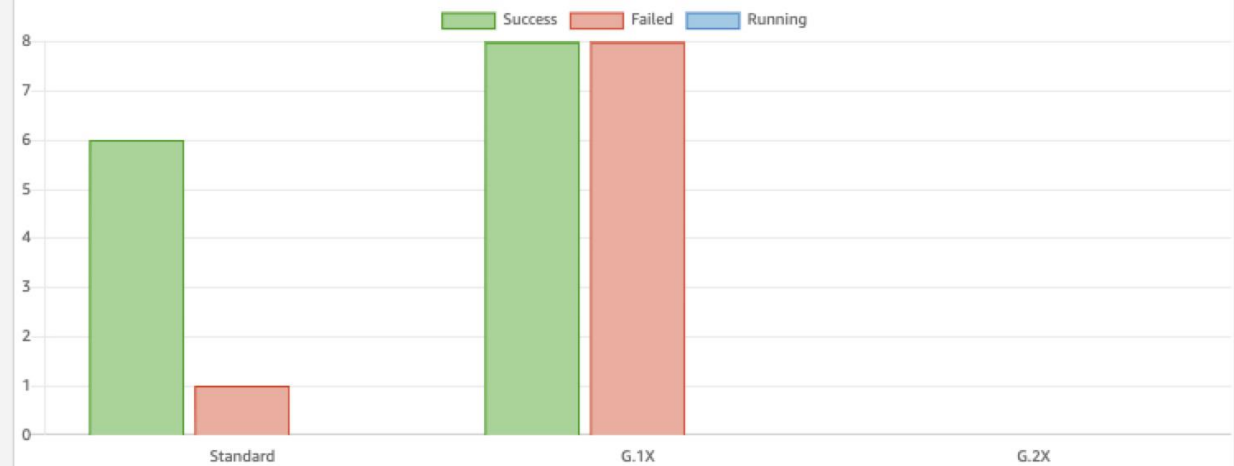
Job Run Success Rate [Info](#)

Success Rate	Status
71%	🟢 Service operating normally

Job Type Breakdown [Info](#)



Worker Type Breakdown [Info](#)



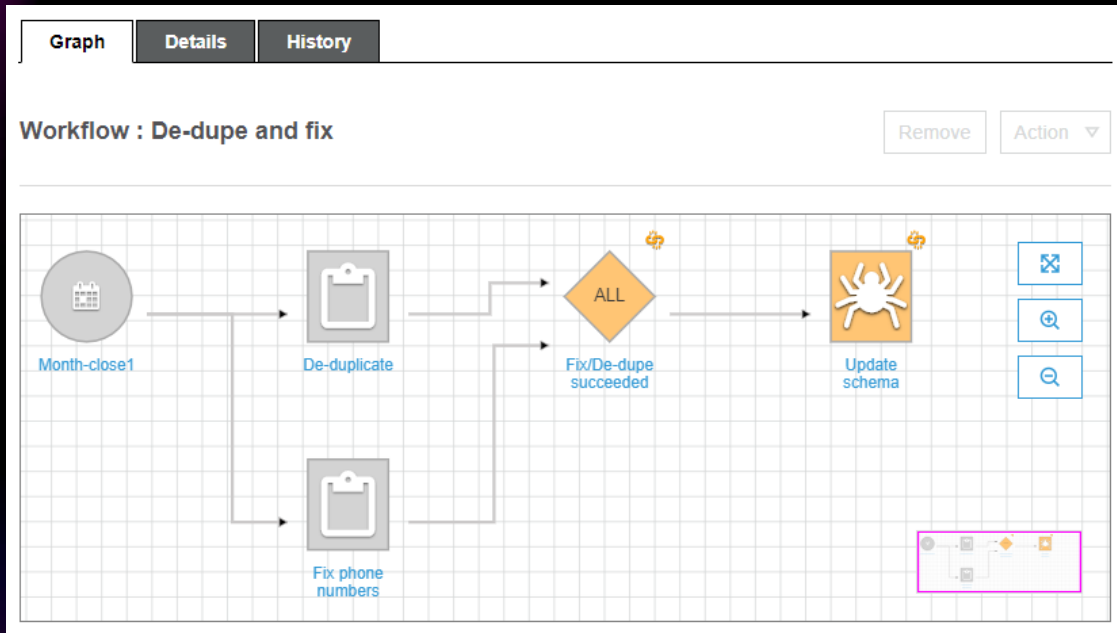
Job Runs Timeline [Info](#)

Legend: Success (Green), Failed (Red), Running (Blue), Canceled (Gray)

Estimated Job DPU Usage [Pricing Info](#)



Orchestrate jobs easily with AWS Glue workflows

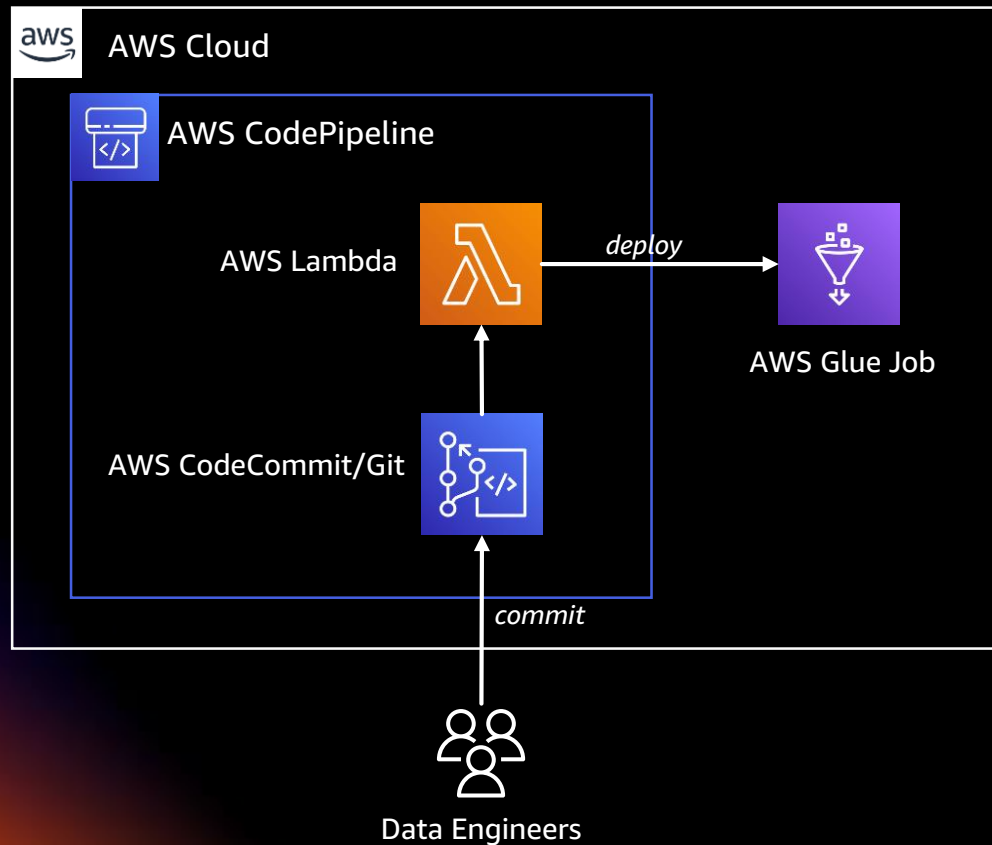


Orchestrate AWS Glue jobs and other AWS services

Schedule jobs or trigger based on events

Monitor execution of the workflows in one place

AWS Glue APIs to build CI/CD pipeline



BOTO3 Endpoints to automate CI/CD pipeline

Automate to save development hours

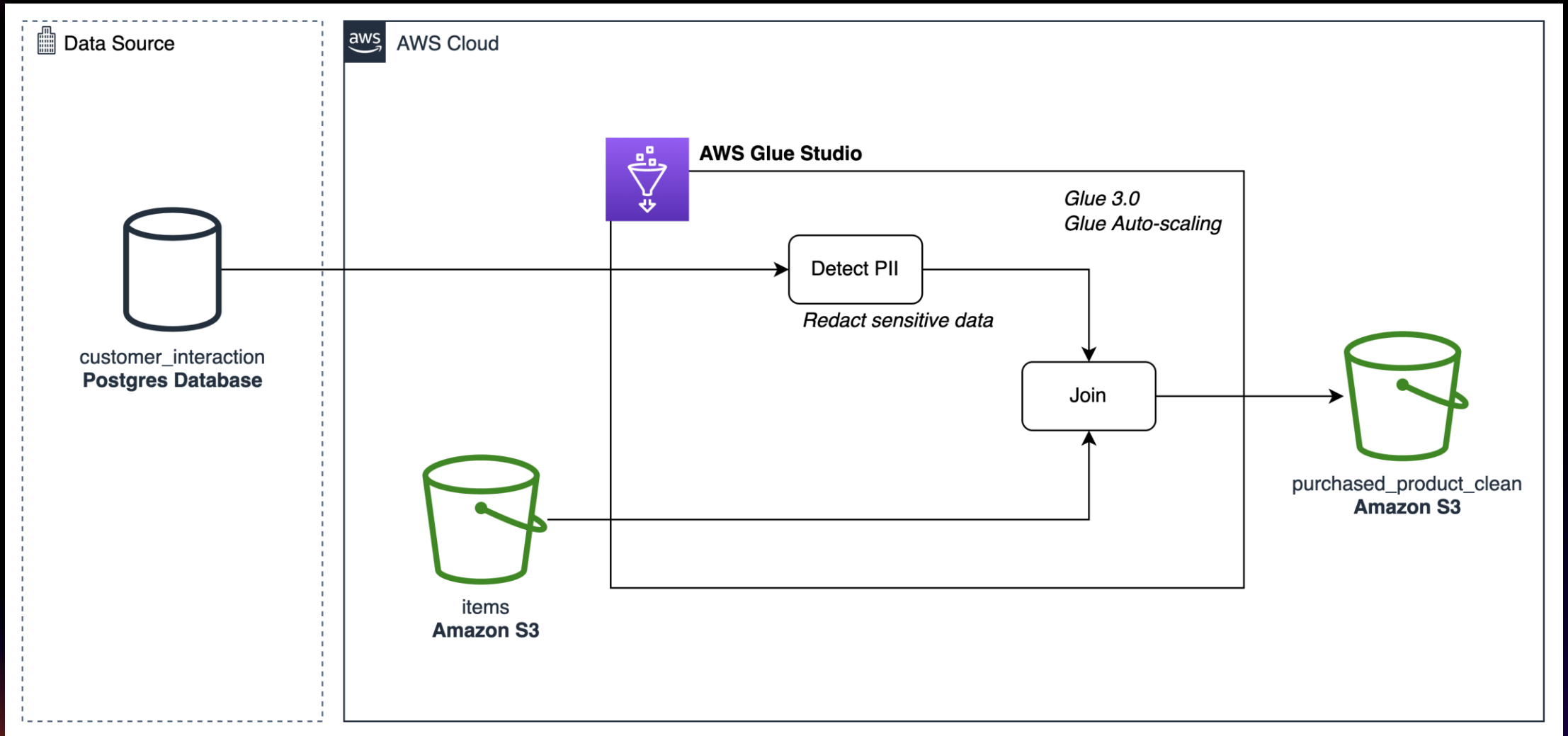
Deploy jobs faster without any manual intervention

Manage Data Catalog **through code snippets**

Demo



Demo architecture



Summary

AWS Glue to simplify data integration in the cloud



No-code to advanced data use cases



Process petabytes of data both in batch and real-time using Apache Spark



Migrate from expensive traditional ETL solutions to gain flexibility and reduce costs



Catalog data assets to make them available to AWS Analytics Services

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](#)

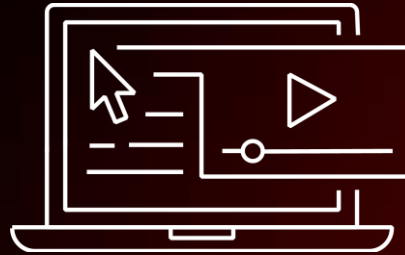
AWS Training and Certification for Data and Analytics



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!