
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ANT 223

Simplify and accelerate data integration and ETL modernization with AWS Glue

Santosh Chandrachood
General Manager
AWS Glue

Moisés Nascimento
Chief Data Officer
Itaú Unibanco

Roberto Figueira
Head of D&A Platform Engineering
Itaú Unibanco

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Trends and challenges in Data integration

AWS Glue overview, new innovations and announcements

Why Itaú choose AWS Glue for data mesh architecture

Conclusion and next steps

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customers

Hundreds of thousands of AWS Glue customers

Financial services













Software & internet











Healthcare











Industrial











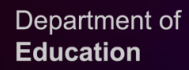
Government











aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



BMW built self service data integration based on AWS Glue into its Data Platform which is used by 5K users every month, running 20K Glue ETL jobs Glues fast startup times allowed BMW to deliver data faster than ever and using Glue UI analysts accelerated the time it takes to create transformation jobs

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key Glue launches in 2022

AWS Glue pace of innovation in 2022 with more than 18+ new capabilities

Q1

- Job Run Insights
- FindMatches 2.0

Q2

- Interactive Sessions
- Auto Scaling
- Glue Studio Notebook
- Sensitive data detection
- Visual job API
- Streaming SASL
- Streaming de compression

Q3

- Flex
- Python Shell 3.9
- Small workers for streaming
- Streaming auto scale

Q4

- Glue git integration

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Gartner Data Integration Magic Quadrant

- Native integration with other AWS cloud services makes it the most suitable data integration tool
  choice when selecting an AWS-native data ecosystem
- Scores well on innovation for its serverless data integration service
- Rich set of metadata capabilities such as cataloging, search, data quality checks, PII detection and data access logging for auditability

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Integration

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration process

1) Motivation (use case examples)

New business report request

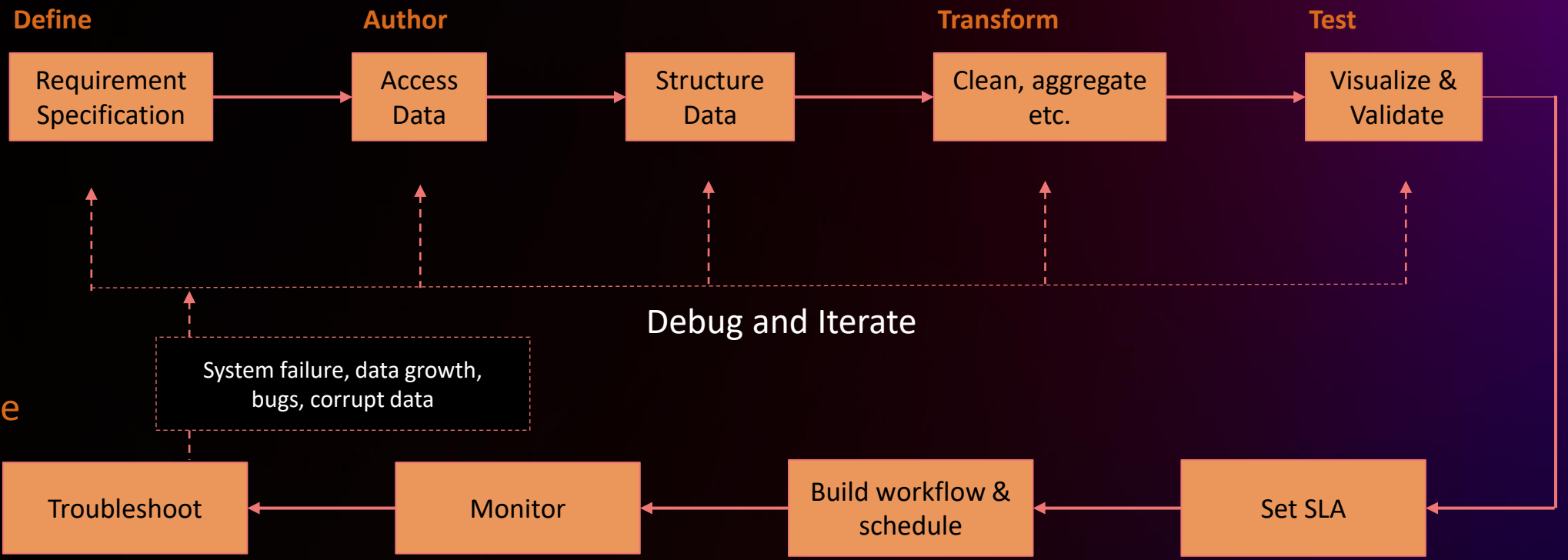
Or

Data pipeline between teams

Or

Research using data from several sources for Machine learning

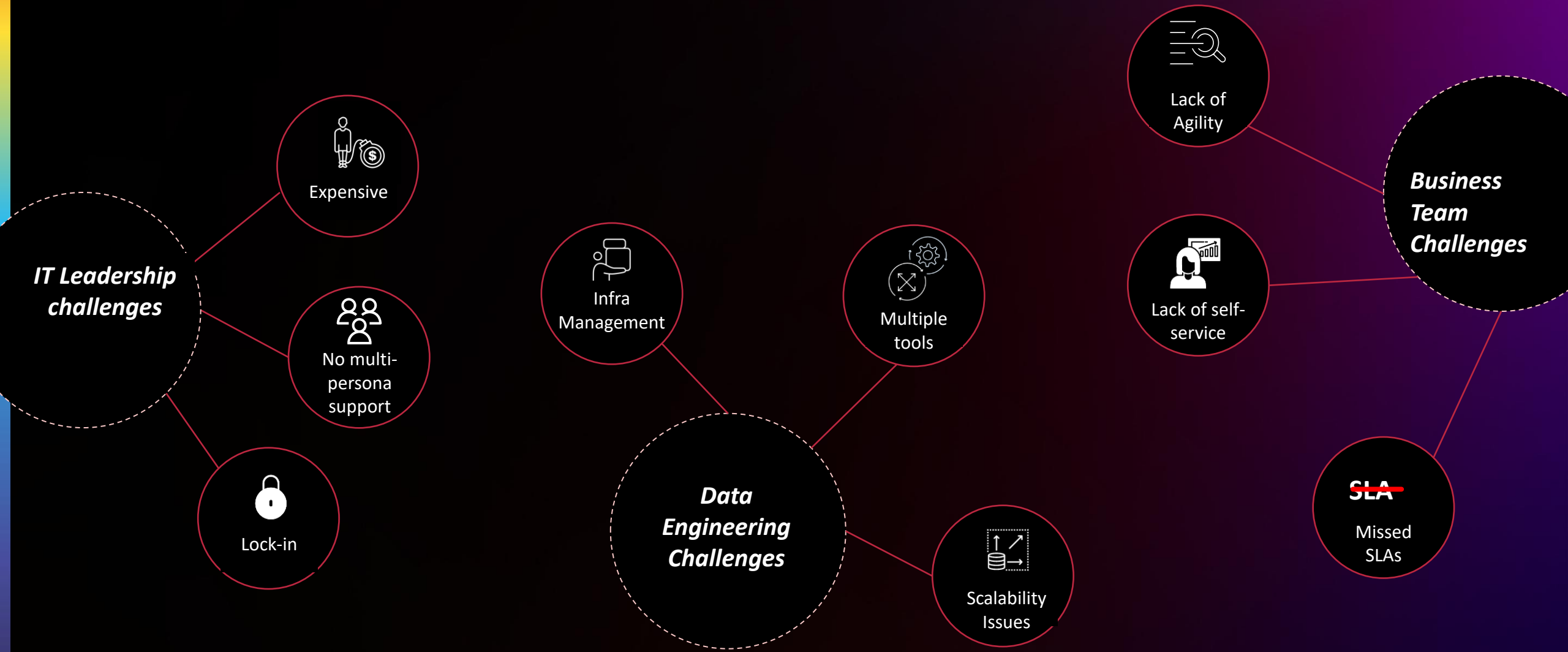
2) Build



3) Operationalize

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customers are facing challenges



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data challenges



Growing exponentially



From new sources



Increasingly diverse



Analyzed by many applications

Limited by scale and cost as data volumes grow

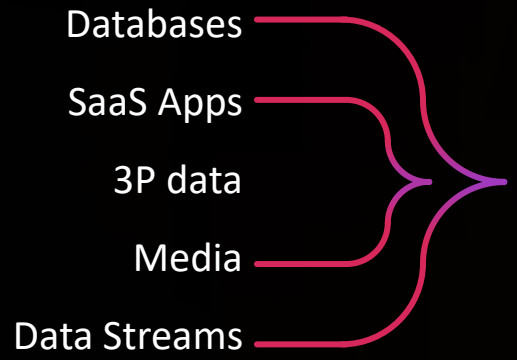
It take months to introduce new data source

Transformation requires more processing power than companies have

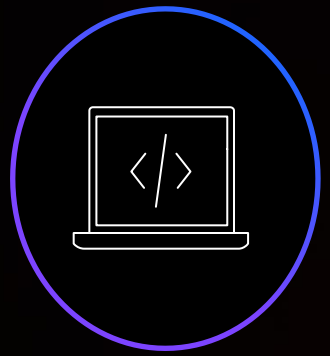
Challenges to manage security and reliability when used by multiple applications

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration and ETL is HARD



DATA SOURCES



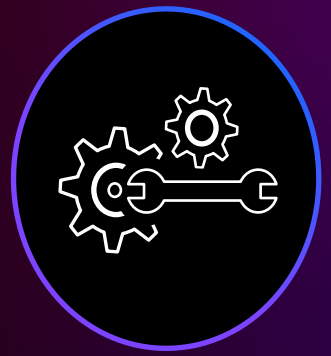
Complex application code



Cumbersome retry logic



Pipeline management



Need specialized ETL skills

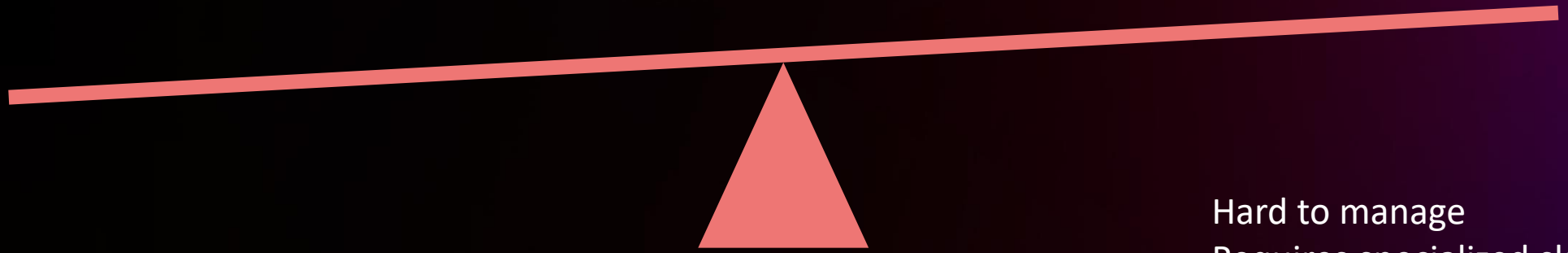
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What are the options

Traditional ETL

Limited scalability
Idle time
High licensing cost
Lock in

DIY big data platform



Hard to manage
Requires specialized skills
Dedicated admin
Pay whether you use it or not

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue





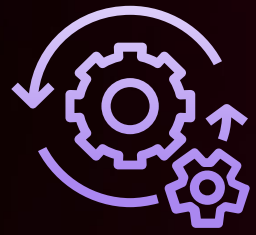
Focus on data

Low maintenance serverless solution



Scale on demand

Avoid licensing cost and infrastructure idle time



Powerful open source engines

No lock in, support from wide innovative eco system



All in one

Support all your users personas and workloads

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue overview

SERVERLESS DATA INTEGRATION SERVICE

Connectors



Data warehouse



Data lakes



Marketplace



NoSQL



Streams

Author



Visual



Notebook



Built-in transformations



IDE

Operationalize



Workflow



Monitoring



Schedule

Data Management



Data Catalog



Data quality



Sensitive data detection

Engines



Serverless infrastructure



Choice of data integration engines



Compliance and security



Stream data processing

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Engines



Serverless infrastructure



Choice of data integration engines



Compliance and security

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue data integration engines



Serverless infrastructure
No infrastructure to manage



Fast startup
5-10 seconds startup-time



Cost-effective
On demand per second billing

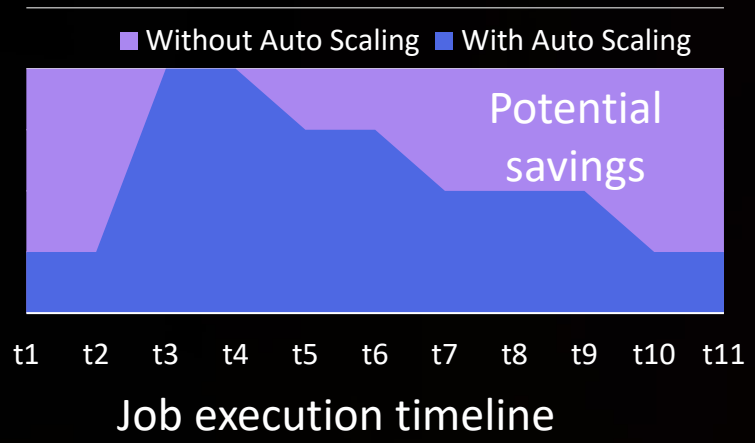
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Cost Effectiveness

AWS GLUE JOB RUNS AT LOWER COST

Auto scaling

Automatically resize compute for lower cost



Job execution timeline

Reduce cost by 20-40%

Simplify capacity planning

Streaming

Small worker types to reduce real time streaming cost

Low volume data processing

G 0.25X

(2 vCPU, 4 GB)
$0.11 per hour

Change data capture streams to implement near real time data warehouse

Flex

spare capacity execution



Up to **34% cost savings**

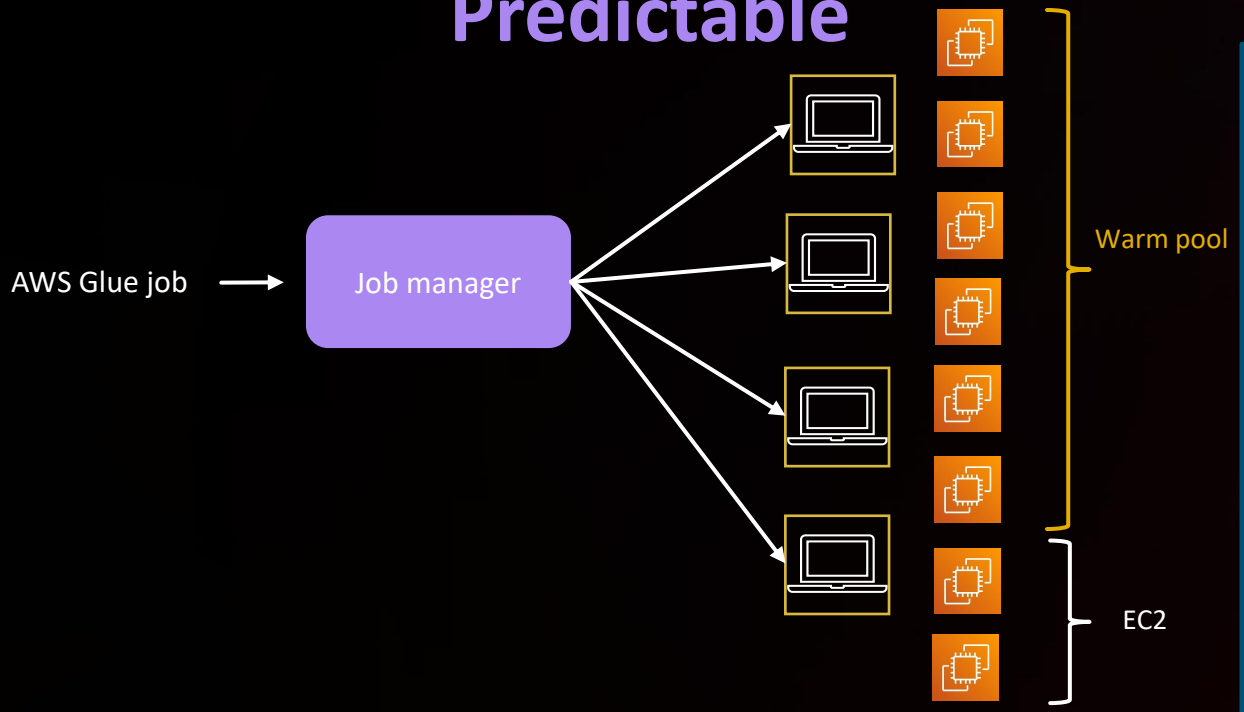
Cost effective for **one-time data-load** workloads

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue scale and reliability

FAST, PREDICTABLE, COST-EFFECTIVE

Predictable



Instantly scale as much as your job needs. Have a 300+ compute nodes in seconds

Reliable

Industry leading SLA 99.9+% uptime

Rapid and dynamic Right-sizing of your resource

Batching, micro-batching, and streaming workloads at scale

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue with Apache Spark

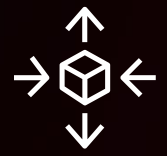
Apache Spark



Support multiple workloads in one application



Accelerate the speed of analytics and processing



Highly fault tolerant

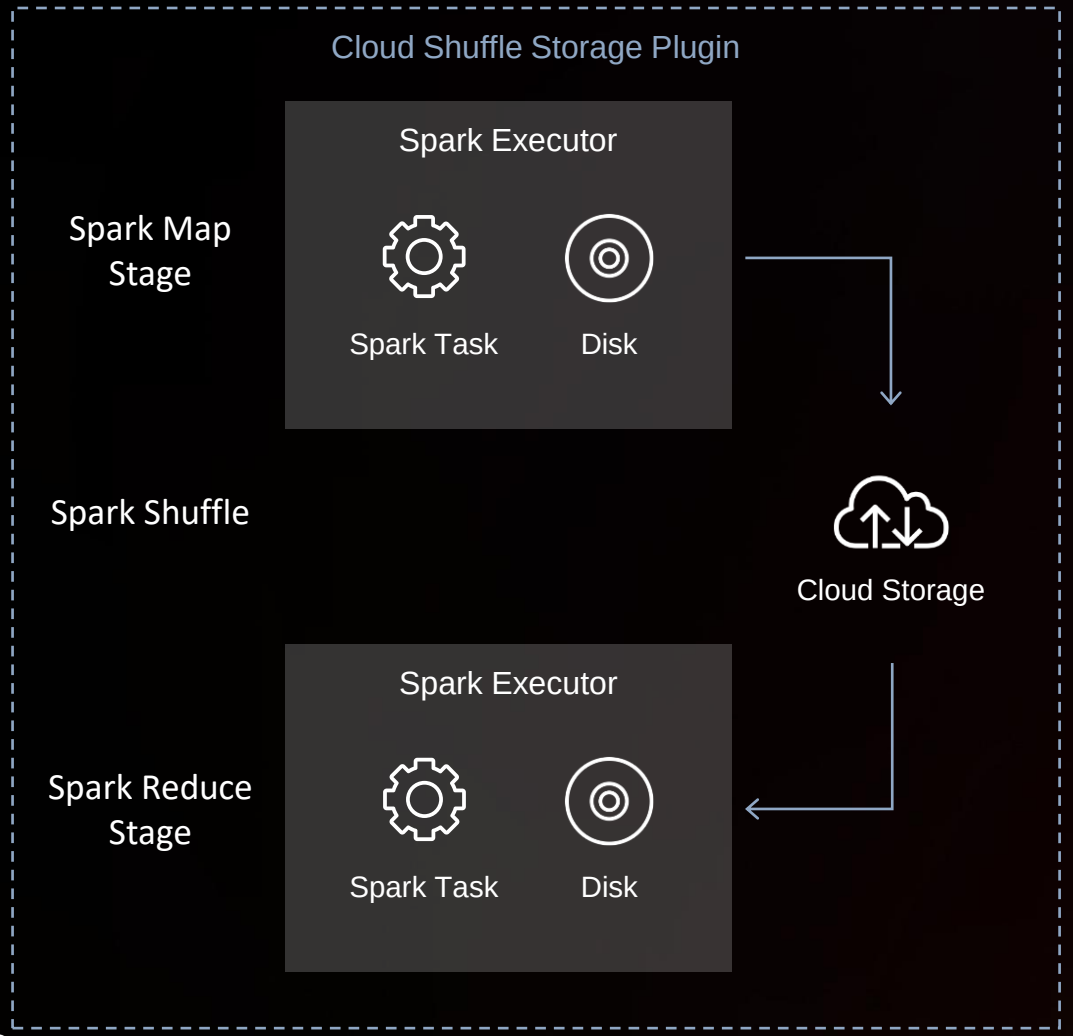


Developer friendly, supporting multiple programming languages

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

New

Announcing:
Cloud Shuffle Storage Plugin improves fault tolerance



- Shuffles in Spark occur when data needs to be redistributed across the cluster
- Shuffles are constrained by local disk capacity
- Cloud shuffle reduces Spark shuffle failures
- Native support on AWS Glue 3.0+ with Amazon S3
- Open binaries under the Apache 2.0 license

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue 3.0

FAST, PREDICTABLE, COST-EFFECTIVE

- Performance-optimized Apache Spark 3.1 runtime for batch and stream processing
- Optimized for fast performance and predictability



Now we want to give you more options to scale, move, and integrate your data

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Announcing Glue 4.0

FAST, PREDICTABLE, AND COST-EFFECTIVE

New

Upgrades Glue engine

Apache Spark — Apache Spark 3.3.0

python™ — Python 3.10

Scala — Scala 2.12

Adds more options for scaling, storing, and running your jobs



Performance optimized Spark 3.3



Distributed Pandas API on Spark - Improved data processing

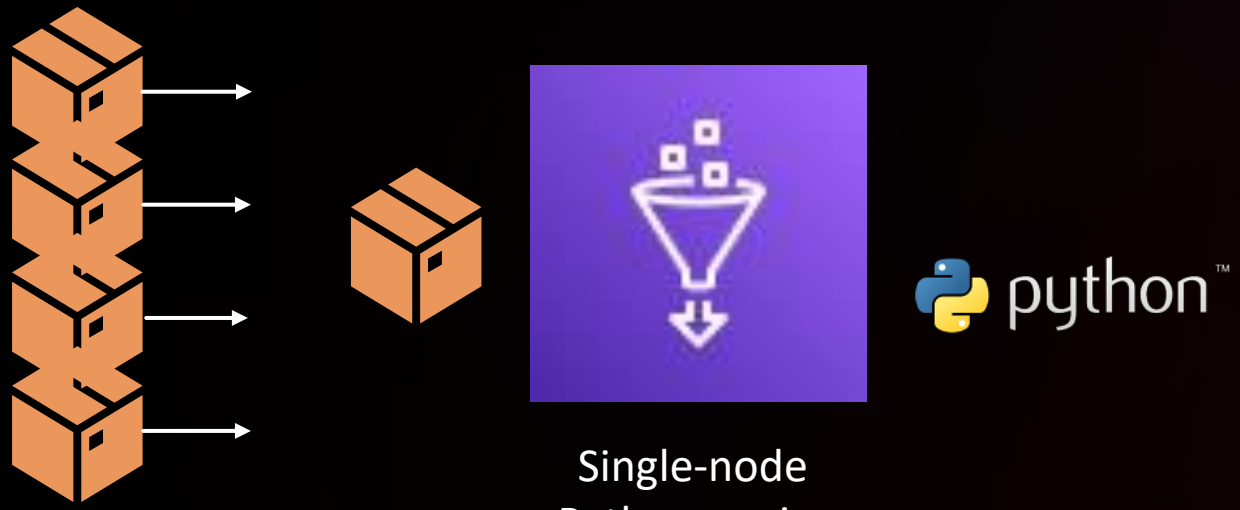


for Amazon Redshift integration Apache Spark – 10x faster in TPC-DS at 3TB scale

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration engine: single node Python



- Glue offers a single node serverless Python engine, for small dataset or latency agnostic workloads
- Glue also offers built-in support for popular data integration libraries
- Customers can bring their own Python libraries

Developers love python

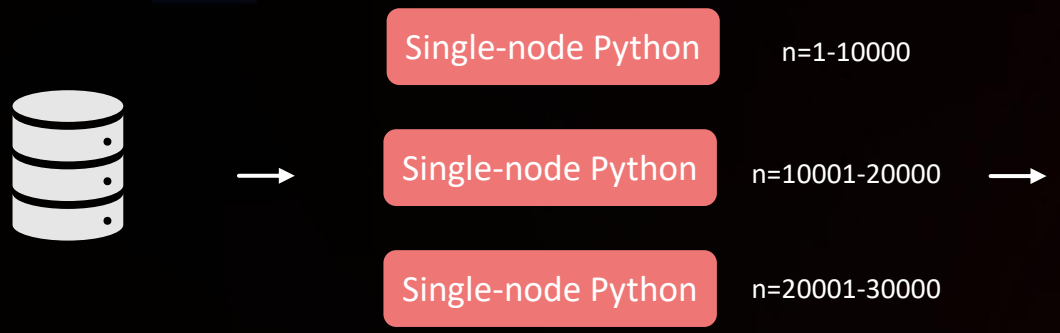
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenge: processing big data workload

1 Process data in chunks



2 Manually distribute compute

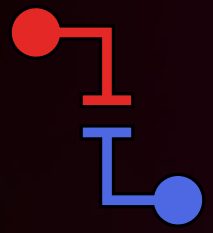




Slow processing time



Operational complexity



Error prone data processing

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Build scalable application with Ray

Python function

```
def some_function(x):
        # does something
```

Single node function

Ray task

```
# define the task

@ray.remote
def some_function(x):
        # does something


# launch the tasks
future = some_function.remote(5)


 # get the results
 ray.get(future)
```

With few lines of code, create a highly distributed tasks

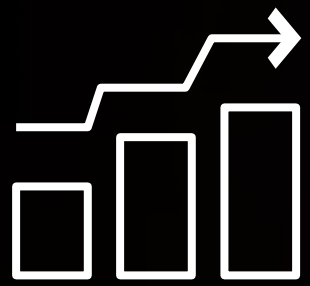
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Announcing AWS Glue with Ray

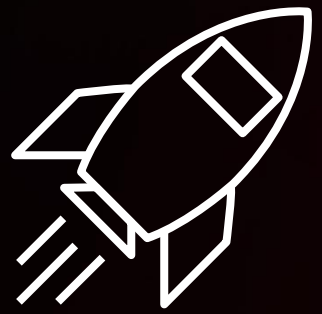
Preview

SERVERLESS DATA INTEGRATION IN DISTRIBUTED PYTHON

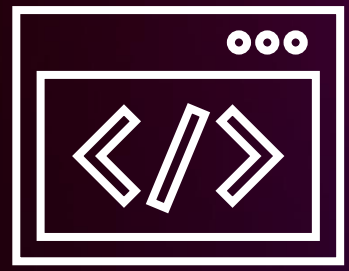
"Ray is an open-source unified compute framework that makes it easy to scale AI and Python workloads"*



Scalable to hundreds of nodes



Superfast start and scale up/down time



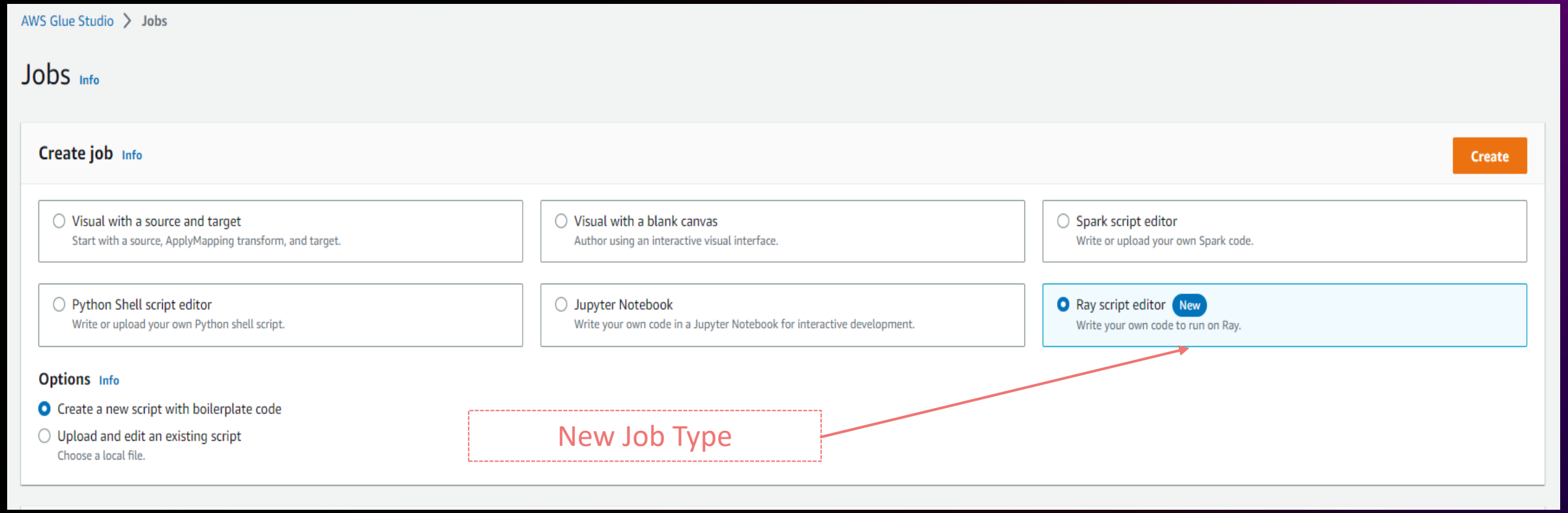
Familiar Python Primitives, built-in libraries

*To learn more - visit Ray.io

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How to create a Ray job: Glue Studio



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary: Engines

- Focus on data and leverage powerful serverless engines
- Choice of data integration engines: Spark, Python Shell, Ray
- Cost effective: Flex and Auto scaling
- Developer friendly
- few seconds to cluster with hundreds of nodes
- Supports batch, micro-batching, event and streaming workloads

ANT343: AWS Glue with Ray session
12/1 Thursday 4:15 MGM – Grand 121

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Connectors



Stream



Database



Marketplace



NoSQL



Data lake



Data warehouse

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Connect with AWS Glue

80+ connectors across a wide range of sources and targets

Built-in Connectors

Out of box connectors **managed by** AWS Glue

Marketplace Connectors

Broad range of subscription based, **3rd party options**

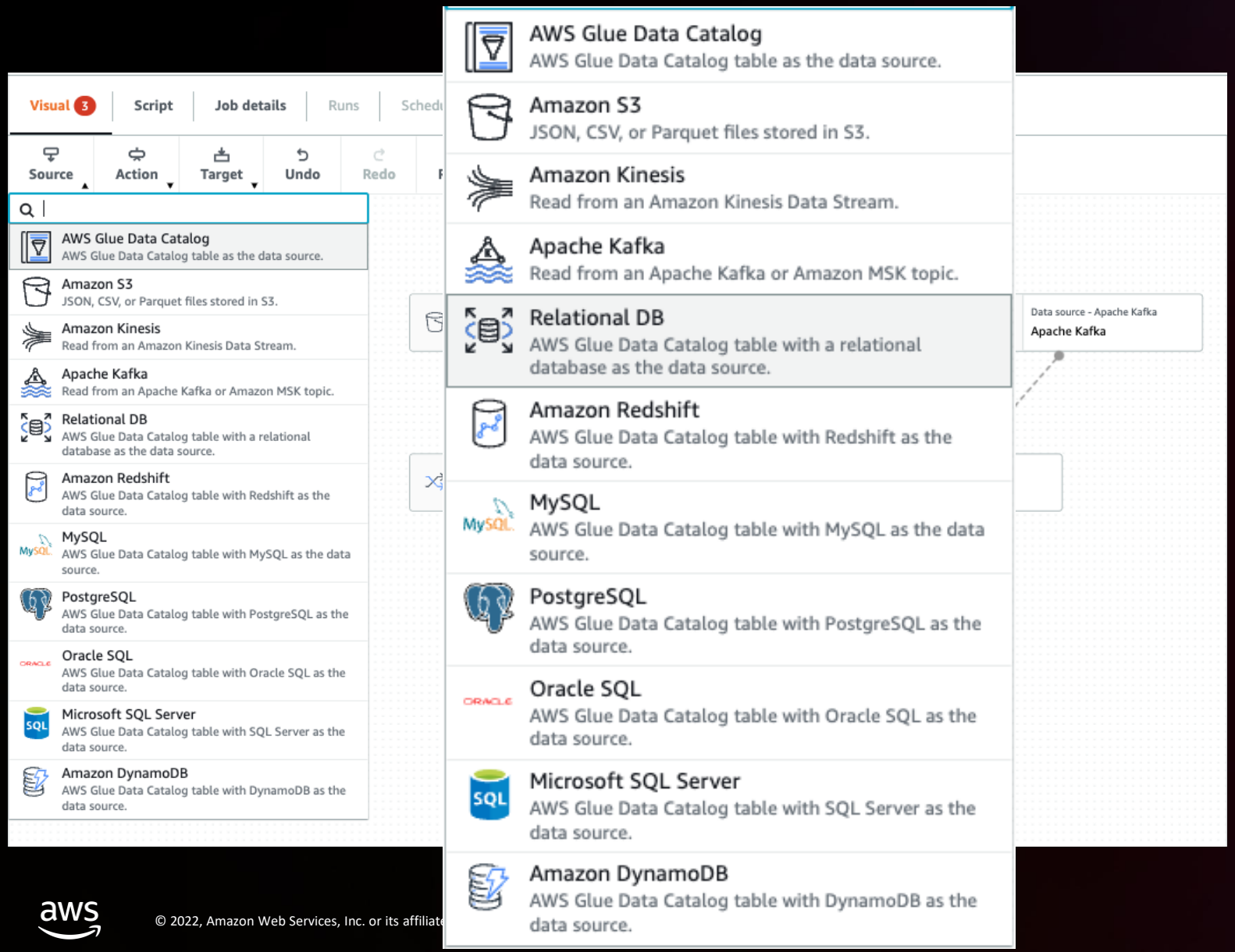
Custom Connectors

Flexibility to build domain specific data access



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scalable and simple to use



Built-in source and target

Efficient data transfer with pushdown, partitioning, and incremental loads

Optimized for native AWS services

Standard interface to manage datatypes

© 2022, Amazon Web Services, Inc. or its affiliate

Amazon AppFlow supports over 50+ applications

MARKETING CONNECTORS

- Facebook Ads*
- Facebook Page Insights
- Google Ads*
- Google Search Console
- Instagram Ads
- LinkedIn Ads

CUSTOMER SERVICE AND ENGAGEMENT CONNECTORS

- Mailchimp
- SendGrid
- Zendesk Sell*
- Zendesk Chat*
- Freshdesk
- Okta
- Typeform

AS WELL AS

- Stripe
- Amazon RDS for PostgreSQL
- QuickBooks Online
- Jira Cloud*
- GitHub

AND MORE!



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

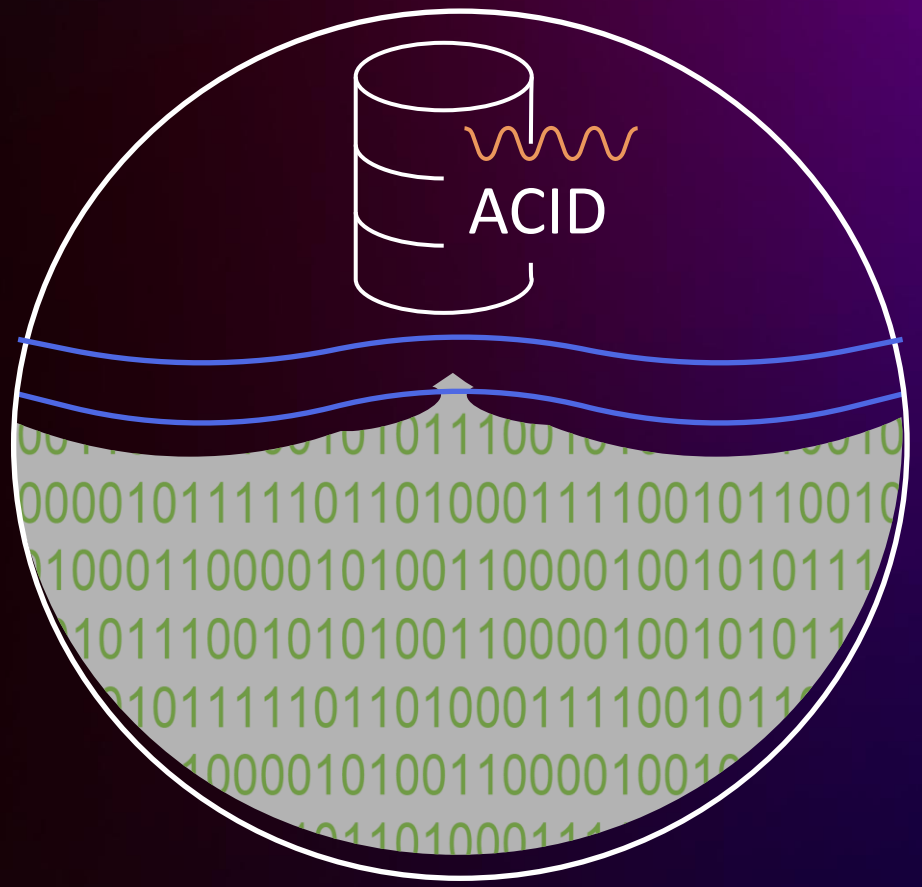
* Launched in 2022 prior to re:Invent

Keeping your data lake up to date

How to manage a data lake in a transactionally consistent manner?

ACID transactions	Time Travel queries
Upsert & Delete	Rollback
Streaming Ingestion	Change Data Capture (CDC)

We heard you -
More community innovation!



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

New

Announcing:
Native support for Hudi, Delta Lake and Iceberg







Simplify incremental data processing in data lakes built on Amazon S3

- Built in support to read and write (insert, update and delete)
- No setup, minimal configuration and easy to deploy
- Latest framework versions available with Glue v4.0

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Authoring Solutions



Built-in transformations



IDE



Visual



Notebook

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Studio: Interfaces for every skillset



Business Analyst
Data scientist



ETL developer
IT/Data professional



Data engineer



Data Analytics

Software Engineering



SQL



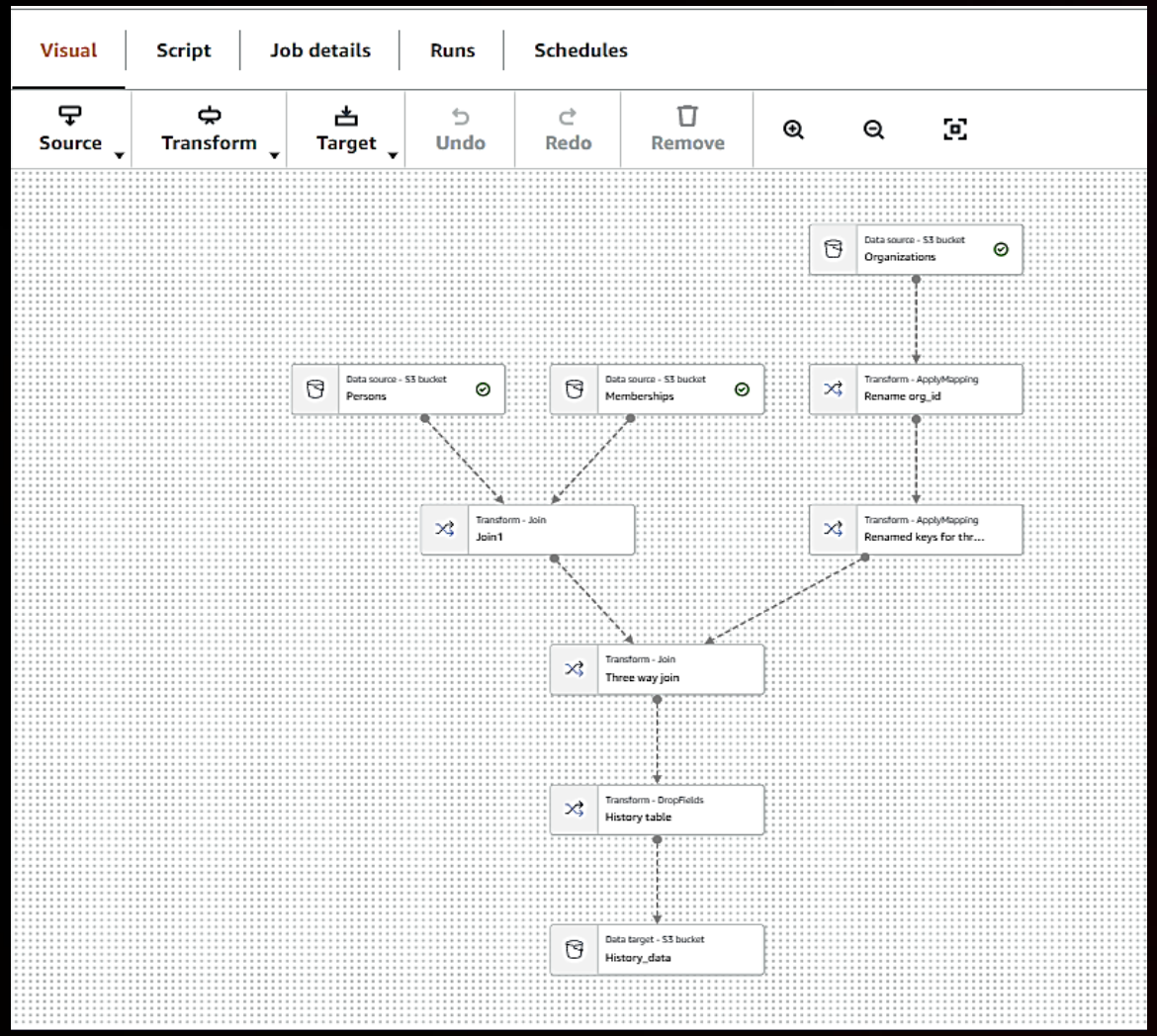
Apache **Spark**

Scala

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Studio: Visual ETL



Design large-scale data integration jobs visually

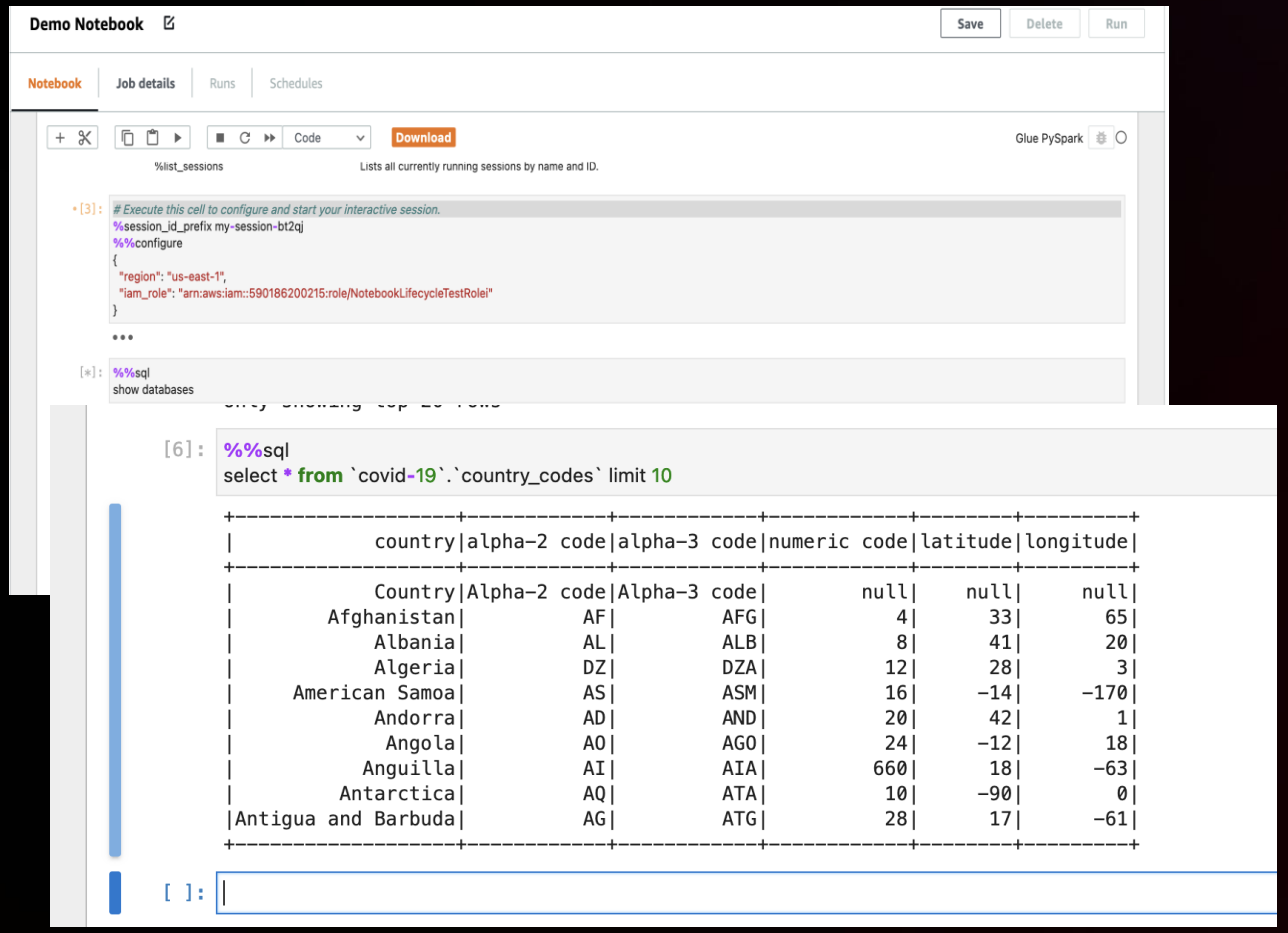
Data integration jobs without the Spark learning curve

Preview your data at each step of the process

Detects schema so no need to catalog source data

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Studio: Notebooks



Built-in interface for AWS Glue Interactive Sessions

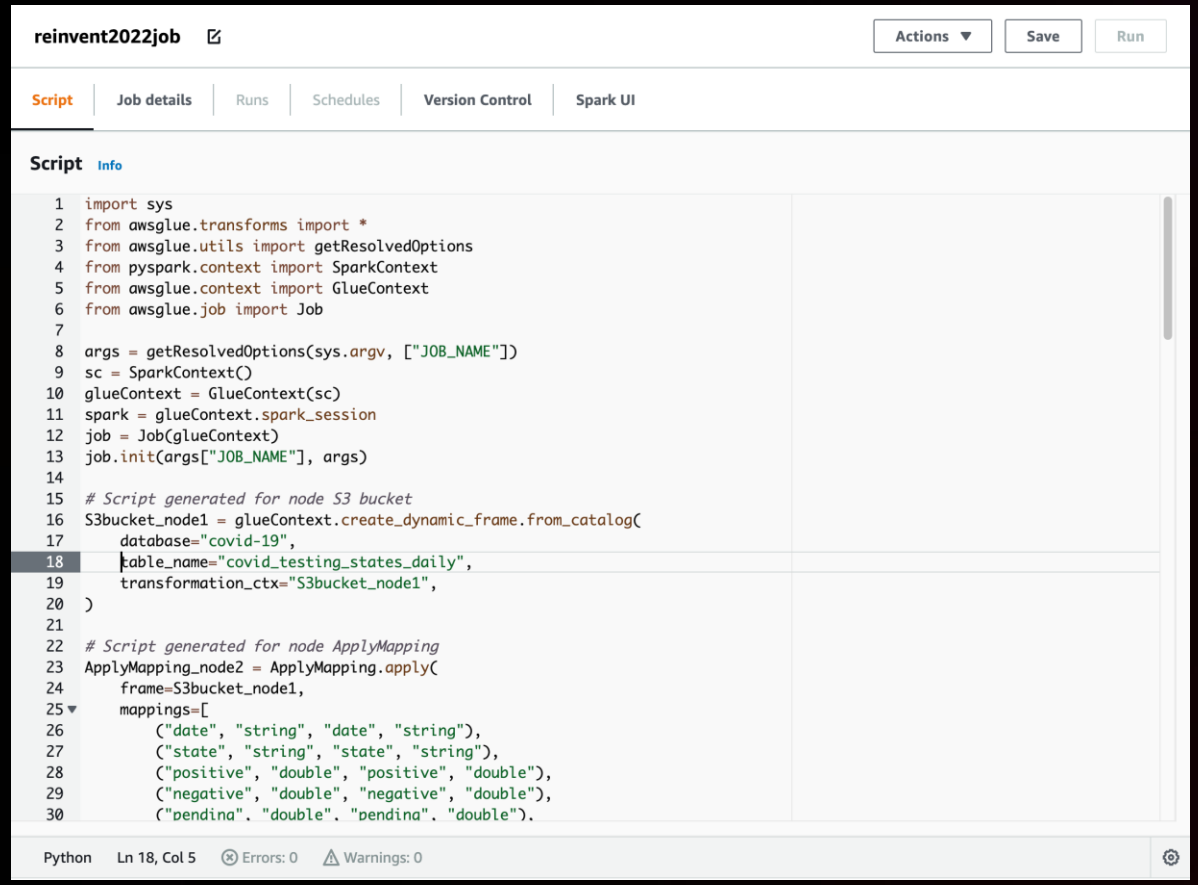
Explore data to see how it needs to be processed

Test ETL code, with immediate results

Convert notebooks into jobs with one click

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Studio: Python & Scala Code Editor



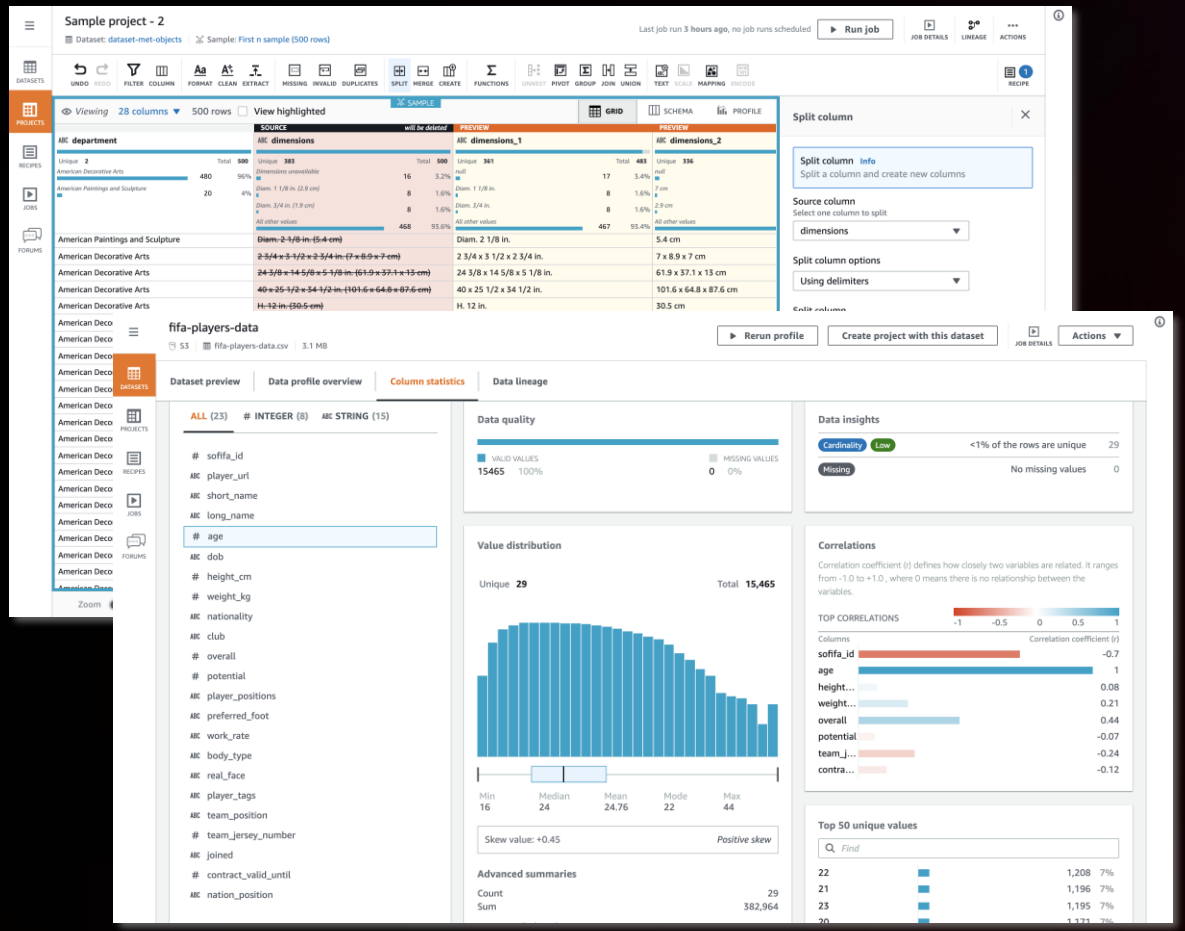
Convert existing Apache Spark code into AWS Glue jobs

Quickly create jobs from code developed locally

No-fuss code and job management

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue DataBrew



- Chose from 250+ transformations to prepare datasets
- Understand the data profile with 50+ statistics
- Share, reuse, schedule transforms across teams
- Interactive previews for fine-grained transformation

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

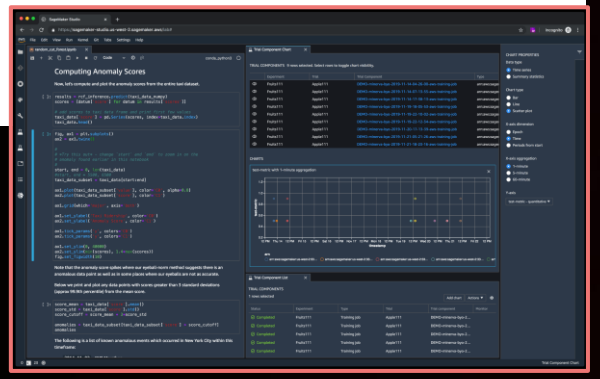
Interactive Session API

DEVELOP DATA INTEGRATION PROCESSES ON THE FLY

Choose your favorite IDE

Jupyter Notebooks & Lab
Visual Studio Code
IntelliJ | PyCharm | DataSpell

And now: Amazon SageMaker



Test, get immediate results, and adjust your code

Jupyter kernel that integrates with your favorite IDEs

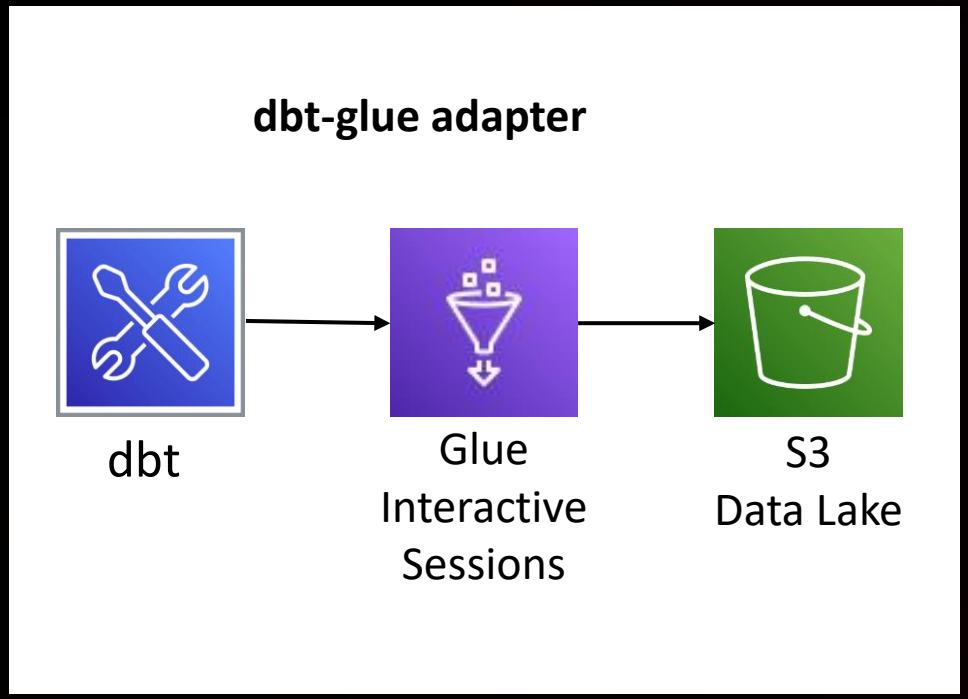
Easy to set up and starts in seconds

Amazon SageMaker

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Session API

Support 3rd party : dbt – Glue integration



dbt is a popular SQL-Native data modeling framework

dbt was primarily built for Data Warehouses, such as Amazon Redshift or Snowflake

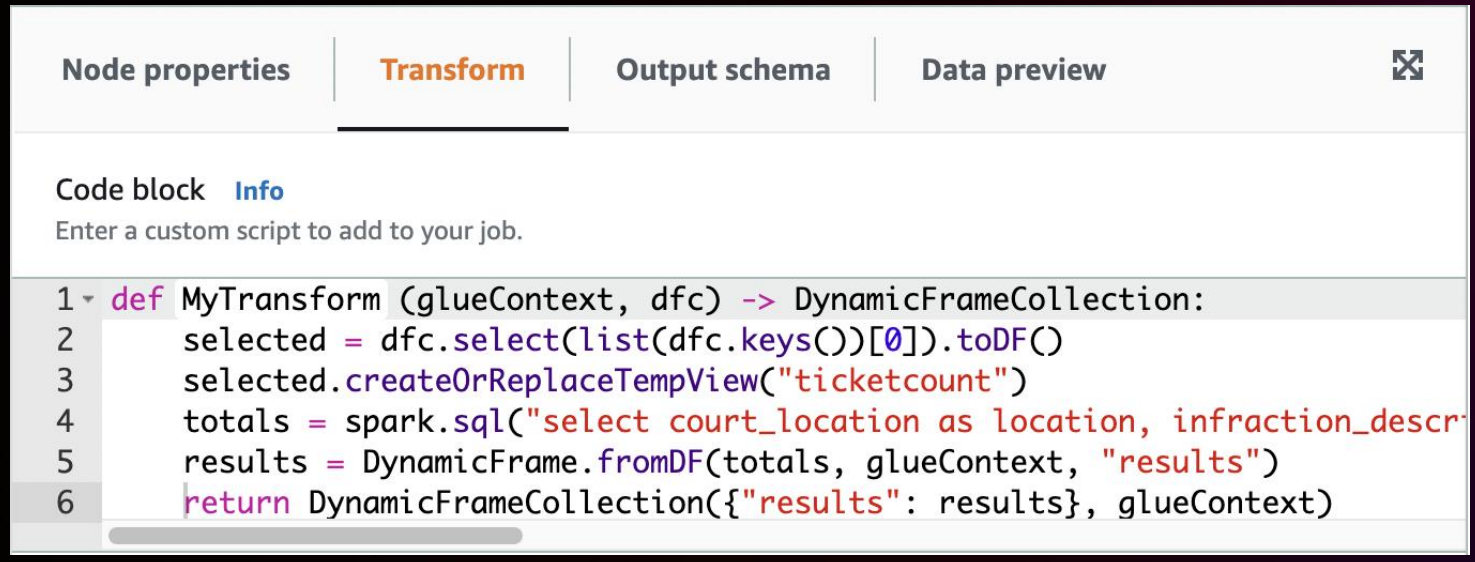
SQL-first integration with Glue Interactive Session brings all dbt features to the AWS data lake

Serverless and Open-source

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Transformation pain points

- Custom business logic requires code development
- Many transforms are built in, but no way to add more yourself
- Users can’t share custom transforms

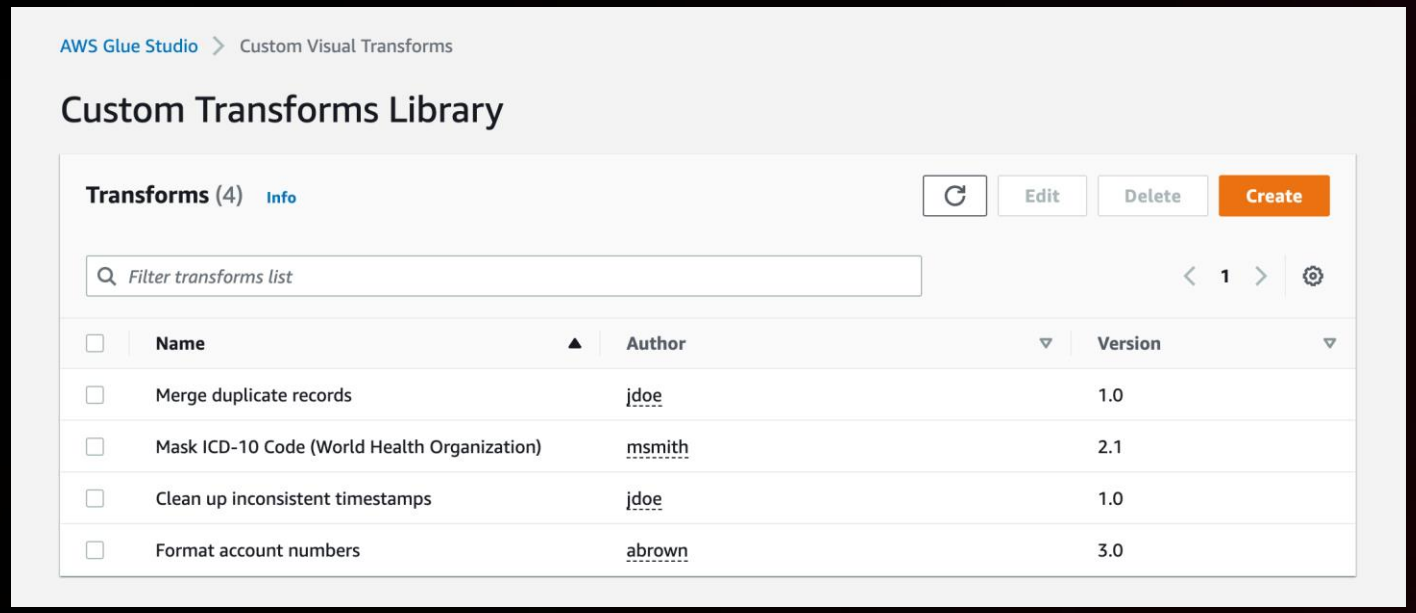


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Announcing: Custom visual transforms

New

- (Re)use custom business logic in visual ETL jobs
- Share custom transformations between teams



Benefits:

- Reduces dependence on Spark developers
- Maintain and update jobs more easily
- Custom visual transforms work in both visual and code-based jobs

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary: Authoring

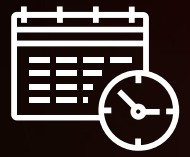
- AWS Glue job authoring supports a variety of technical skillsets
- We abstract away complexity for users who are not Spark developers
- Our code-based interfaces offer flexibility to choose the underlying engine
- Custom visual transforms help scale your team and help users build faster

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Operationalize



Workflows



Schedule



Monitor

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

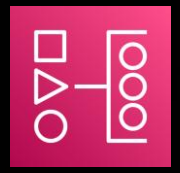
Flexible orchestration of AWS Glue jobs

Multiple options to meet your needs



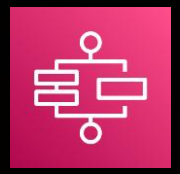
AWS Glue Workflows

Visually sequence Glue specific jobs based on triggers



Amazon Managed Workflow for Apache Airflow

Programmatically Integrate with Apache Airflow

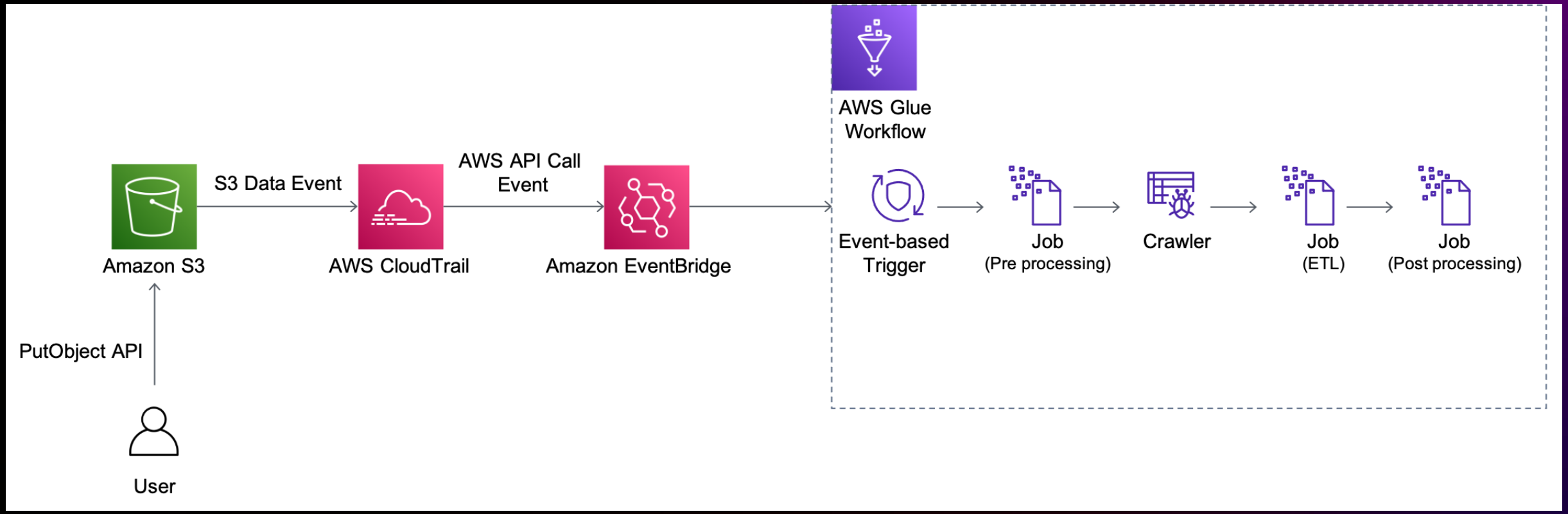


AWS Step Functions

Build workflows spanning multiple AWS services

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Event-driven workflows
Amazon EventBridge + AWS Glue



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Git integration

New

GitHub



AWS Glue



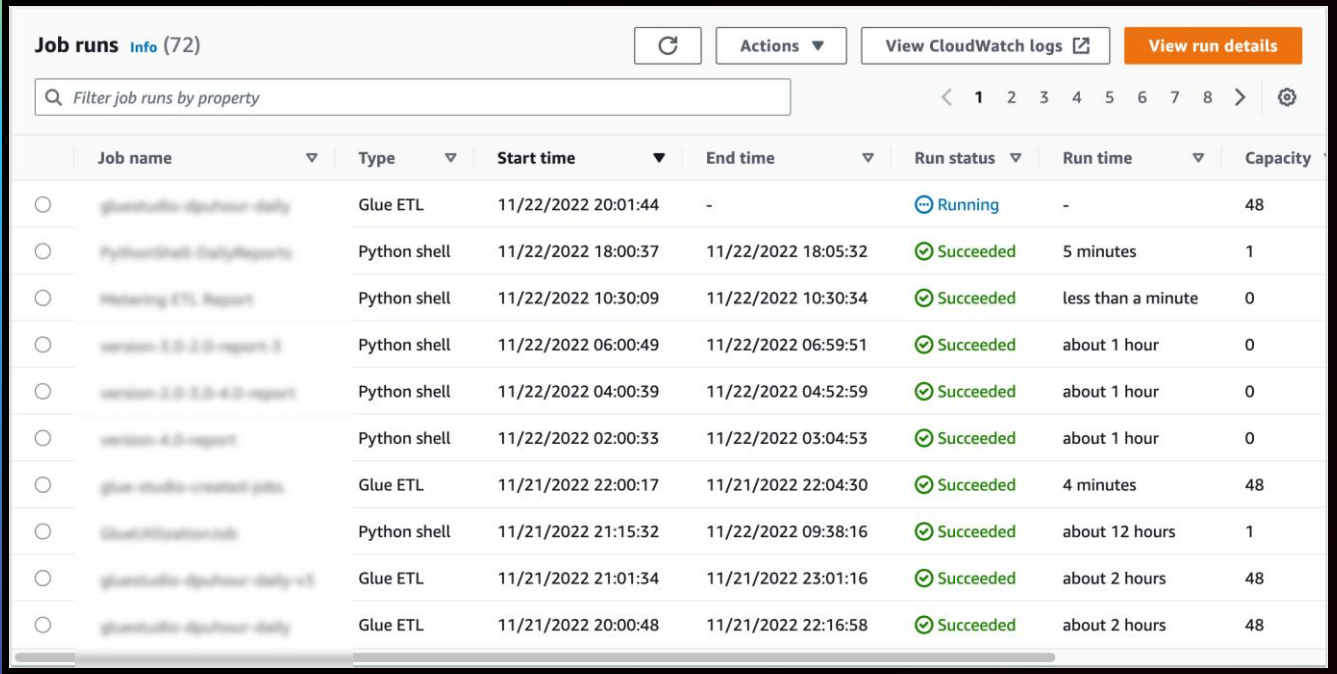
AWS CodeCommit

- Track changes to your ETL code and Visual ETL jobs
- Deploy AWS Glue jobs with your existing pipelines and DevOps tools
- Supports GitHub and AWS CodeCommit
- Download/upload AWS Glue jobs in AWS Management Console

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue job monitoring



Monitor 1000s of jobs through a single pane of glass

View debugging recommendations and logs

Track resource usage and identify expensive jobs

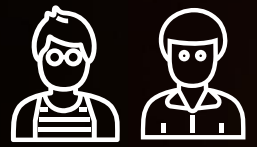
Filter by job details and time periods

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Management



Data Catalog



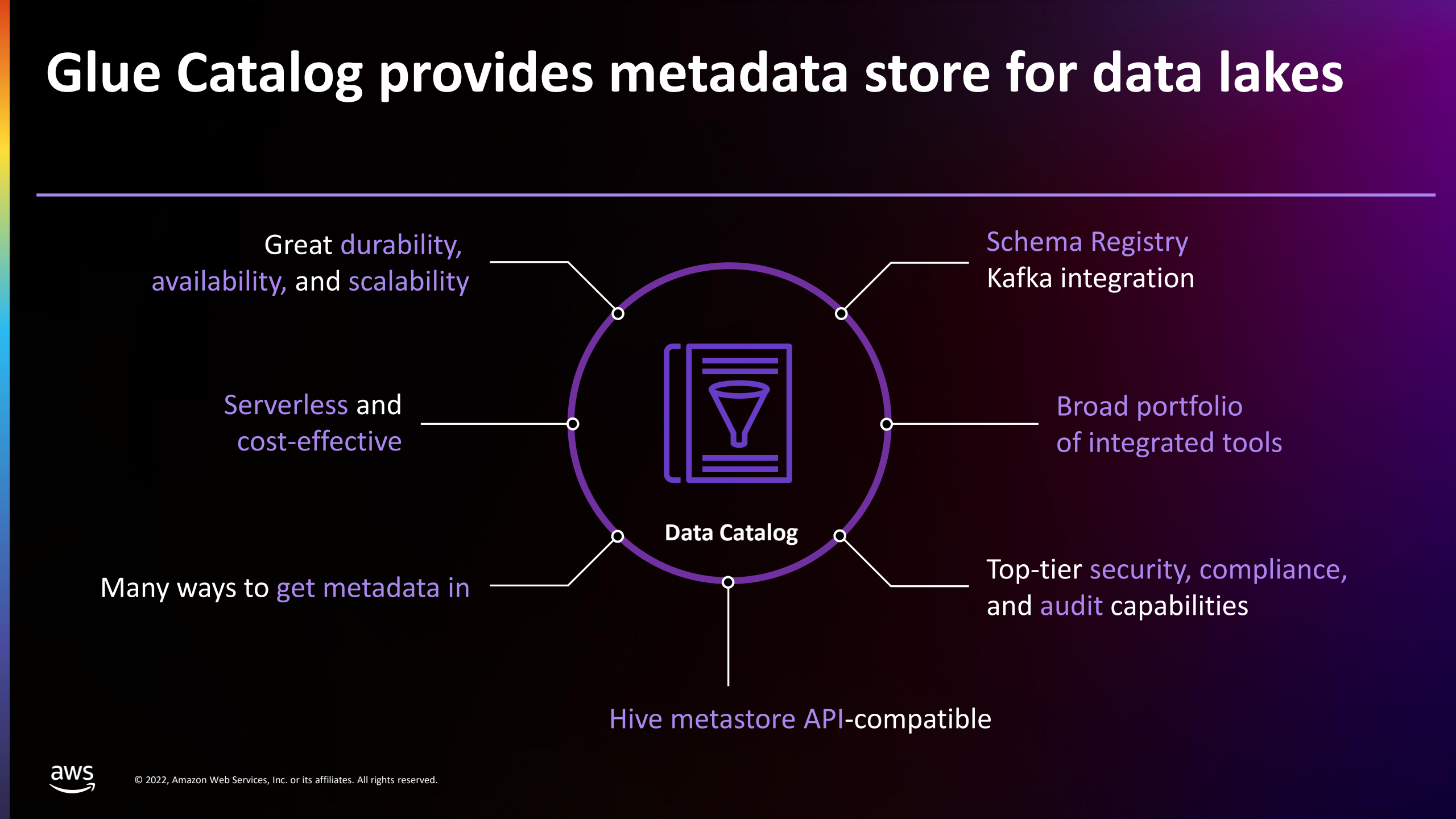
Sensitive data detection



Data quality

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

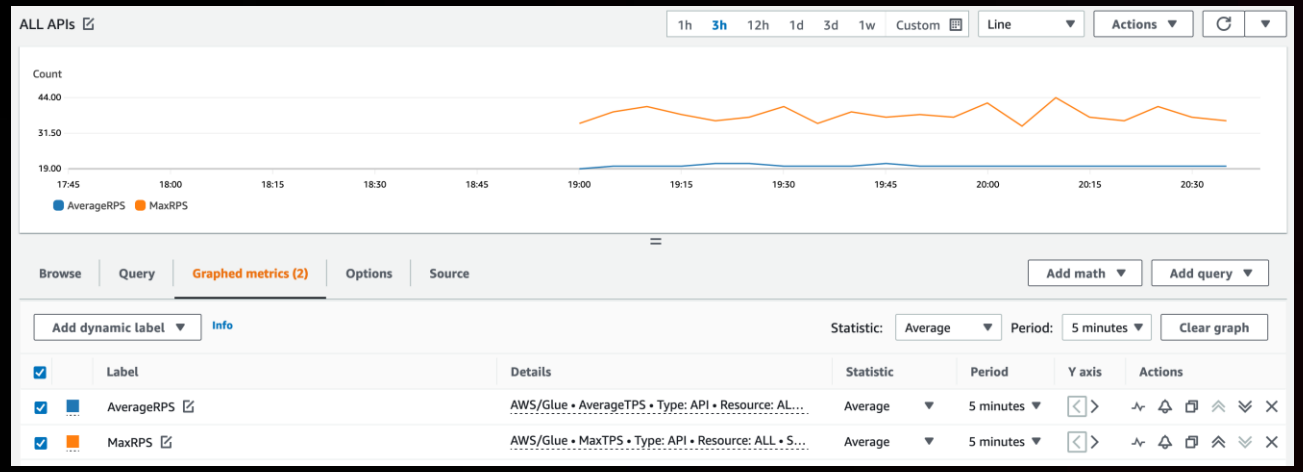
Glue Catalog provides metadata store for data lakes



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Catalog – Capabilities

New



Audit with CloudWatch Catalog Metrics

Federate Hive MetaStore and Access with Athena, RedShift and Glue

Database, Connection Tagging

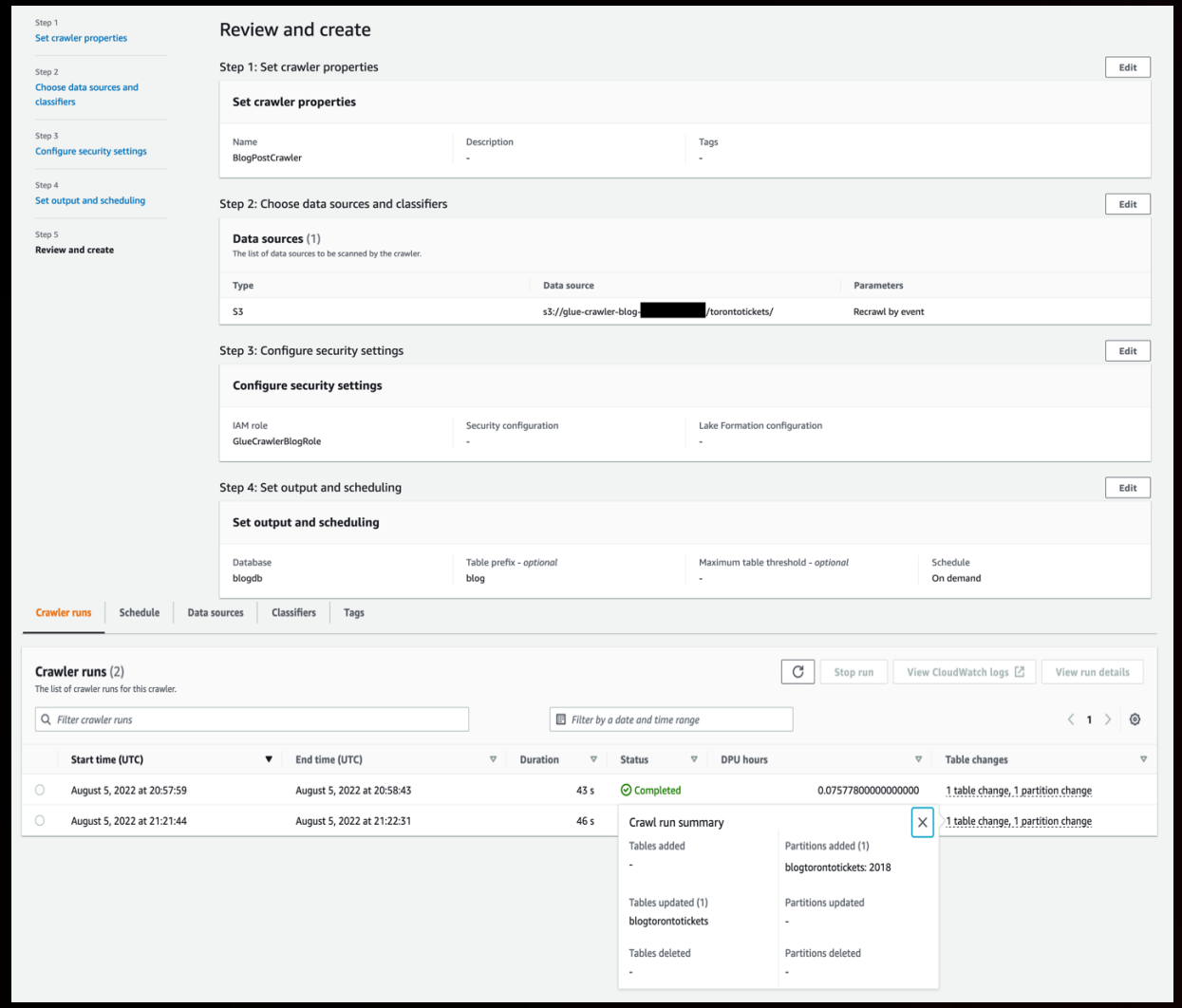
Sync Delta Lake Metadata in Catalog to enable column permissions

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Glue Data Catalog - Crawlers

New



Simplified monitoring with properties and metrics associated with past crawler runs

Cross-account crawling with Lake Formation

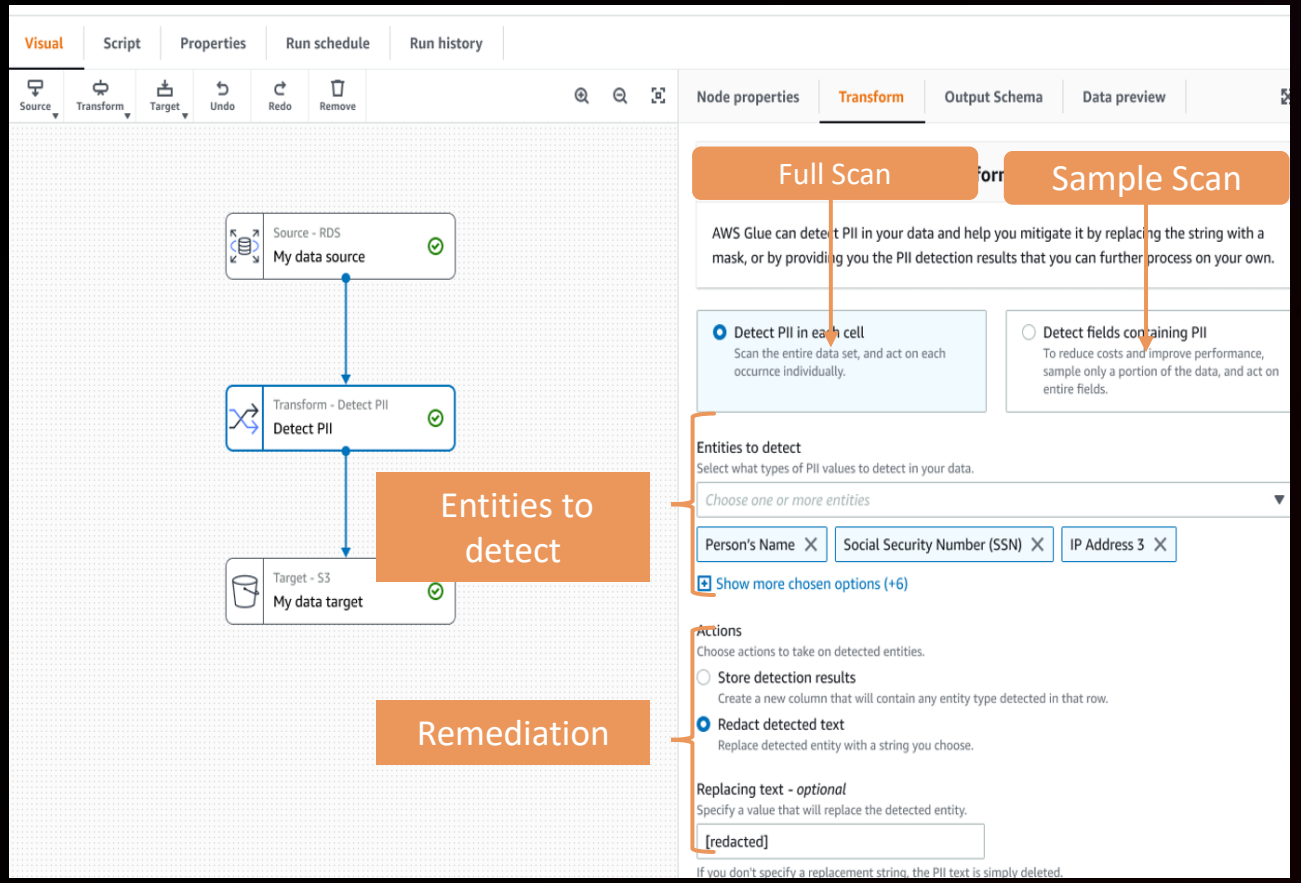
Faster Incremental crawling for existing Catalog Tables

Table and column stats collection

Catalog SnowFlake, MongoDB Atlas

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sensitive data detection



Identify PII and sensitive data that goes through your pipelines

Simple detection: Out of the box entities

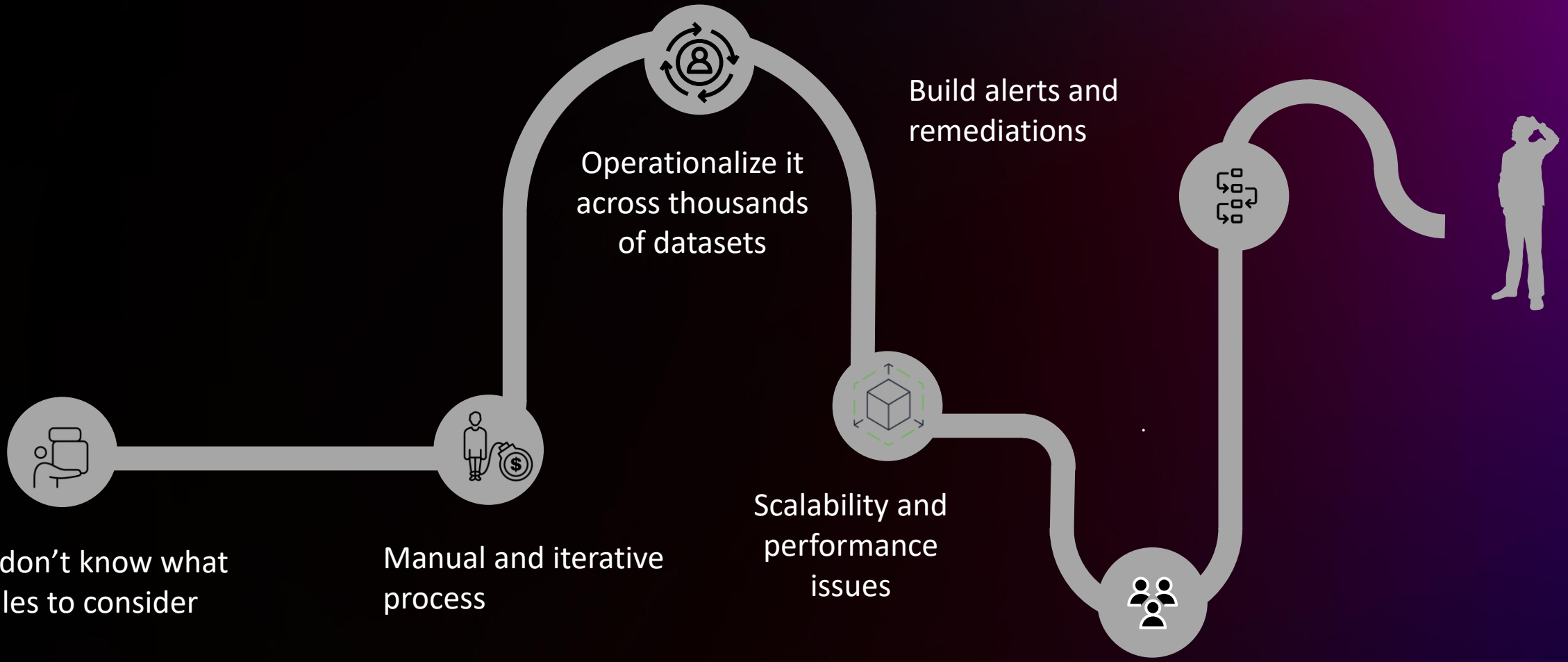
Remediate: ability to store, redact, of mask sensitive data

Detect across TB of data in scalable way

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managing quality is hard

Preview



You don’t know what rules to consider

Manual and iterative process

Operationalize it across thousands of datasets

Scalability and performance issues

Build alerts and remediations

Multiple personas, infrastructure management

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Preview

Announcement:

AWS Glue Data Quality

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Data Quality Features and Benefits

Preview



Serverless, scalable and performant



Data Quality Recommendations



Out-of-the-box data quality rules and actions



Data Quality Definition Language



Data Quality for data-at-rest and data-in-transit



Supports multiple persona

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works: Data Quality on AWS Glue Data Catalog

Preview



Data Steward

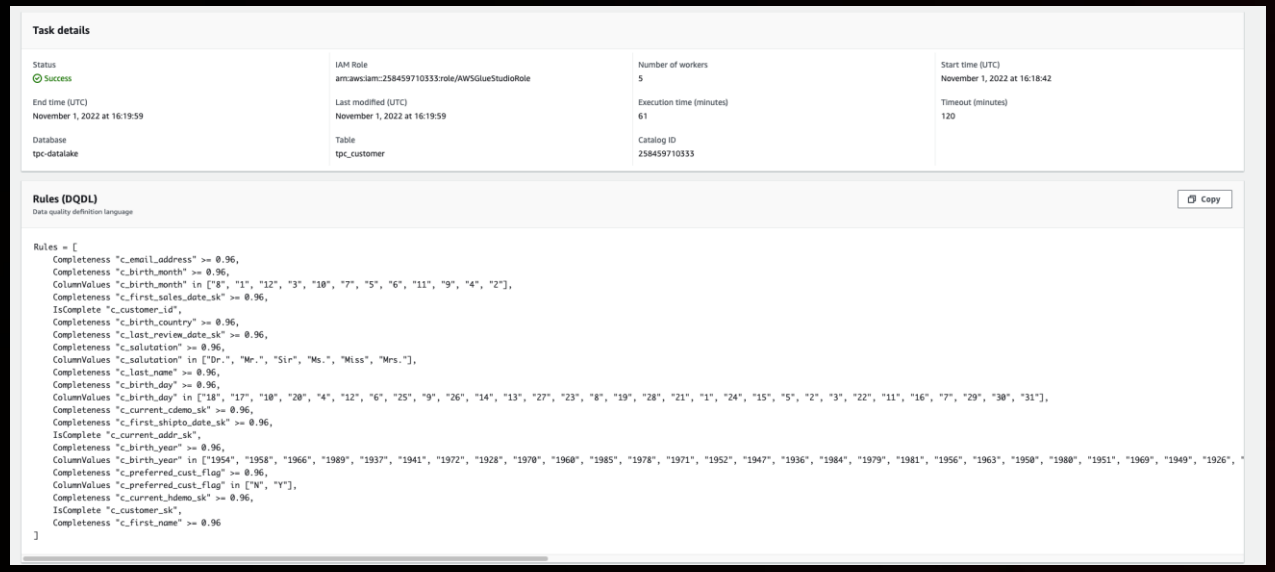
Selects a dataset in AWS Glue Data Catalog

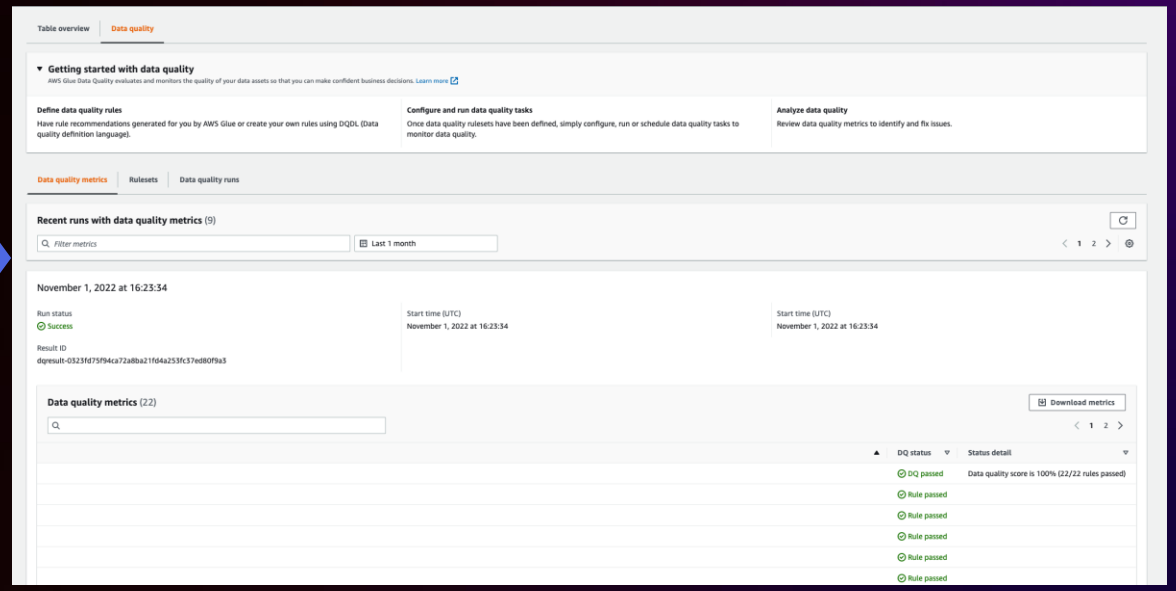
AWS Glue Data Quality analyzes data and recommends rules

Data Steward refines the rules to create finalized rules

AWS Glue Data Quality Evaluates rules

Data Steward reviews results and alerts, takes appropriate action

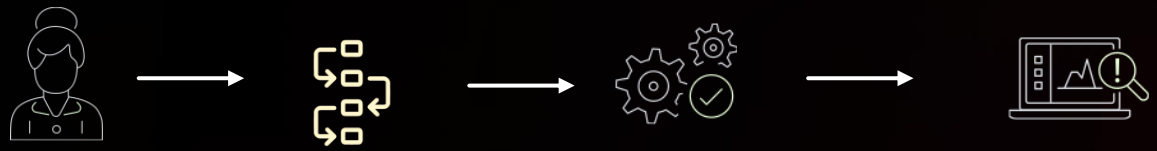




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works: Data Quality on AWS Glue ETL Pipelines

Preview

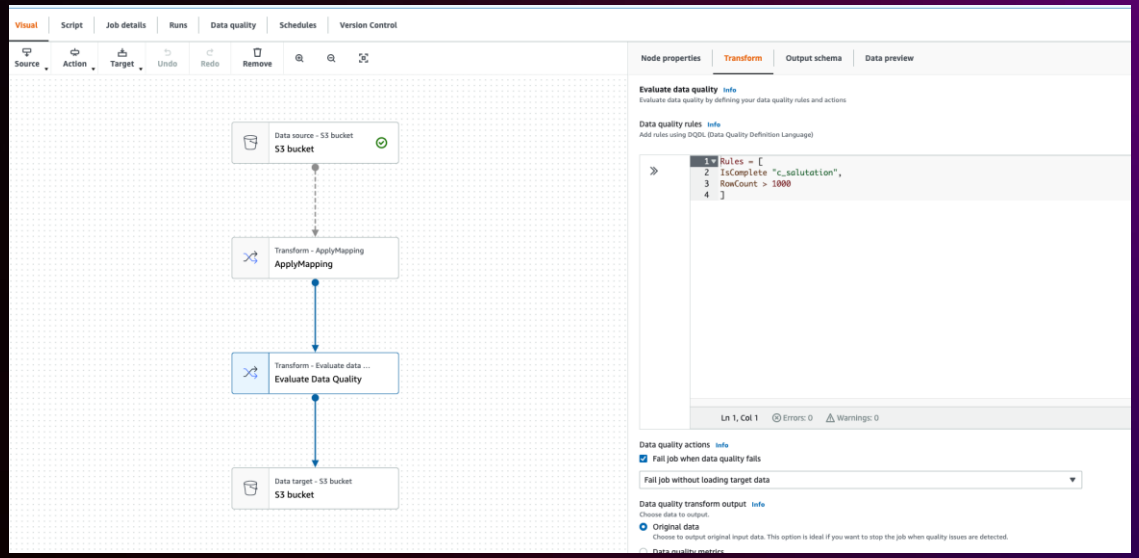


Data Engineer

Selects an ETL Job and adds Data Quality rules

AWS Glue Data Quality Evaluates rules

Engineer reviews results and alerts, take appropriate action



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Data Quality Demo

Shiv Narayanan

Product Manager, AWS Glue

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



"We are excited about AWS Glue Data Quality, which will automatically identify, analyze, and take act on data quality issues in a matter of minutes. This will help us make informed, timely, and accurate decisions and save countless hours in manually identifying and fixing all data issues."

Ashok Srinivas, Director of ML Engineering & Sarang Bapat, Director of Data Engineering

ANT222: AWS Glue Data Quality
11/30 Wednesday 6:15
Level 3, Chairmans 368, MGM Grand

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Why Data Integration and ETL modernization

Challenges with data integrations

AWS Glue new innovations and announcements

Why Itaú choose AWS Glue for data mesh architecture

Conclusions

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Itaú Unibanco

A universal bank | 98 years of history | Latin America's largest financial institution



Latin America

- Argentina
- Brazil
- Chile
- Colombia
- Mexico
- Panama ①
- Paraguay
- Peru
- Uruguay ①

Rest of the world

- Bahamas ①③
- Cayman ①②③
- England ①②
- France ①
- Germany ①
- Portugal ①
- USA ①②③
- Spain ①
- Switzerland ③

Key

- Multiple Bank
- Other operations

① Corporate & Investment Banking ② Asset Management ③ Private Banking

Brazil's most valuable brand
+ US$ 7 billion

+ 65 MM customers

+ 20 MM digital customers

~ 100 K employees

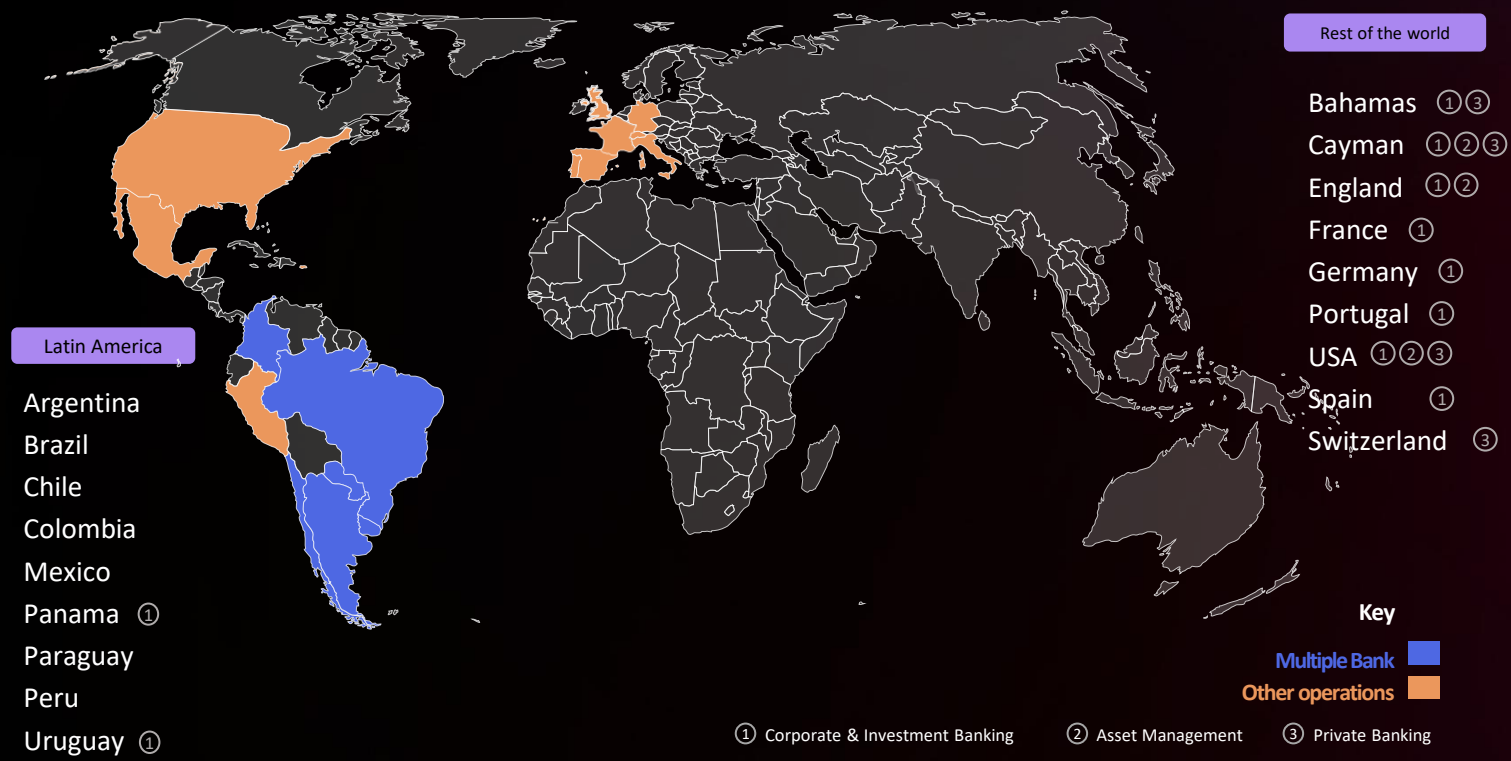
4 K business services

Largest bank in market value; (2) 2022 Brand Finance Brazil Ranking; (3) In 2Q22
https://www.itau.com.br/download-file/v2/d/42787847-4cf6-4461-94a5-40ed237dca33/ff1ae935-d478-6f69-0660-1f2f526badd9?origin=1
Dollar exchange rate (Nov. 25, 2021)

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Itaú Unibanco

A universal bank | 98 years of history | Latin America's largest financial institution



Brazil's most valuable brand
+ US$ 7 billion

+ 65 MM customers

+ 20 MM digital customers

~ 100 k employees

4 k business services

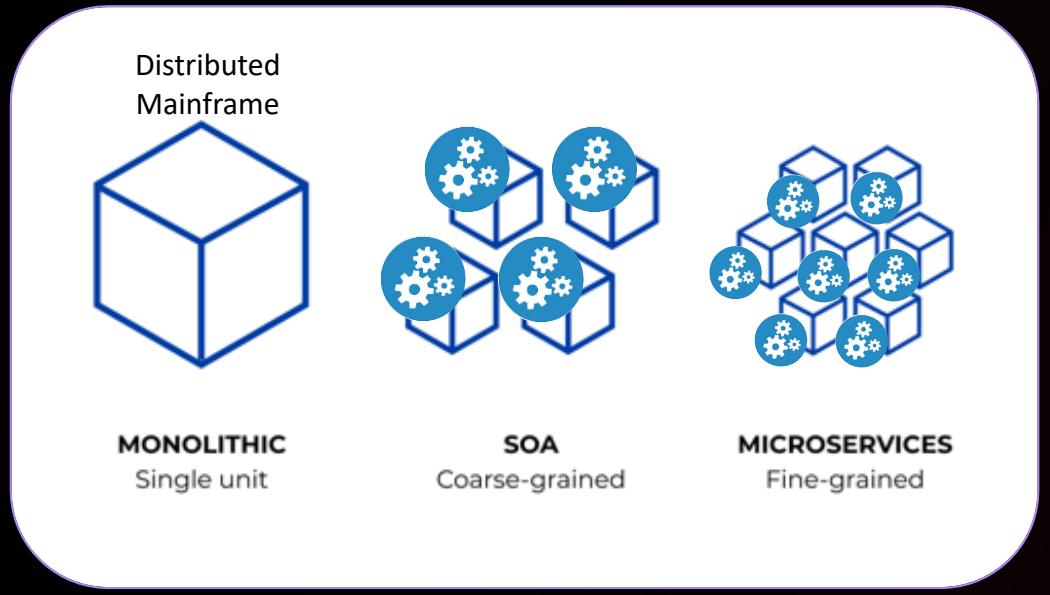
1.Customer-centered **2.**People and Processes **3.**Data and Technology

Largest bank in market value; (2) 2022 Brand Finance Brazil Ranking; (3) In 2Q22
https://www.itau.com.br/download-file/v2/d/42787847-4cf6-4461-94a5-40ed237dca33/ff1ae935-d478-6f69-0660-1f2f526badd9?origin=1
Dollar exchange rate (Nov. 25, 2021)

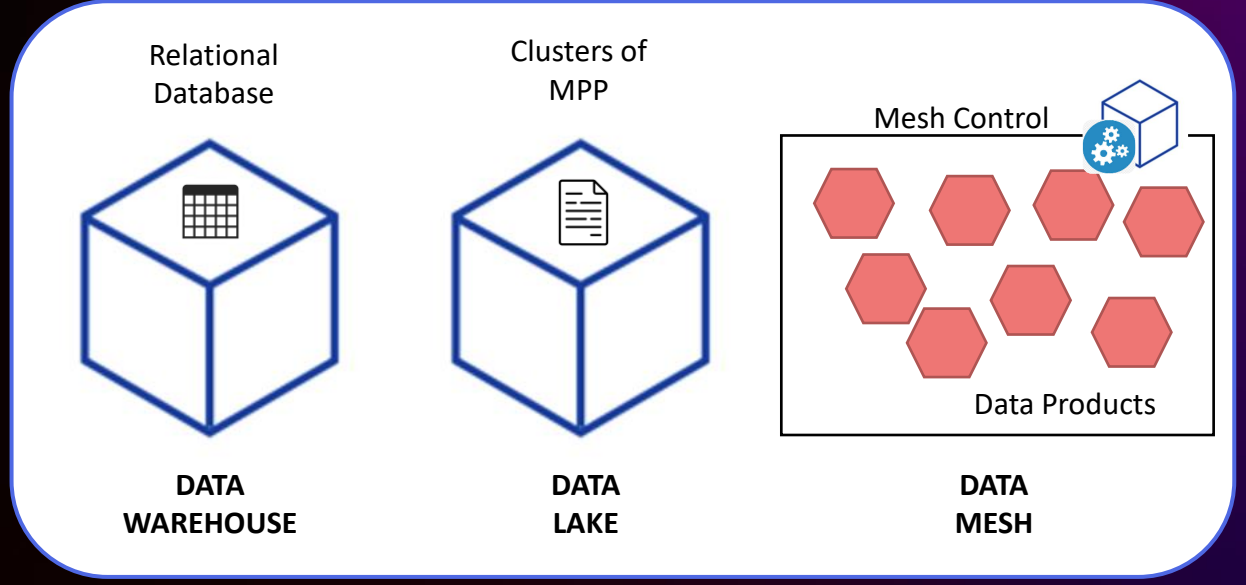
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Systems Architecture Evolution

Transactional

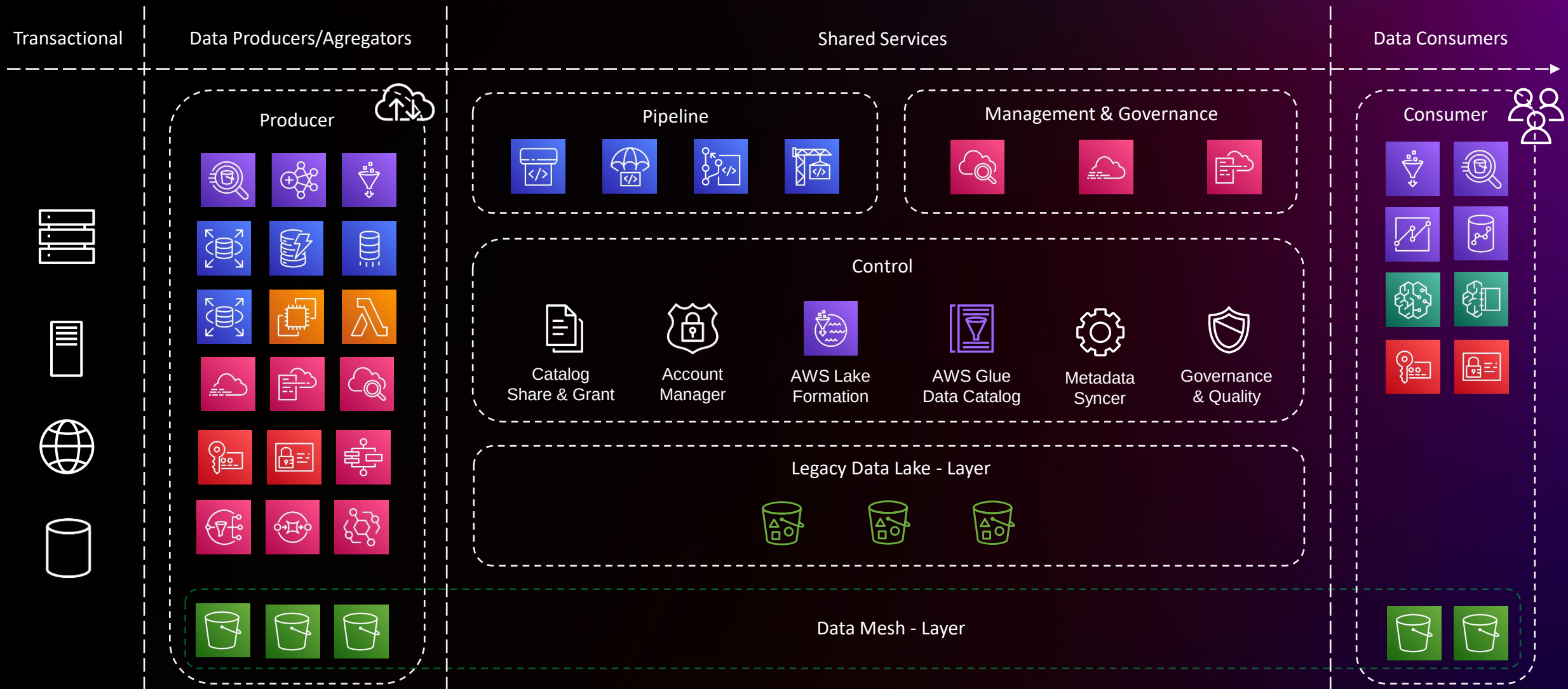


Analytical



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Platform - Itaú



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Platform – Itaú - Available Resources

Analytics


Amazon EMR


Amazon Athena


Amazon QuickSight


AWS Glue


AWS Lake Formation


Amazon Redshift

Management & Governance


Amazon CloudWatch


AWS CloudTrail


AWS CloudFormation

Database


Amazon DynamoDB


AWS Database Migration Service


Amazon Relational Database Service

Application Integration


AWS Step Functions


Amazon EventBridge


Amazon Simple Notification Service


Amazon Simple Queue Service

Developer Tools


AWS CodePipeline


AWS CodeBuild


AWS CodeDeploy


AWS CodeCommit

Security, Identity, & Compliance


AWS Key Management Service


AWS Identity and Access Management

Machine Learning


Amazon SageMaker Studio Lab


Amazon SageMaker

Compute


Amazon Elastic Compute Cloud


AWS Lambda

Storage


Amazon Simple Storage Service

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

.01

Background

.02

How we build a cloud data platform

.03

Formulating a Data Strategy

.04

Addressing Data Architecture Dilemmas

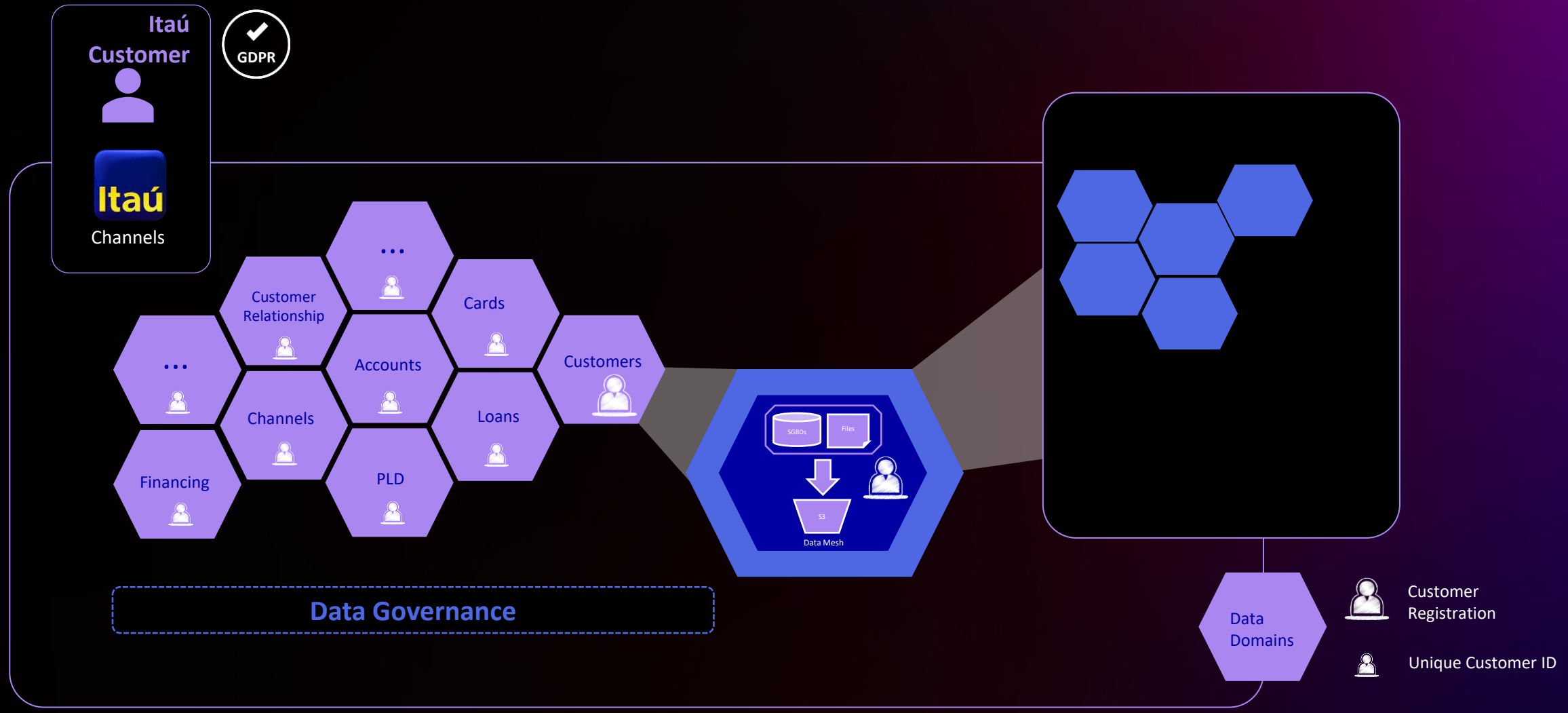
.05

Lessons learned and best practices

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Work with a Canonical Data Model and set your domains



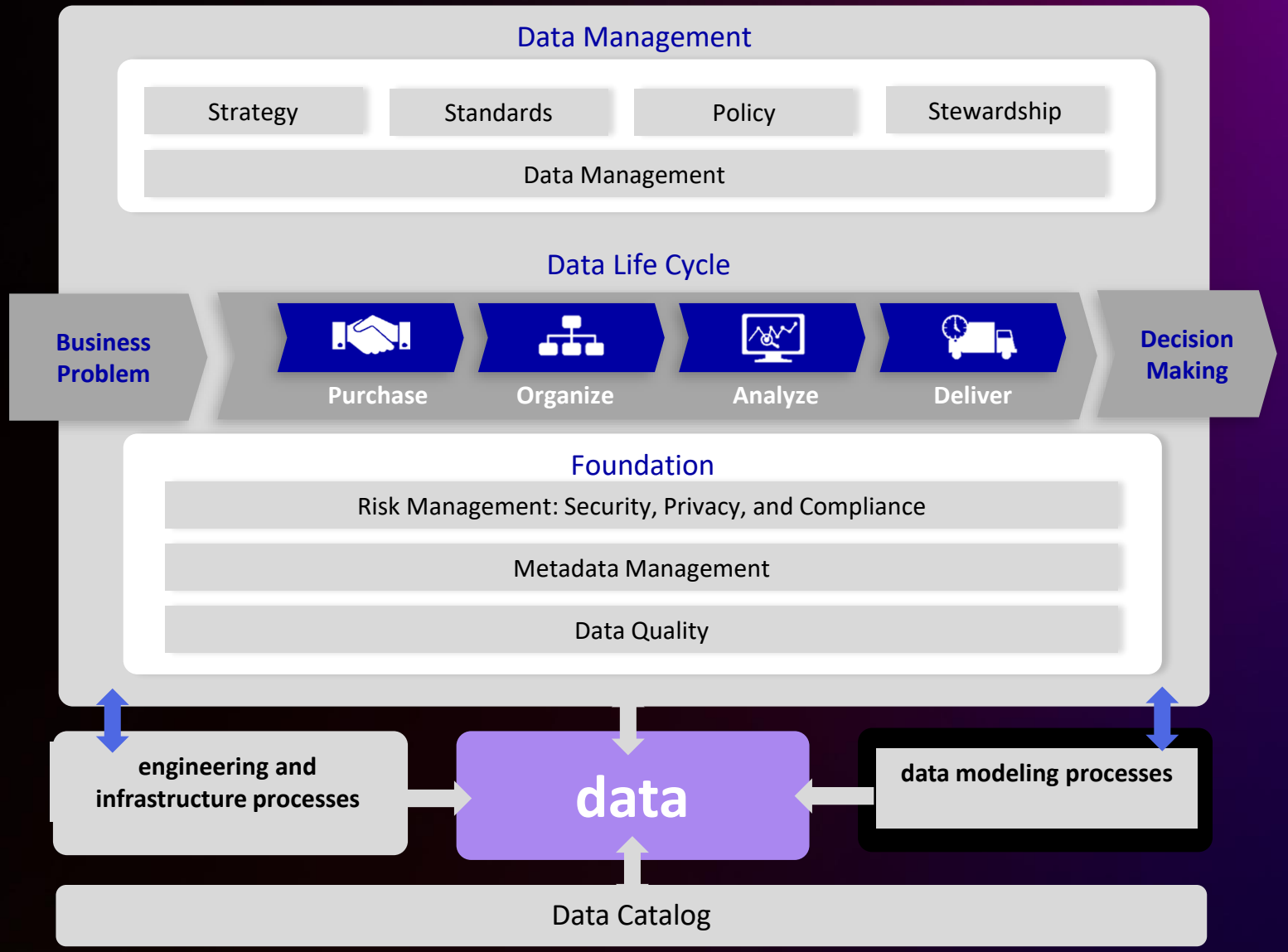
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Establish Data Governance

As a **strategic program** of the organization

+

Integrated with the **technical components** of the Data Platform



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Have a control tower with clear indicators for data excellence in the organization and business impact measurement

The control tower

- **Organizes and generates visibility** on all fronts necessary for the evolution of the subject
- Generates **cadence** in follow-up and evolutions
- Clearly communicates the **strategy**
- Manages transformation and evolution through **indicators**



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

.01

Background

.02

How we build a cloud data platform

.03

Formulating a Data Strategy

.04

Addressing Data Architecture Dilemmas

.05

Lessons learned and best practices

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Security versus Democratization

What to consider when protecting data?

Data security, access control, and data protection may seem similar, but there are differences between them



Data security

Are you really who you say you are?



Access control

Prove that you are who you say you are



Data protection

How can we ensure that data are protected?

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Privacy versus massive data usage

With the decentralization and scale we want to obtain with Data Mesh,
the discipline of privacy needs to be part of the day-to-day of teams

Balanced strategy



Centralized governance

- Policies and guidelines
- Access control
- Expert advice

Teams with autonomy

- Use case assessment
- Definition of technical solutions
- Privacy champions in the squads

- Greater **transparency** for holders
- Adherence to the concept of **minimization**
- Use of data suitable for the desired **purposes**
- Application of **Privacy by Design** concepts

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Governance versus agility

With Data Mesh, both data ingestion and management are naturally integrated into the teams responsible for the data domains, but guardrails and good practices are necessary to maintain governance.

Federated governance for agility

Centralizes → Federates

- Data Management
- Data Quality
- Metadata Management
- Domain data risk management
- Data Security (Authorization)



Adaptive governance for efficiency

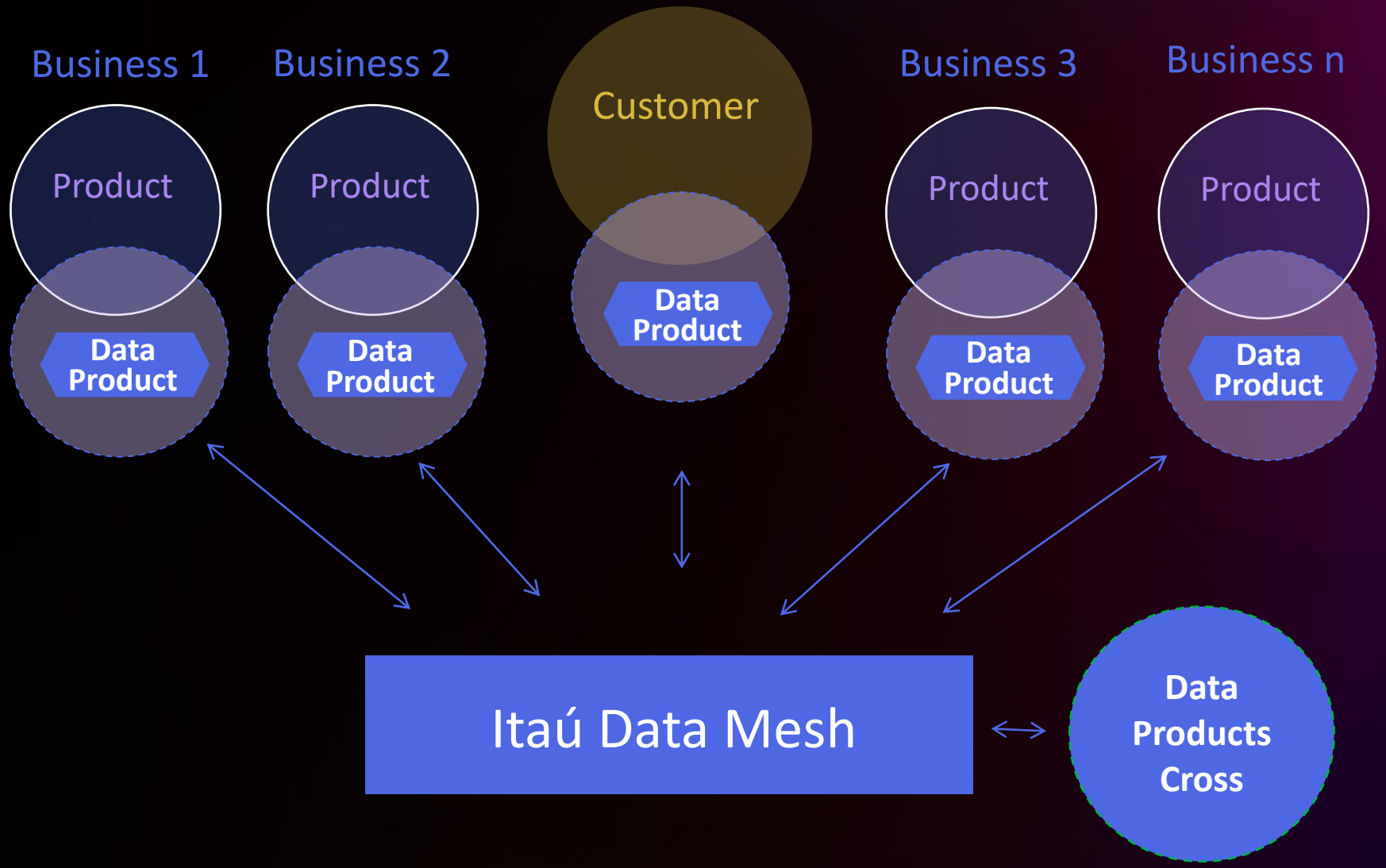
Executes processes → Governs and standardizes

- Policy
- Practices (Governance, Quality, Privacy)
- Standards for Interoperability (universal domain standards)
- Global risk management
- Guardrails and monitoring integrated into the Data Platform
- Data Security (Authentication)

Data Governance

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Centralization versus Scalability



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

.01

Background

.02

How we build a cloud data platform

.03

Formulating a Data Strategy

.04

Addressing Data Architecture Dilemmas

.05

Lessons learned and best practices

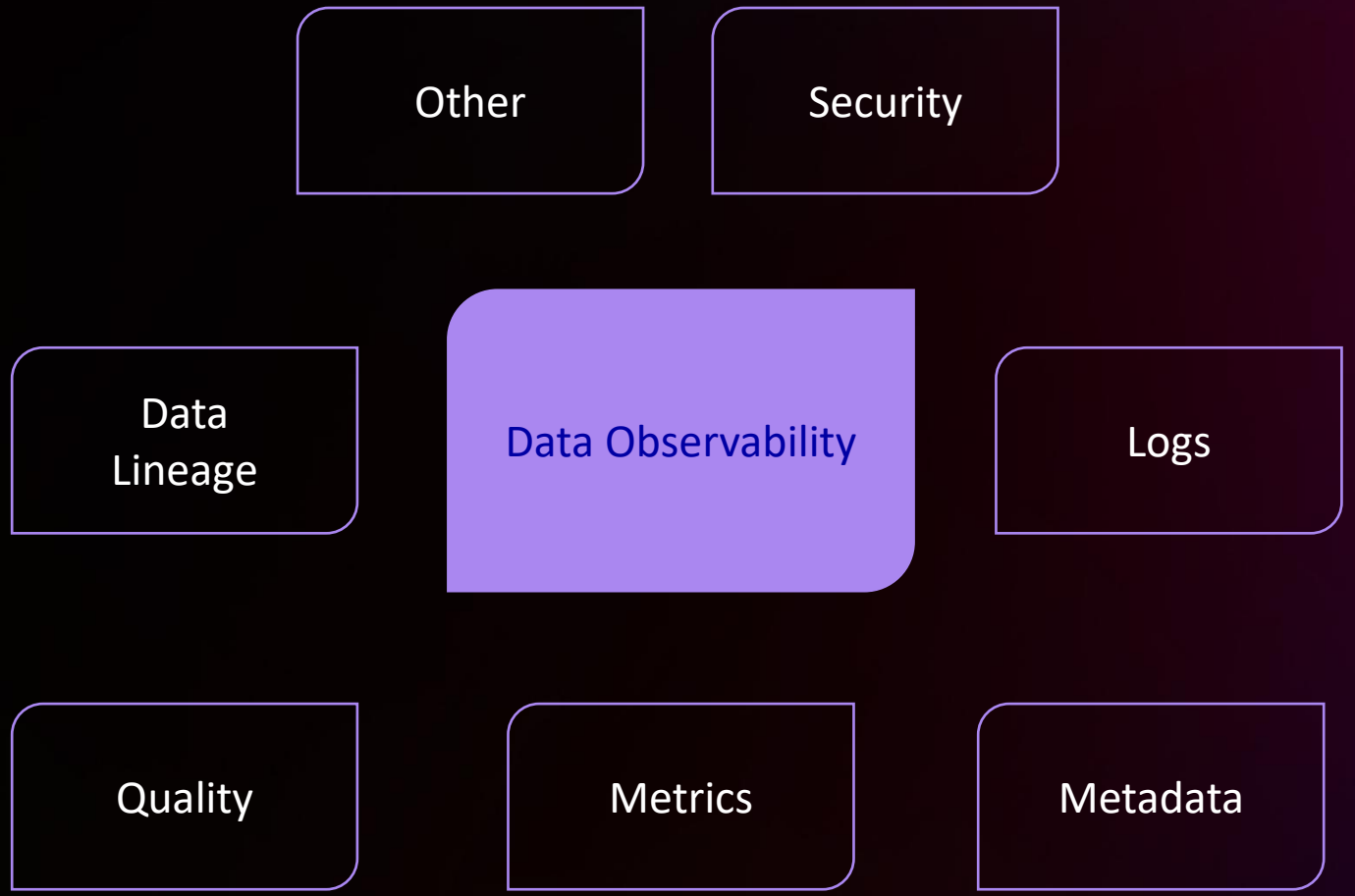
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Successfully architect and deploy your infrastructure: network design, cloud guardrails, and FinOps



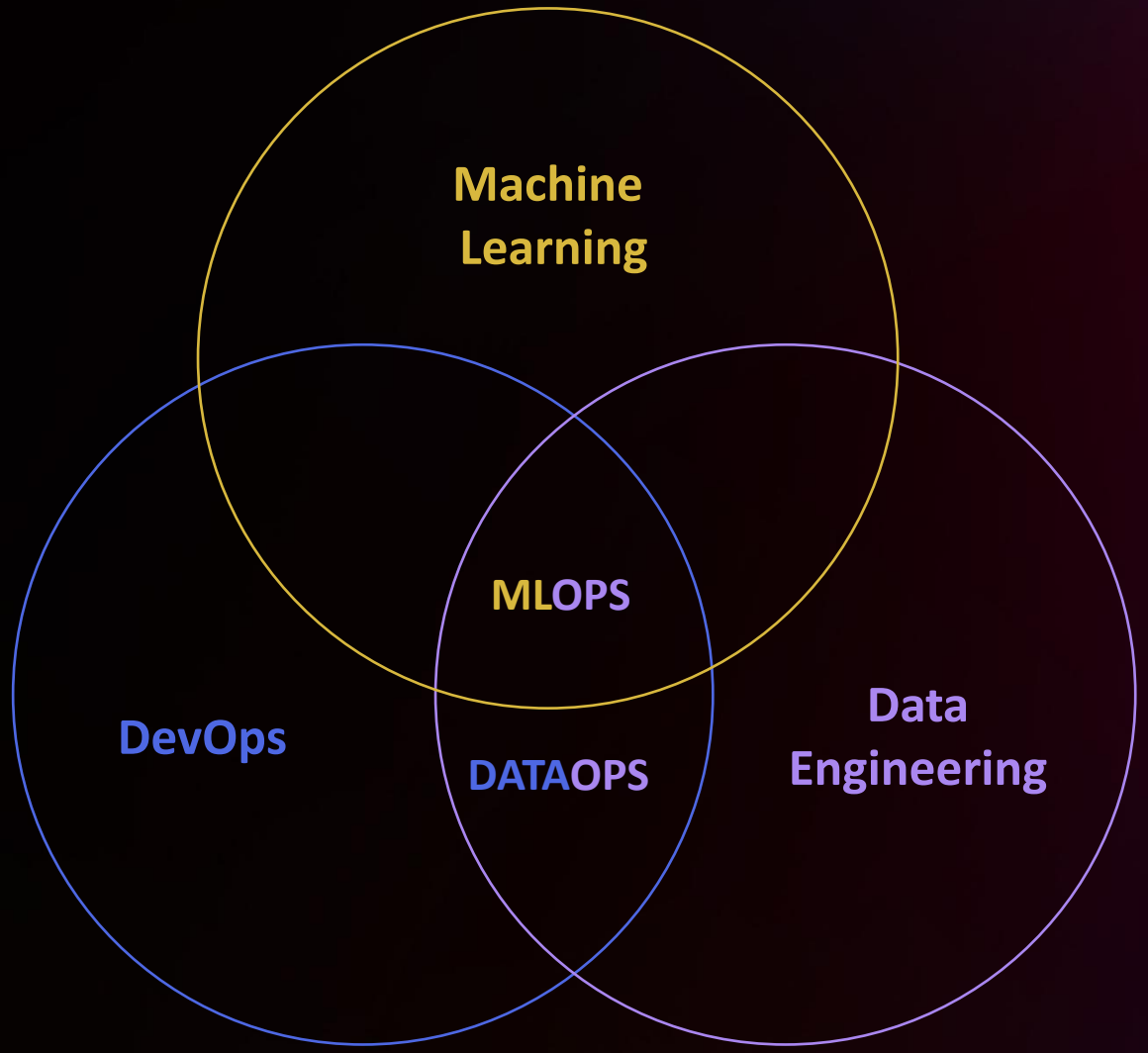
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

In design, enable broad and intuitive observability: platform, components, data pipelines, quality, etc.



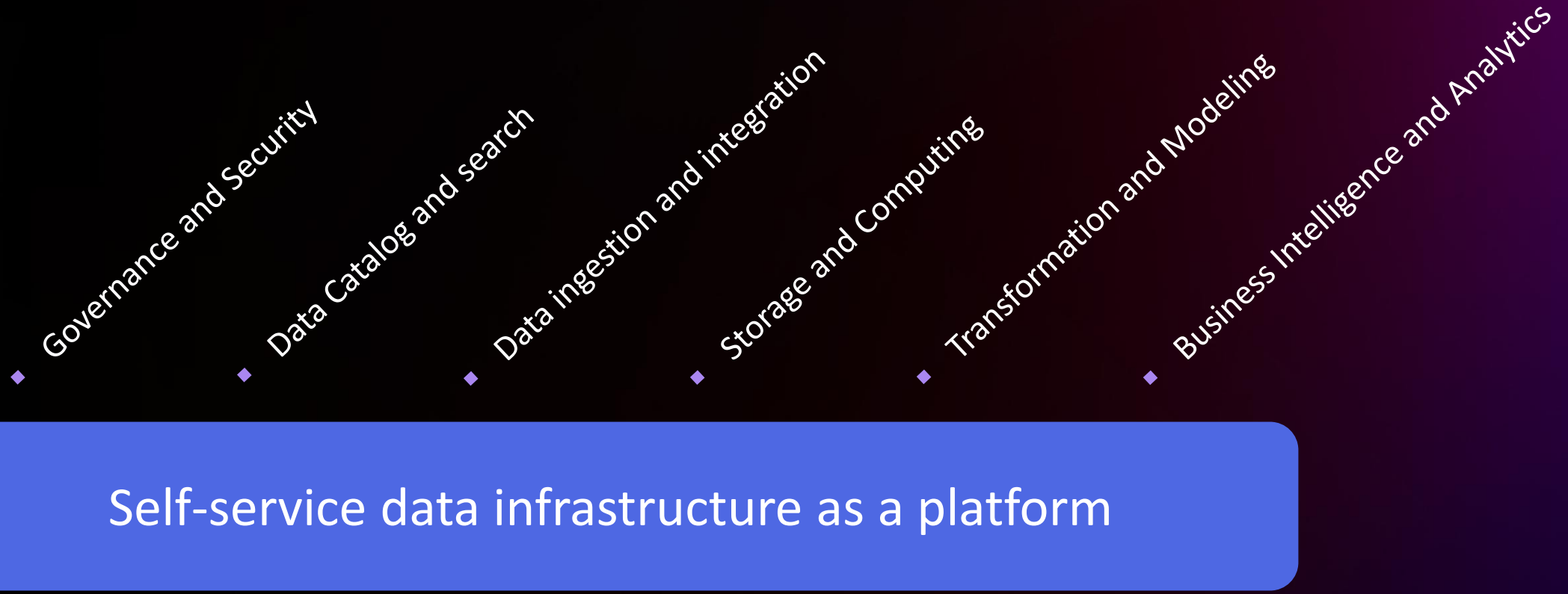
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Establish DataOps and MLOps pipelines



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

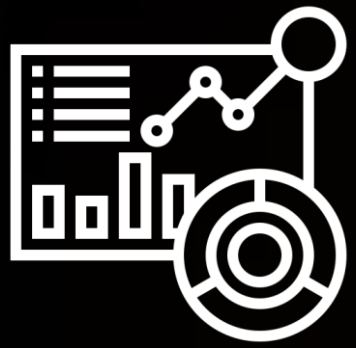
Build the platform with self-service and simple search capabilities



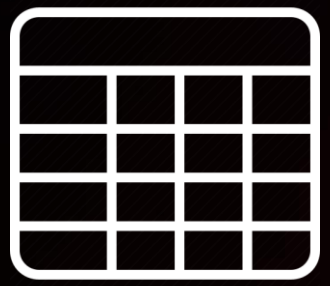
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Think Product When You Think Data

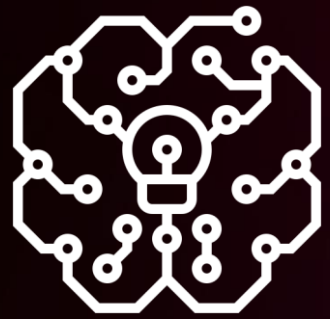
Dashboard



DataSet



Model

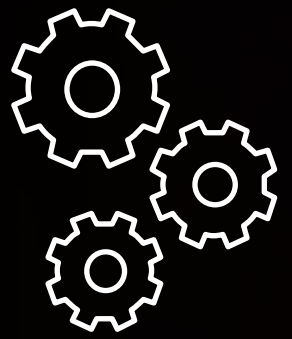


- Shared | Findable
- Self-descriptive
- Addressable
- Interoperable
  governed by global standards
- Reliable
  SLOs Defined and Monitored
- Safe
  Governing by global access control

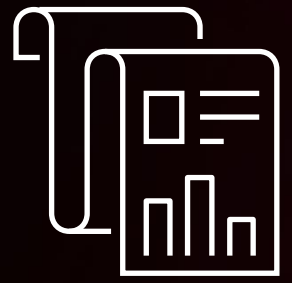
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

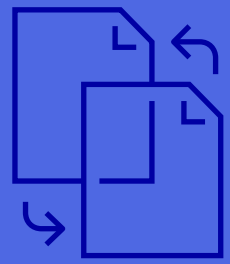
Deliver business value as you build the platform

Technical Components



Business Case





Execute platform components with business use cases in mind, establishing successful partnerships with shared responsibility.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Strategy

2.1 Data Architecture

2.2 Governance, Privacy, and Quality

2.3 Change Management, Training, & Indicators

2.4 Data-Driven Organization

Platform

4.1 Infrastructure and FinOps

4.2 Observability

4.3 Security & Metadata

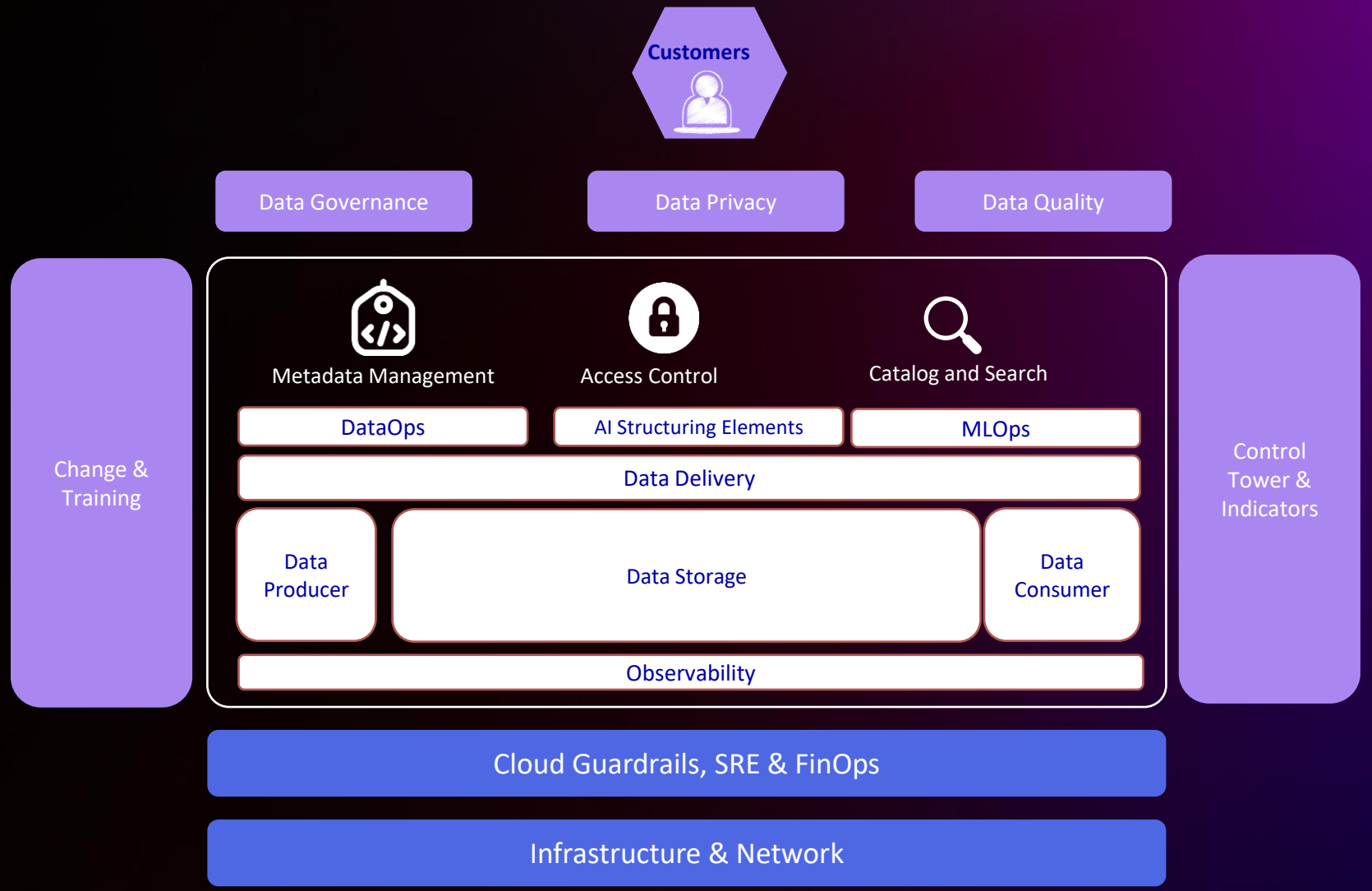
4.4 DataOps and MLOps

4.5 Self-service

4.6 Data Product

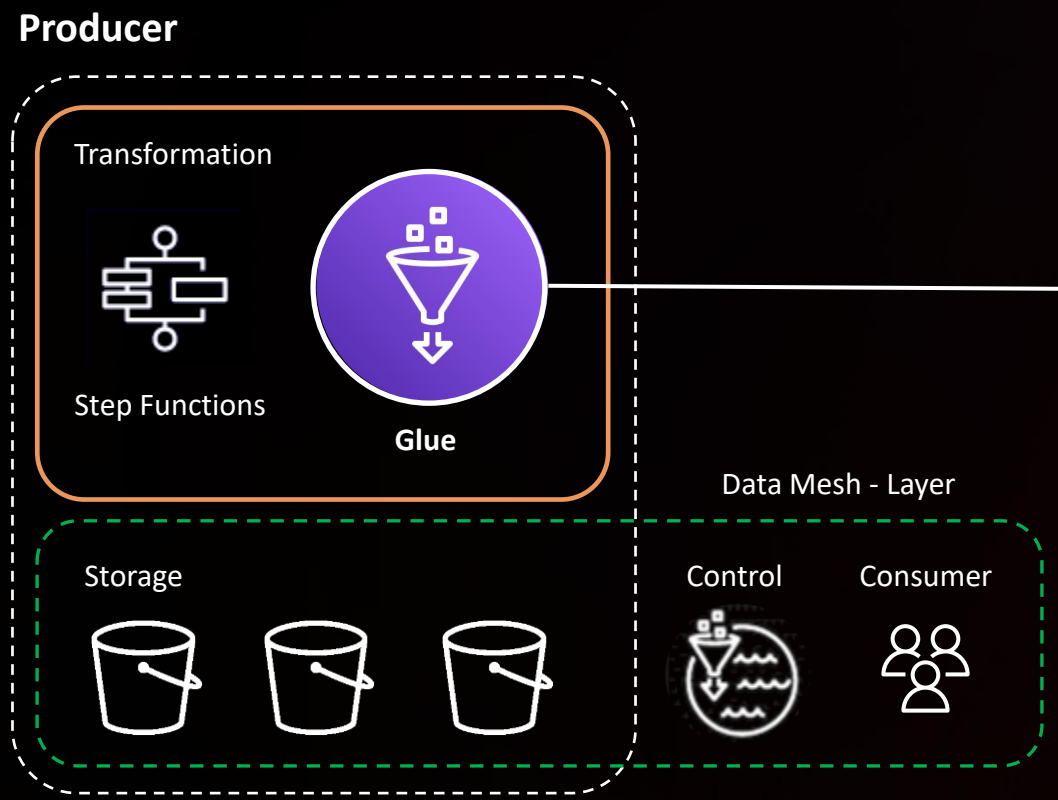
4.7 Value for the business

Data Architecture



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Case AWS Glue - Payments



52 jobs

~24 workers per job

114 Terabytes of processed data

18 Terabytes of democratized information

24 hours data processing

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Case AWS Glue - Payments

Increase payments market share

Easily Accessible Data

Access all the necessary data to make decisions in a unique repository.

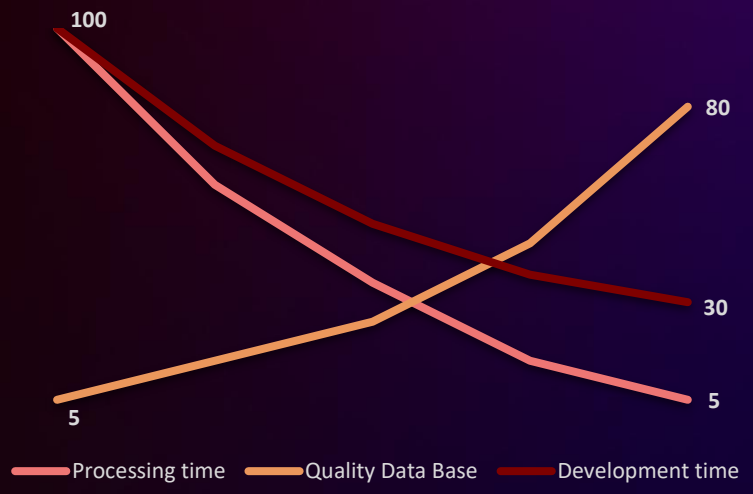
Timely Data

Data made avaiable in a time manner apropriated to the timing to marketing.

Data Quality

Use of governed data, with observability and validated concepts.

95% efficiency in processing time, 70% shorter development time and 75% increase in the number of governed databases available in data mesh



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

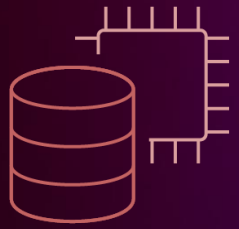
Summary



Broad Adoption



Cost Effective Scale



Embrace Open Source



Glue across AWS Services



Rapid Innovation



Data analytics & machine learning

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary

- Glue combines the best of data integration and ETL tools with big data open source engines
- Full solutions for all data integration and delights multiple personas
- Scale, reliability, and performance
- Innovations and announcements



ANT222: AWS Glue Data Quality
11/30 Wednesday 6:15
Level 3, Chairmans 368, MGM Grand

ANT343: AWS Glue with Ray session
12/1 Thursday 4:15 MGM – Grand 121

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How to get started?

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Get Started today

Thinking of migration?

AWS teams and programs can help you with

Building your business case, creating migration plan

Deciding on budget, scope, etc.

Conduct a proof of concept or architect your data strategy

Get Help

Migration partner program with more than 20 partners:

bitwise

cognizant

wavicle DATA SOLUTIONS

IMPETUS

accenture

blade bridge

quantiphi

next PATHWAY

This is not a complete list. To view all AWS Partners for this category, visit AWS Partner Solutions Finder. This list of partners is current as of November 29, 2022.

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Get Started

New to AWS Glue?

- ✓ Explore demos, customer stories, and the latest features on aws.amazon.com/glue
- ✓ Get trained
- ✓ Ask your account team for a 10-minute demo



Already using AWS Glue?

- ✓ Ask your account team for a free AWS Glue session for further deep dive
- ✓ Learn more about what's new https://aws.amazon.com/glue/whats-new/

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Santosh Chandrachood
sanchas@amazon.com

Moisés Nascimento

Roberto Figueira



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.