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

ANT223

Simplify & accelerate data integration & ETL modernization w/AWS Glue

Speakers

- Roberto Santos Filho, IT Superintendent, Banco Itau Unibanco
- Santosh Chandrachood, General Manager, Amazon
- Moises Nascimento, Director, Itau Unibanco S/A

Announcements

- Lots of incremental announcement. See notes and slide 12 for a summary.

Takeaways

- Great Overview of Glue. If you don't know what Glue entails this is a great session. Also Itau's approach to Data Strategy is interesting.

Azure

- Azure's Data ETL/Pipelining tools is Data Factory.

< Section Header

< Table of Contents

Trends & Challenges in Data Integration

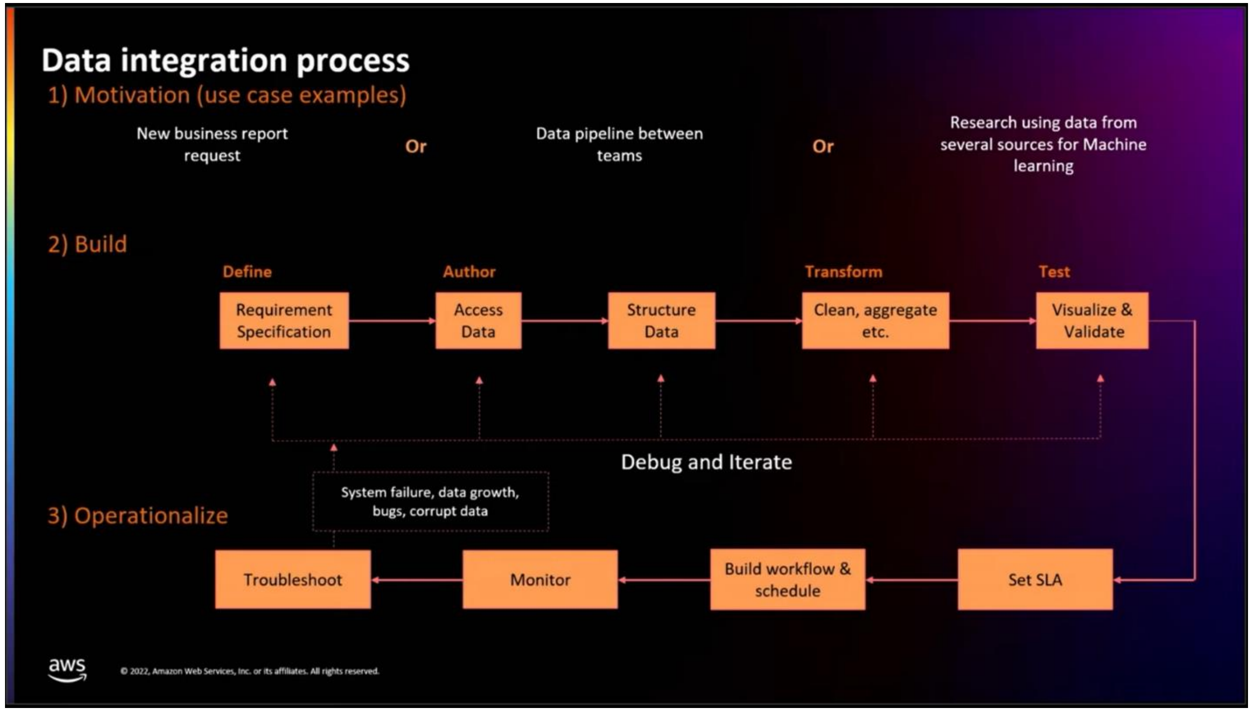
- Business Challenges - Lack of Agility, Lack of Self-Service, Missed SLAs
- Data Engineering Challenges – Infrastructure Management, Multiple Tools, Scalability Issues
- Leadership Challenges – Expensive, No multi-persona support, Lock-in
- Data & Pipeline Challenges (See Slide 2 & 3)

Glue Overview

- Why Glue?
  - Focus on Data not Infrastructure
  - Serverless so scales on Demand
  - Based on Open-Source Engines
  - Works for a variety of workloads and personas
- New – Support for Cloud Shuffle Plugin
  - Improves reliability for Spark jobs by allowing rebalance of storage across clusters
- New – Glue 4.0
  - Upgrades Glue Engine to latest tools for Spark, Python, Scala
  - Spark Connector for Redshift Integration
- New – Glue with Ray
  - Upgrades Glue Engine to latest tools for Spark, Python, Scala
  - Allows Python jobs to scale across multiple nodes for large workload processing.
- New – Native support for Hudi, Delta Lake, Iceberg
  - These technologies bring transaction capability to Data Lake stores
  - Glue now supports this capability natively, no connectors needed
- New – Custom Visual Transforms
  - Build and Share Custom transformations across the organization
- New – Glue Git Integration
  - Manage version and configuration with your favorite Git repository.
- New – Data Catalog Capabilities
  - Makes it easier to track, audit and manage data & access.
- New – Data Catalog Crawler Capabilities
  - Cross Account Capability, Snowflake, MongoDB.
- New – Data Quality
  - Demo.

Why Itau Chose Glue for Data Mesh

- Watch Here.
  - Leveraged AWS and Glue to build a Modern Data Mesh platform.
  - Formulating a Data Mesh Strategy
  - Built a canonical data model to identify domains of data
  - Establish Governance
  - Data Security vs Democratization
  - Data Privacy vs Massive Data Usage
  - Governance vs Agility
  - Centralization vs Scalability
  - Think Product when you think data
  - Example of how Glue helped them achieve visibility of Payments.



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

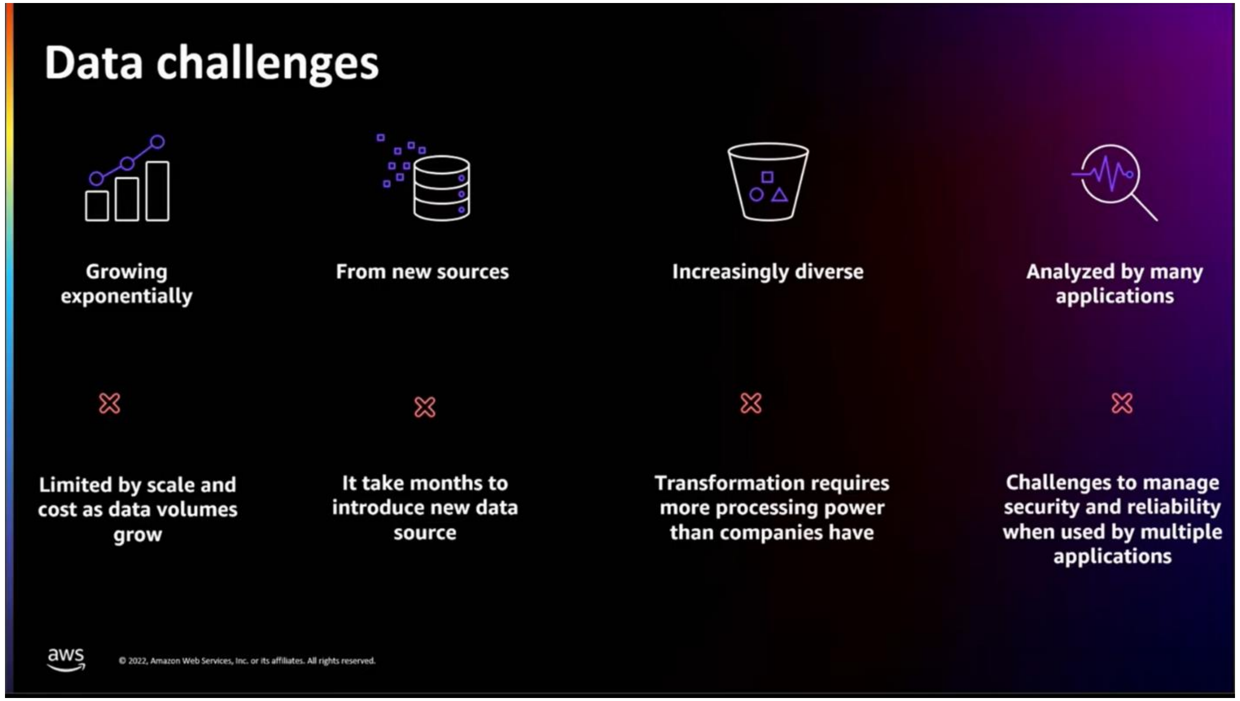
Define: Requirement Specification → Author: Access Data → Structure Data → Transform: Clean, aggregate etc. → Test: Visualize & Validate

Debug and Iterate

System failure, data growth, bugs, corrupt data

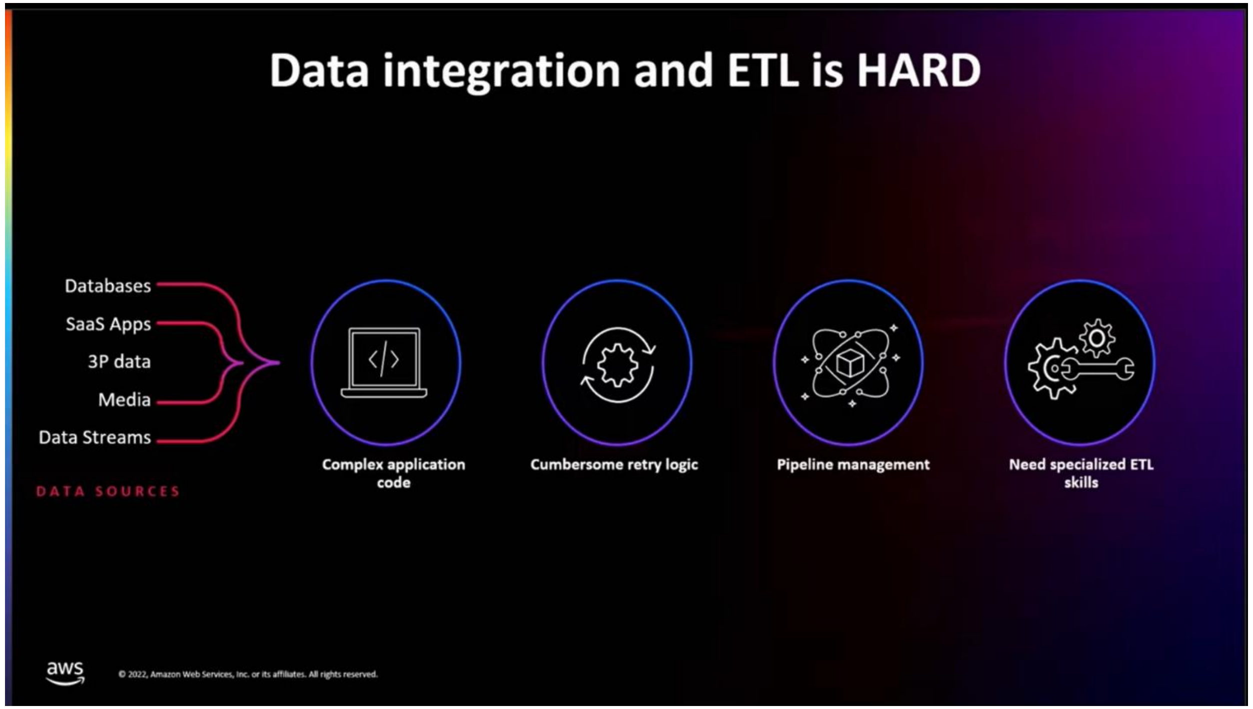
3) Operationalize

Set SLA → Build workflow & schedule → Monitor → Troubleshoot



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



Data integration and ETL is HARD

Databases
SaaS Apps
3P data
Media
Data Streams

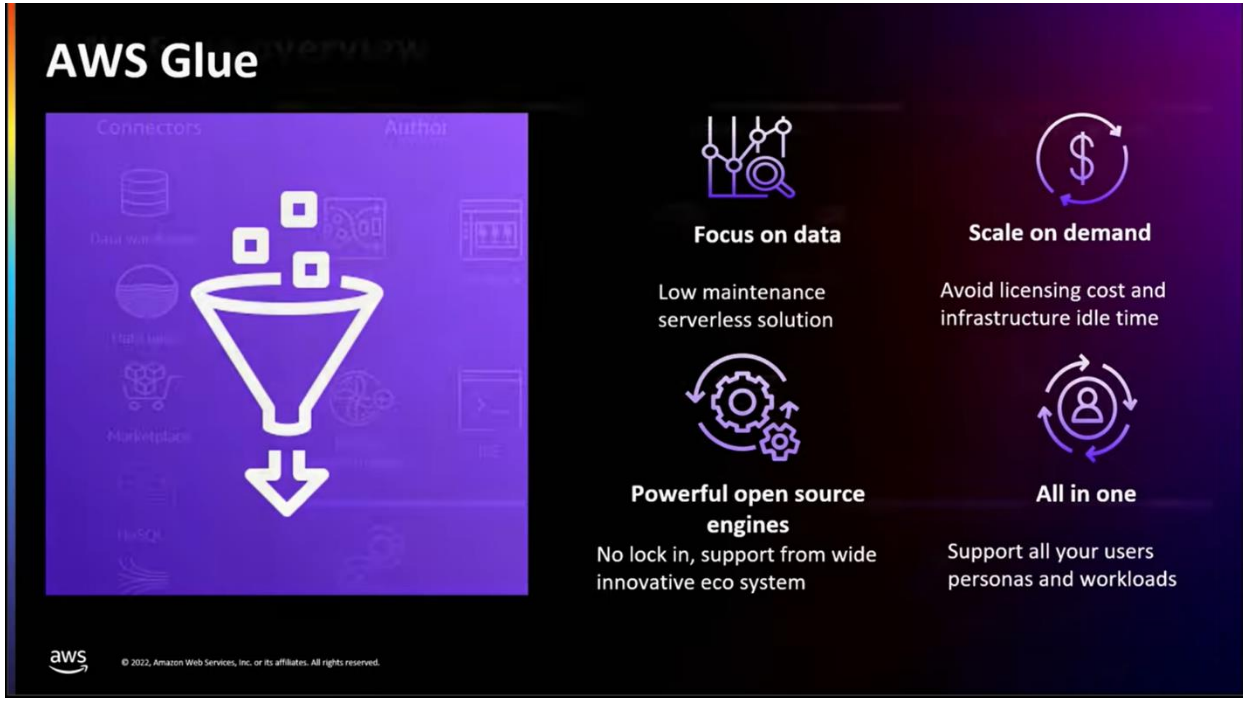
DATA SOURCES

Complex application code

Cumbersome retry logic

Pipeline management

Need specialized ETL skills



AWS Glue

Focus on data

Low maintenance serverless solution

Scale on demand

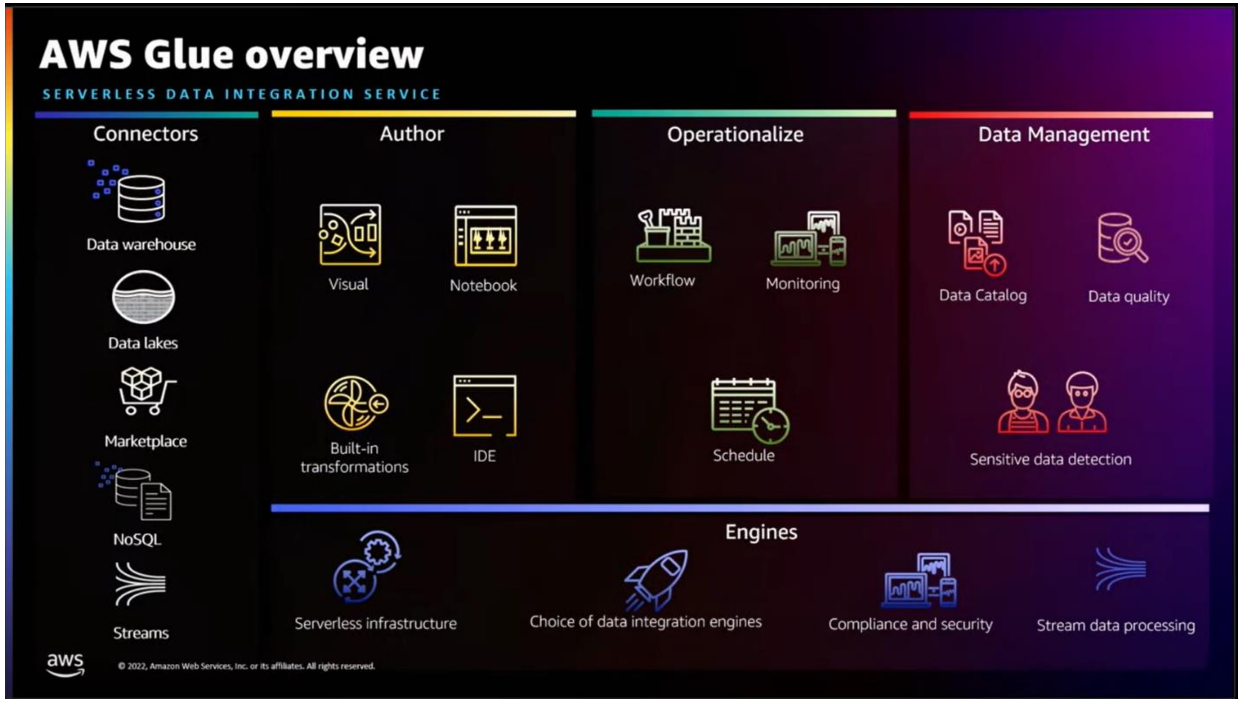
Avoid licensing cost and infrastructure idle time

Powerful open source engines

No lock in, support from wide innovative eco system

All in one

Support all your users personas and workloads



AWS Glue overview

SERVERLESS DATA INTEGRATION SERVICE

Connectors
Data warehouse
Data lakes
Marketplace
NoSQL
Streams

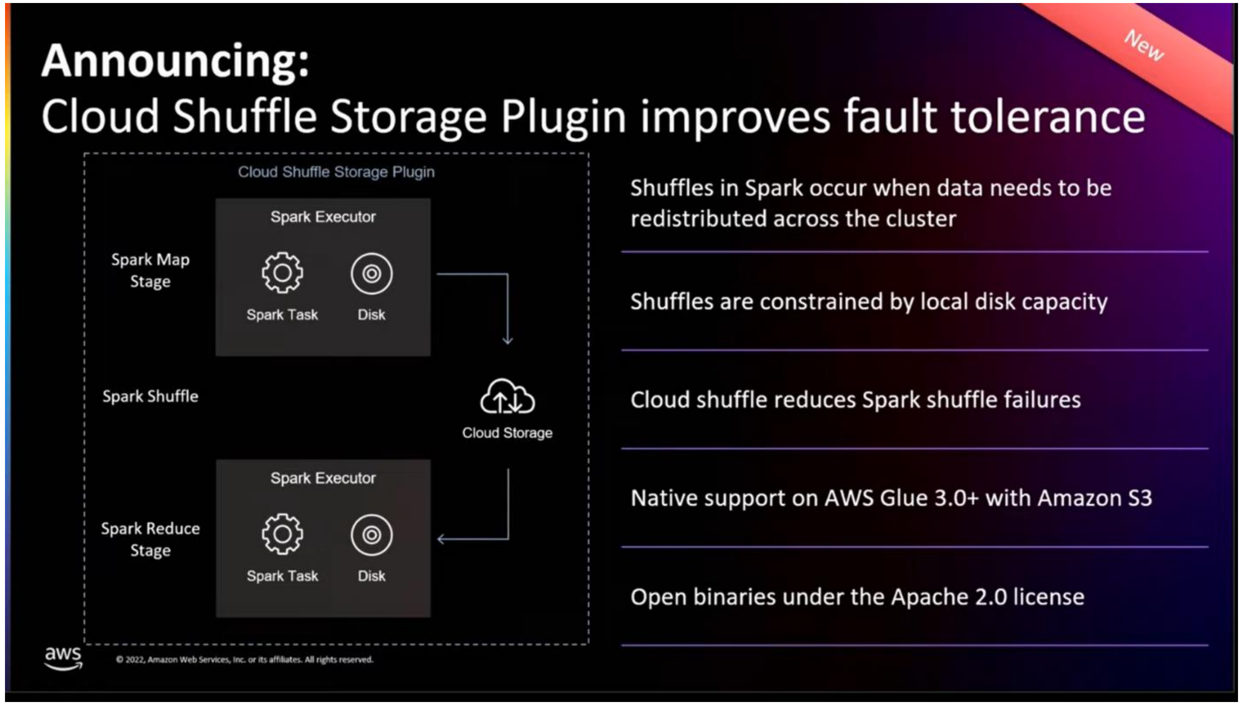
Author
Visual
Notebook
Built-in transformations
IDE

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

aws



Announcing:
Cloud Shuffle Storage Plugin improves fault tolerance

New

Cloud Shuffle Storage Plugin
Spark Map Stage
Spark Executor
Spark Task
Disk
Spark Shuffle
Cloud Storage
Spark Reduce Stage
Spark Executor
Spark Task
Disk

Shuffles in Spark occur when data needs to be redistributed across the cluster

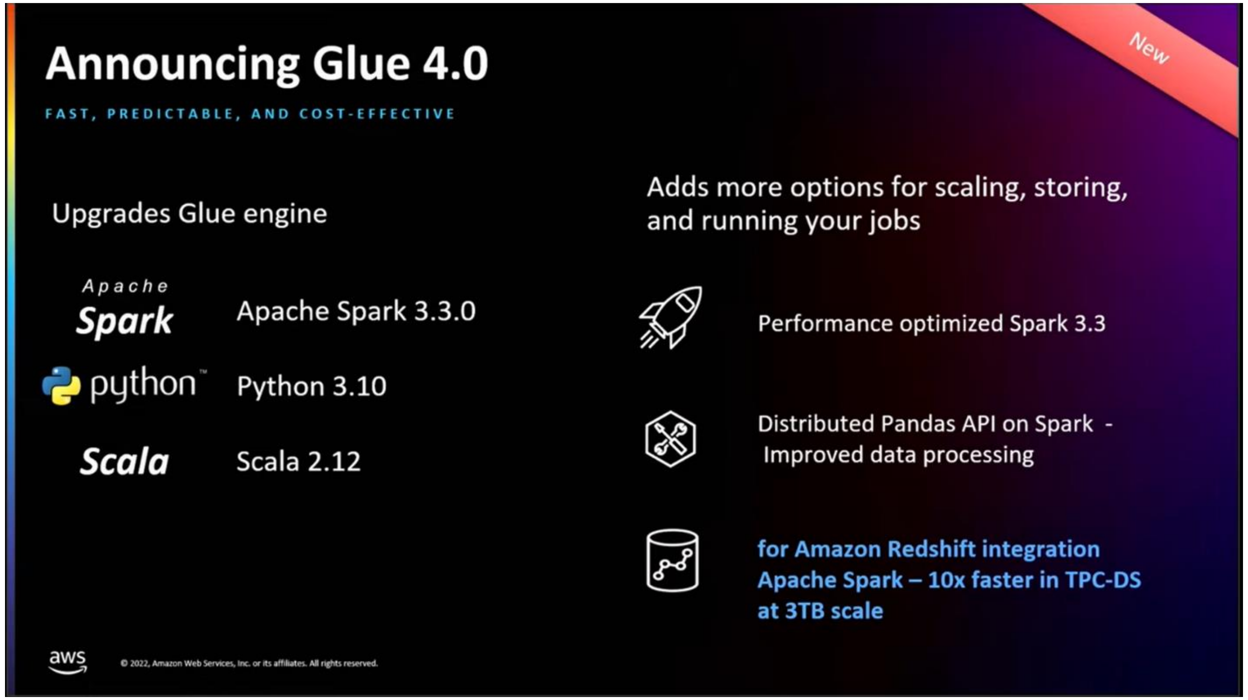
Shuffles are constrained by local disk capacity

Cloud shuffle reduces Spark shuffle failures

Native support on AWS Glue 3.0+ with Amazon S3

Open binaries under the Apache 2.0 license

aws



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

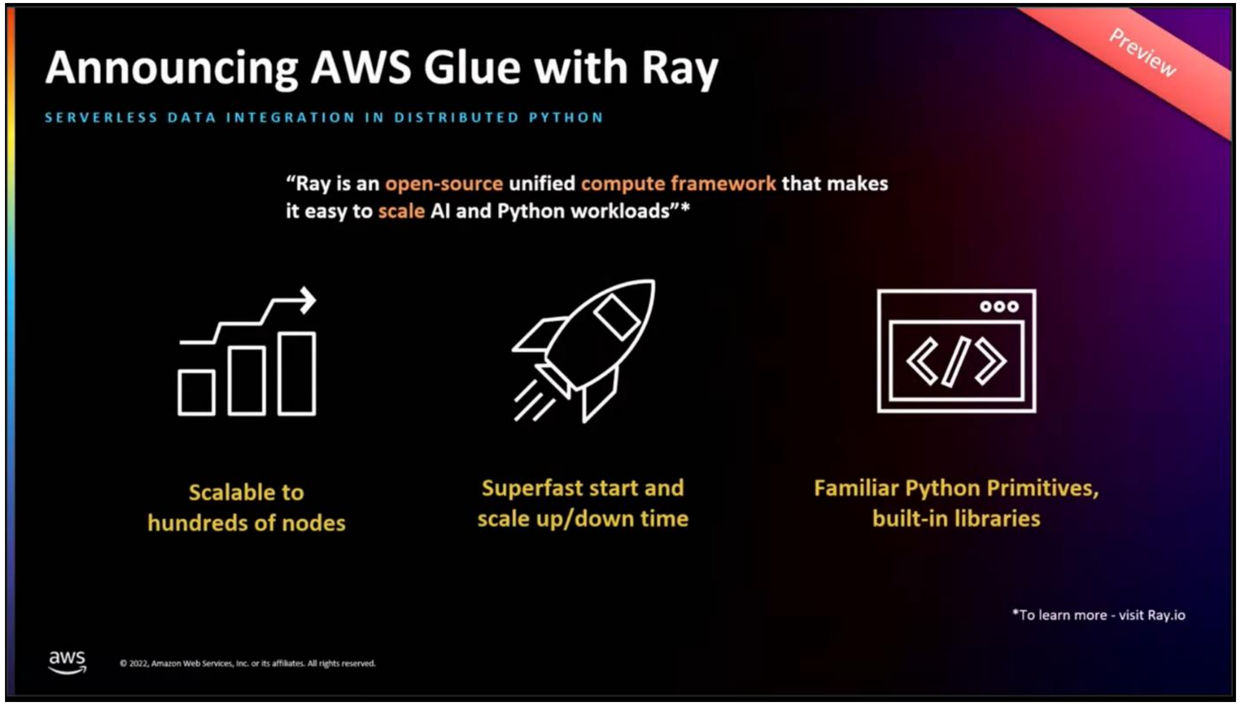
Adds more options for scaling, storing, and running your jobs

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



Announcing AWS Glue with Ray

SERVERLESS DATA INTEGRATION IN DISTRIBUTED PYTHON

Preview

"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

New

Announcing: Native support for Hudi, Delta Lake and Iceberg

Apache hudi

Linux Foundation Delta Lake

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

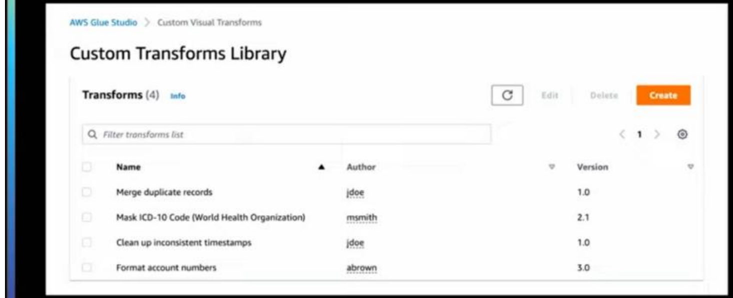
Latest framework versions available with Glue v4.0

aws

New

Announcing: Custom visual transforms

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



AWS Glue Studio > Custom Visual Transforms

Custom Transforms Library

Transforms (4) Info

	Name	Author	Version
□	Merge duplicate records	jdoe	1.0
□	Mask ICD-10 Code (World Health Organization)	msmith	2.1
□	Clean up inconsistent timestamps	jdoe	1.0
□	Format account numbers	abrown	3.0

Benefits:

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

aws

New

AWS Glue Git integration

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

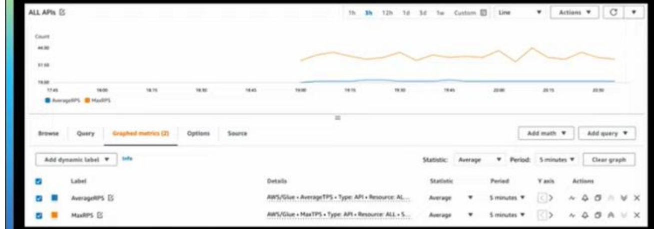
Supports GitHub and AWS CodeCommit

Download/upload AWS Glue jobs in AWS Management Console

aws

New

Data Catalog – Capabilities



☑	Label	Details	Statistic	Period	Y axis	Actions
☑	AverageRPS	AWS/Glue • AverageTPS • Type: API • Resource: AL...	Average	5 minutes		
☑	MaxRPS	AWS/Glue • MaxTPS • Type: API • Resource: ALL • S...	Average	5 minutes		

Audit with CloudWatch Catalog Metrics

Federate Hive MetaStore and Access with Athena, RedShift and Glue

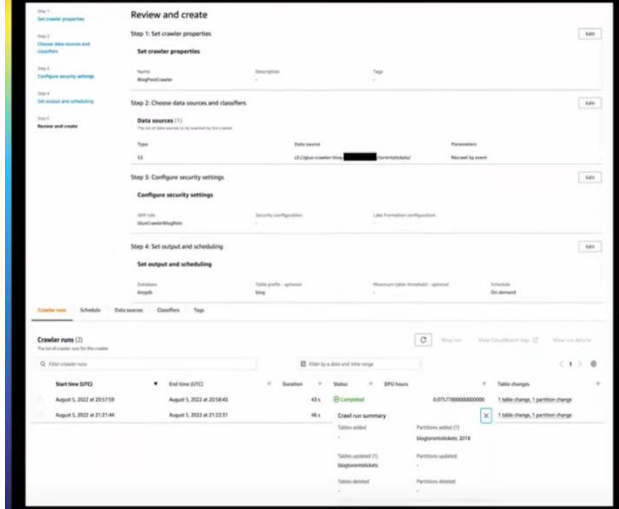
Database, Connection Tagging

Sync Delta Lake Metadata in Catalog to enable column permissions

aws

Glue Data Catalog - Crawlers

New



Review and create

Step 1: Set crawler properties

Set crawler properties

Name: BlogPostCrawler | Description: - | Tags: -

Step 2: Choose data sources and classifiers

Data sources (1)

Type	Data source	Parameters
S3	s3://glue-crawler-blog[redacted]/torontotickets/	Recrawl by event

Step 3: Configure security settings

Configure security settings

IAM role: GlueCrawlerBlogRole | Security configuration: - | Lake Formation configuration: -

Step 4: Set output and scheduling

Set output and scheduling

Database: blogdb | Table prefix - optional: blog | Maximum table threshold - optional: - | Schedule: On demand

Crawler runs (2)

Start time (UTC)	End time (UTC)	Duration	Status	DPU hours	Table changes
August 5, 2022 at 20:57:59	August 5, 2022 at 20:58:45	45 s	Completed	0.075777800000000	1 table change, 1 partition change
August 5, 2022 at 21:21:44	August 5, 2022 at 21:22:31	46 s			1 table change, 1 partition change

Crawl run summary

Tables added: - | Partitions added (1): blogtorontotickets: 2018

Tables updated (1): blogtorontotickets | Partitions updated: -

Tables deleted: - | Partitions deleted: -

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

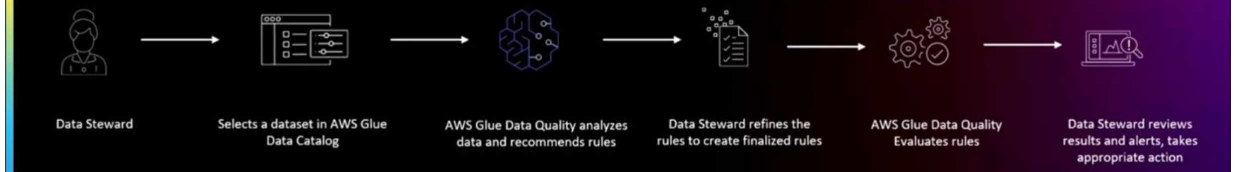
AWS Glue Data Quality Features and Benefits

Preview

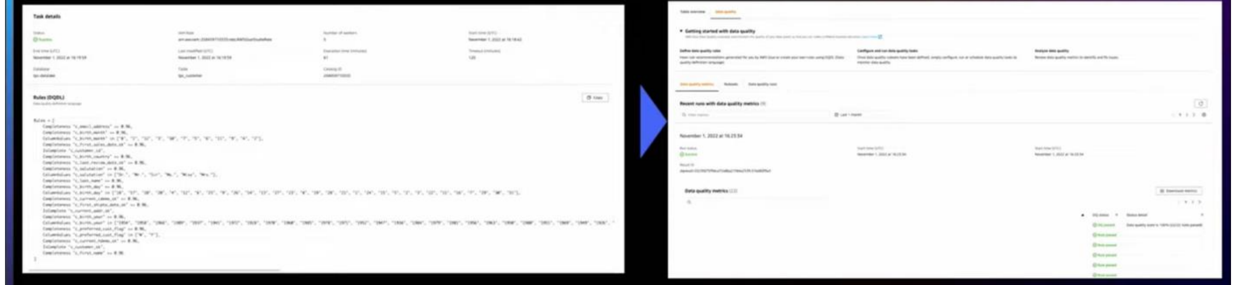
- Serverless, scalable and performant
- Data Quality Recommendations
- Out-of-the-box data quality rules and actions
- Data Quality Definition Language
- Data Quality for data-at-rest and data-in-transit
- Supports multiple persona

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



Task details

Rules (DQDL)

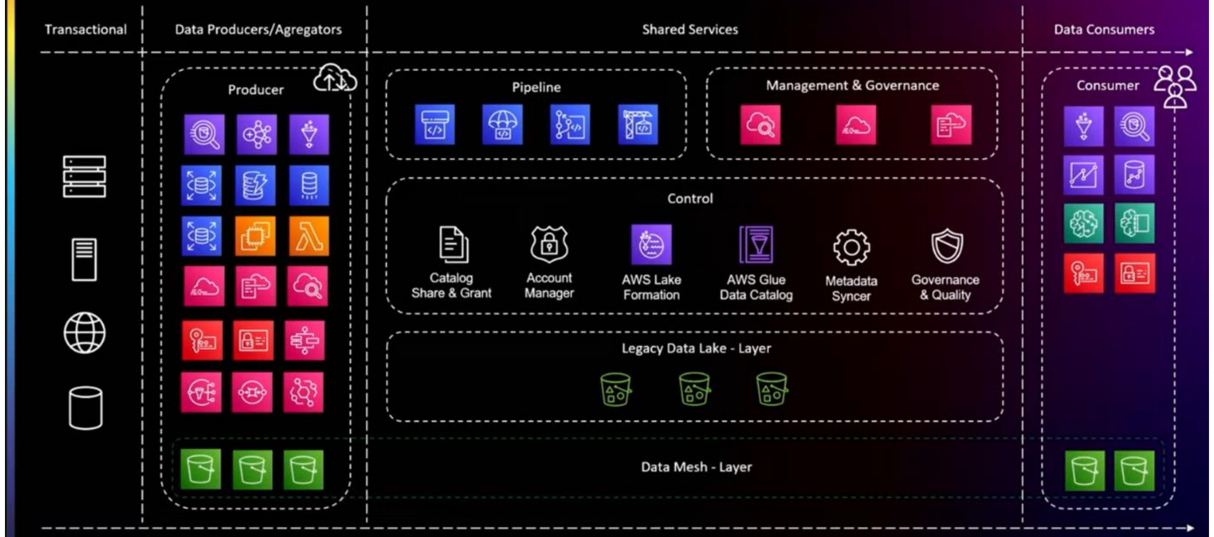
```
Rules = [
    Completeness "c_email_address" >= 0.96,
    ...
]
```

Getting started with data quality

Recent runs with data quality metrics

Data quality metrics

Data Platform - Itaú



Transactional
Data Producers/Agregators
Shared Services
Data Consumers
Producer
Pipeline
Management & Governance
Control
Catalog Share & Grant
Account Manager
AWS Lake Formation
AWS Glue Data Catalog
Metadata Syncer
Governance & Quality
Legacy Data Lake - Layer
Data Mesh - Layer
Consumer