Simplified Data Integration with AWS Glue

DISCOVER, PREPARE, AND INTEGRATE ALL YOUR DATA AT ANY SCALE

ROHAN GHOSH

Enterprise Solutions Architect, AWS

Event Agenda

09.30	Introduction and Orientation
09.45	Introduction to AWS Glue
10.30	Getting started with AWS
10.45	Lab 01: Working with Glue Data Catalog
11.30	Break
11.45	Lab 05: Working with Glue Studio
12.30	Unleashing the power of Data Integration with Glue
13.15	Lunch
14.00	Demo: Building a transactional data lake with Apache Iceberg tables
14.45	Introduction to DataBrew
15.15	Lab 11: Working with Glue Databrew
16.15	Break
16.30	Demo: AWS Glue Data Quality
17.00	Q&A
17.15	Conclusion & wrap-up

Introduction to AWS Glue

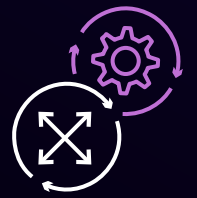
- Trends and challenges in data integration
- What is AWS Glue and how do customers use it?
- Authoring data integration jobs
- Data management
- Data integration engines
- Operationalise
- Better together: Amazon Redshift and AWS Glue
- Simplifying ETL migration
- New features


AWS Glue is a serverless data integration service for easy to discover, prepare, and combine data for analytics, machine learning, and application development.

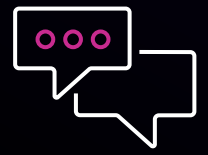
Trends and challenges in data integration

Put data **to work**

Make better decisions



Improve efficiencies



Respond faster

Uncover opportunities

Customers want more value from their data

Growing exponentially

From new sources

Increasingly diverse

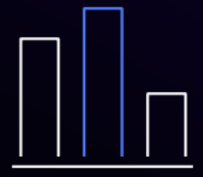
Used by many people

Analyzed by many applications

More demanding workloads

Batch

Data warehousing



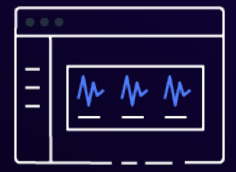
Business intelligence

Reporting

Nightly or hourly ETL

Latency-agnostic, long-running jobs

Real-time



Responsive dashboards

Monitor and alert

Streaming apps

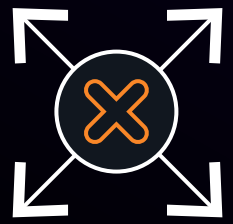
Continuous operation

Latency-sensitive micro-batch jobs

Traditional ETL solutions can’t handle demands

Require multiple solutions

Multiple solutions are needed to get the job done



Non-scalable infrastructure

On-premises ETL tools are complex to install, manage, and scale

High cost

Advanced features–like centralized data catalog, handling streaming data–are licensed separately

Unique user skill sets

Users have varying technical abilities to interact with data

What is AWS Glue and how do customers use it?

AWS Glue

DISCOVER, PREPARE, AND INTEGRATE ALL YOUR DATA AT ANY SCALE

All-in-one data integration service

Cost effective, serverless, and scalable

Tailored tools to support all data users

Support all workloads in one place

How customers use

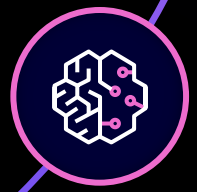
AWS Glue

Apply **modern data strategies** (for example, data lakes, data mesh) for scalable data analysis

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

Catalog data assets to make them available across a modern data architecture

Process petabytes of data in batch, streaming, and real time using Python and Apache Spark



Prepare data for analytics and Machine Learning

Break down data silos across your organisation

ETL

Visual data preparation

Data replication

Data warehouse to/from data lake

Federated query

AWS Glue: Key capabilities

Serverless data integration service

Connectors

Data warehouse

Data lakes

Marketplace

NoSQL

Streams

Author

Visual

Notebook

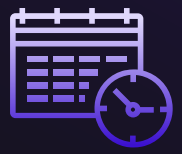
Built-in transformations

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

Complete data integration capabilities in one place

Stream processing

More easily integrate with **Amazon MSK, Amazon Kinesis,** and **Apache Kafka**

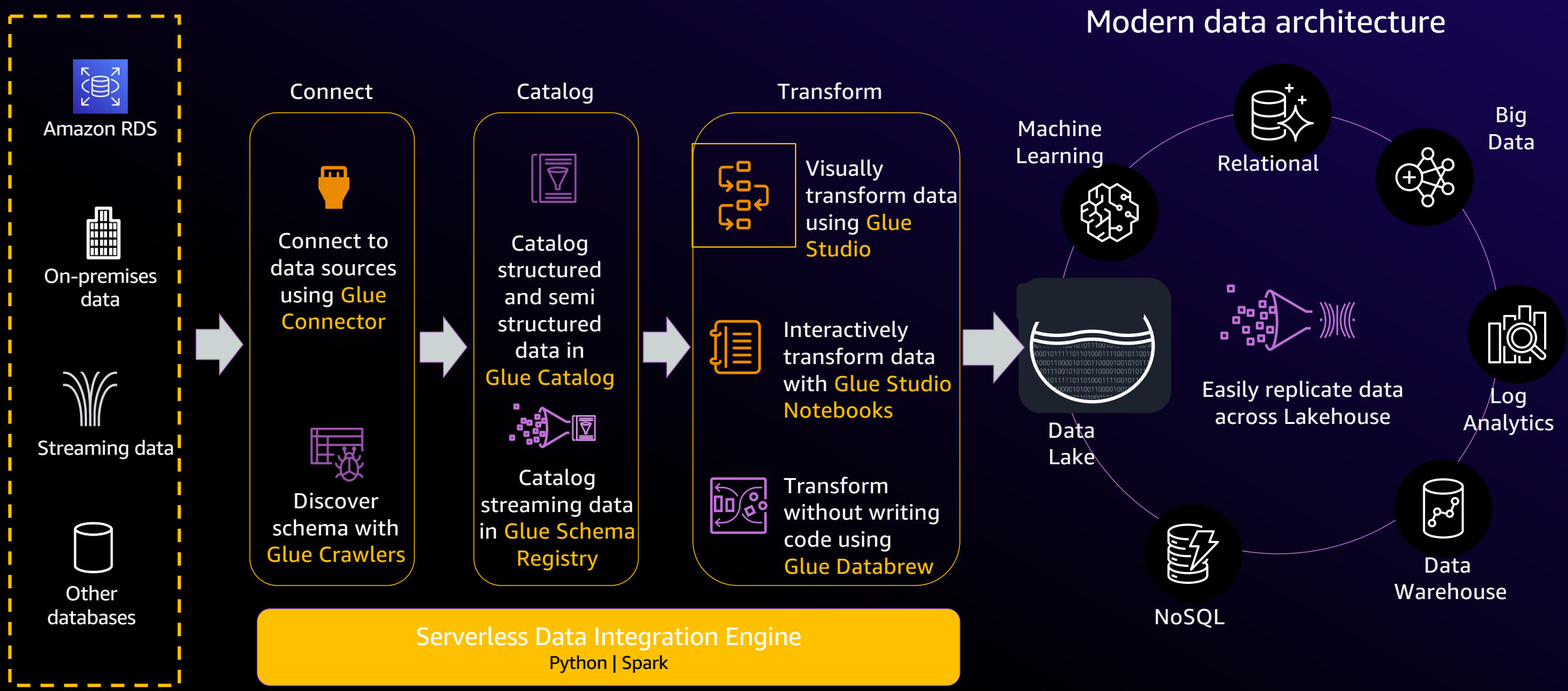
Batch processing

Combine, enrich, and transform **data using AWS Glue Studio**

Event processing

Adapt to **varied** and **changing schemas**

A Data Integration ecosystem for building Lake Houses faster



Optimize data structure for performance

Faster queries at lower cost when you use columnar formats, partitions, and compression

Use Columnar Data

Query performance improves if you convert your data into columnar formats. Use Athena's CTAS feature or use AWS Glue jobs to convert your existing data to a columnar format

Supported Types

- Parquet
- Orc

Partition Data

By partitioning your data, you can restrict the amount of data scanned by each query, thus improving performance and reducing cost

Available Methods

- Partition Projection
- Glue Partition Indexing for multi-engine support
- Bucketing
- Use MSCK repair table command to add partitions for existing data

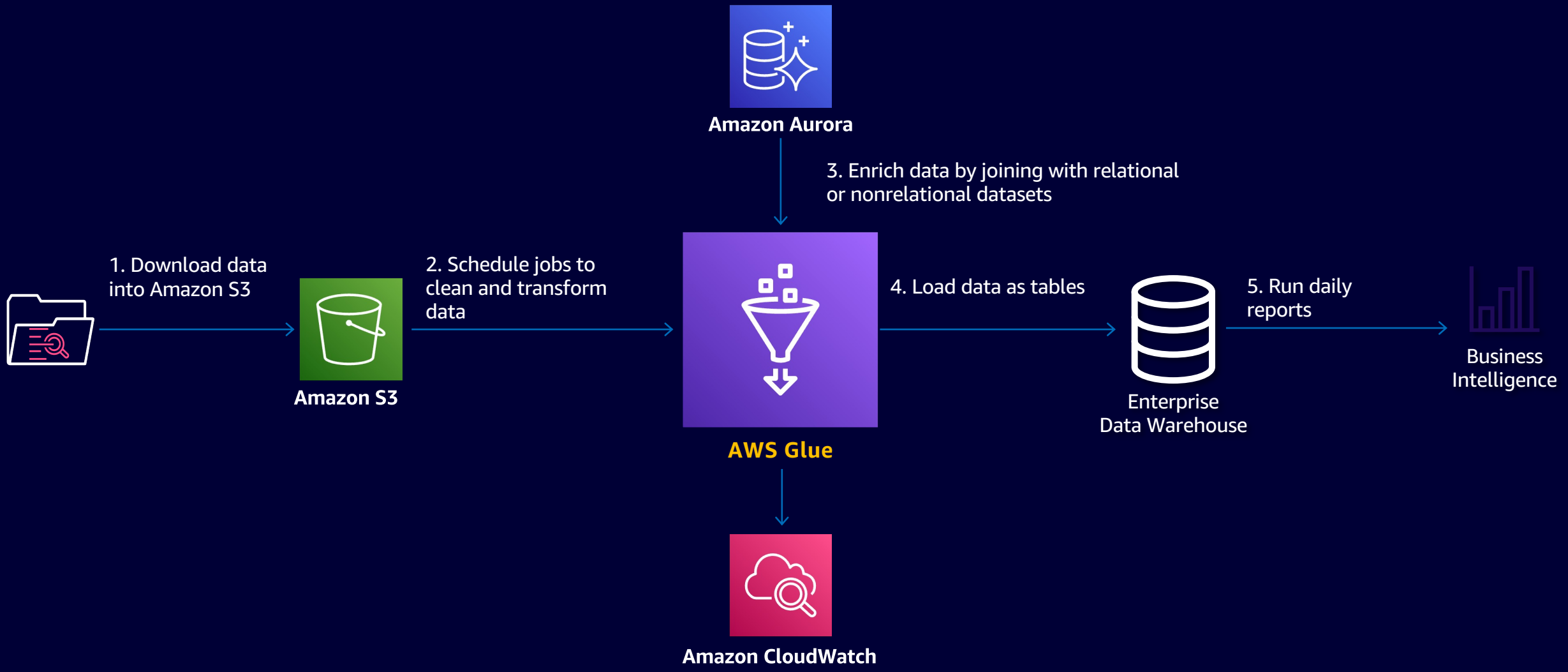
Compression

Athena bills on compressed bytes, so you can save up to 90% on queries that scan data through compression

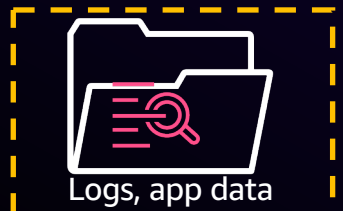
Supported Types

- BZIP2
- DEFLATE
- GZIP
- LZ4
- LZO
- SNAPPY
- ZLIB
- ZSTD

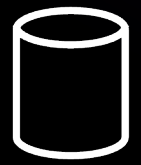
Modernize on-premises ETL tools to load Data Warehouses

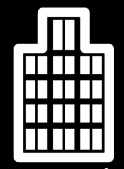


Prepare raw data for Machine Learning


Logs, app data

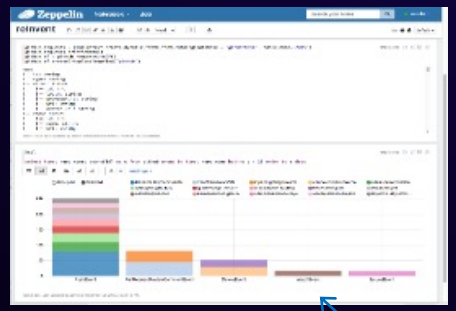

Amazon RDS


Other databases


On-premises data


Streaming data

Notebooks:
data exploration,
experimentation




AWS Glue
Ingest


Raw data


AWS Glue
Transform


Cleaned and enriched data


Glue Catalog


AWS Glue
DataBrew
Transform


Extracted features


Amazon Forecast


Amazon Comprehend


Amazon SageMaker


Amazon Rekognition


Amazon Lex


AWS Glue
DataBrew
Transform


Training data

BMW powers self-service Cloud Data Hub platform with AWS Glue

CHALLENGE

BMW's on-premises Hadoop platform for data ingestion and processing was a hard-to-scale platform, leading to attrition of internal customers

They relied on DevOps to constantly provision clusters, and tuning was time-consuming

They also lacked self-service options for data ingestion, leading to longer data ingestion cycles

SOLUTION

BMW built a self-service Cloud Data Hub platform and used AWS Glue for structured data ingestion

They created self-service features with a customized UI and used AWS Glue APIs to automatically provision AWS Glue jobs for data ingestion

They also built a customized level data catalog on top of the AWS Glue Data Catalog and Catalog APIs

OUTCOME

- ✓ Significant adoption of the Cloud Data Hub data platform with internal users
- ✓ Users self-serve data instead of relying on data engineers, shortening data ingestion cycles

Jack in the Box improves performance and reduces costs with AWS Glue

CHALLENGE

Jack in the Box's data platform built on a competitor's SQL Server had high license costs, and those costs increased as data volumes increased

The proliferation of traditional, on-premises ETL tools led to an unmanageable code base

SOLUTION

Jack in the Box centralized data pipelines in AWS Glue in phases

OUTCOME

- ✓ Reduced baseline software and hosting costs by 50%
- ✓ Improved performance, scalability, and ability to focus on business problems rather than infrastructure challenges

ENGIE builds Common Data Hub on AWS, accelerates zero-carbon transition

CHALLENGE

ENGIE's decentralized global customer base accumulated lots of data, and it required a smarter, unique approach and solution to align its initiatives and efficiently provide data across its global business units

SOLUTION

ENGIE built its Common Data Hub data lake on AWS, allowing the company's business units to collect and analyze data to support a data-driven strategy and lead the zero-carbon transition

OUTCOME

- ✓ Collected 95 TB of data across 351 projects
- ✓ Automated energy predictions
- ✓ Maximized wind farm energy production

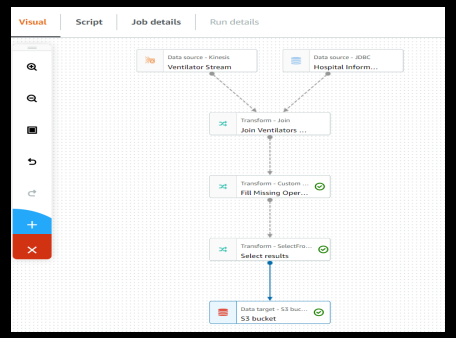
KEY SERVICES: Amazon Kinesis Data Streams, Amazon Redshift, AWS Glue, Amazon Athena, Amazon S3, Amazon SageMaker

Authoring data integration jobs: tools for everyone

Designed for all data users

ETL developer

AWS Glue Studio



Use a rich visual interface

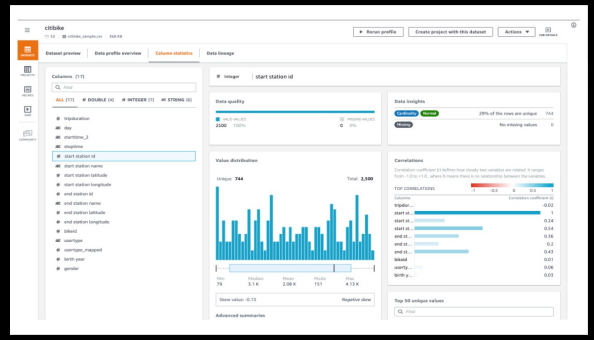
Choose from 250+ built-in transformations

Profile data to understand data patterns and anomalies

Work on large datasets at scale

Business analyst | Data scientist

AWS Glue DataBrew



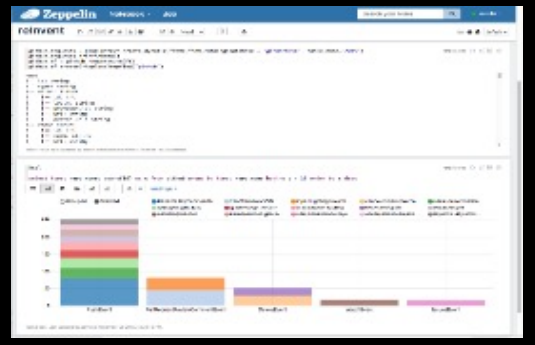
Run ETL jobs without writing code

Monitor thousands of jobs in one place

Apply advanced transforms though code snippets

Data Engineer/Scientist

AWS Glue notebooks



Clean and normalize data with a rich visual interface

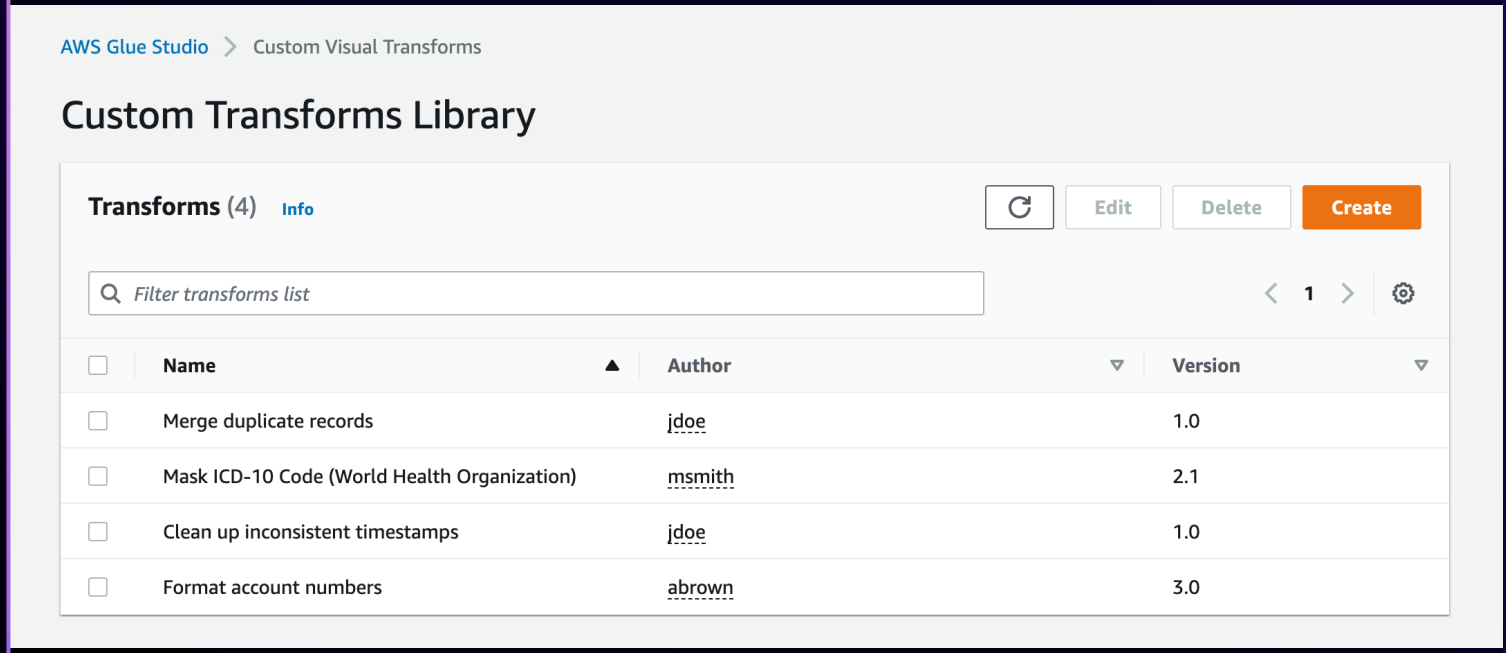
Choose from 250+ built-in transformations to automate tasks

Profile data to understand data patterns and anomalies

Custom visual transforms

(Re)use custom business logic in visual ETL jobs

Share custom transformations between teams



Benefits

- Reduce dependence on Spark developers
- Maintain and update jobs more easily
- Use custom visual transforms in both visual and code-based jobs

NEW

Amazon CodeWhisperer integration with AWS Glue Studio

Automatically generate code to accelerate development of data pipelines

Generate code suggestions in real time and build data integration pipelines faster

Increase productivity of authoring AWS Glue Studio jobs

Scan code for hard-to-find **vulnerabilities**

Get **high-quality suggestions** for AWS services

Interactive sessions and job notebooks

Next-generation interactive data exploration and job development

Time to first query: 10 minutes → **30 seconds**

Use **AWS Glue Studio** or your preferred **IDE or notebook**

Live serverless data integration

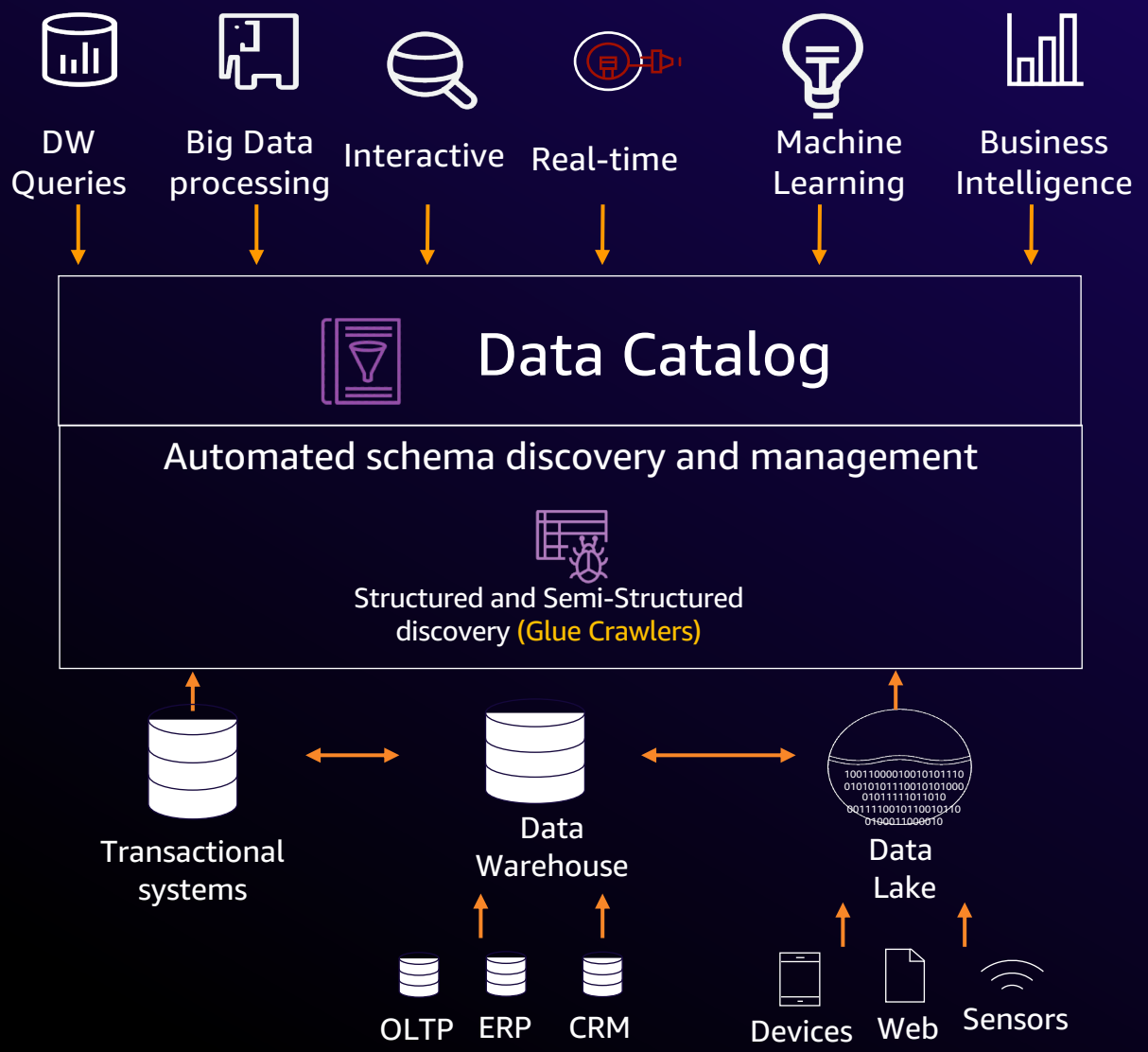
Ephemeral infrastructure with built-in **cost controls**

Dedicated resources for no noisy neighbors

Data management: improving data accessibility and data quality

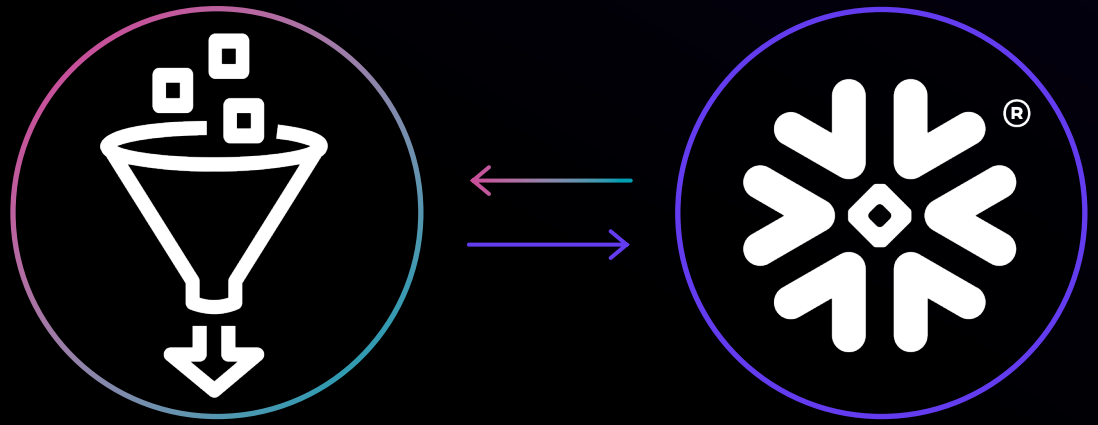
Unified Data Catalog with automated schema discovery

- 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 native connector for Snowflake

NEW



- Out of the box high performance Spark connector
- Flexible authoring using Snowflake SQL
- Visually author common ETL patterns across AWS and Snowflake
- Easily manage price-performance optimized data pipelines

Connectors with AWS Glue

Built-in Connectors

Out of box connectors to support high performance ingestion

Custom Connectors

Flexible method to build your own connectors

Marketplace Connectors

Subscription based low cost connectors

Elasticsearch

JIRA

Amazon S3

amazon REDSHIFT

salesforce

Apache hudi

snowflake

Couchbase

Azure Cosmos DB

SAP HANA

mongoDB

cassandra

Google Big Query

Microsoft Dynamics 365

HIVE

ORACLE SALES CLOUD

Microsoft SQL Server Analysis Services

Amazon Athena

Amazon DynamoDB

splunk>

SharePoint

Google Ads

HubSpot

servicenow

100 + connectors

Native support for Hudi, Delta, and Iceberg

Apache Hudi

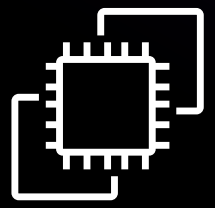
Delta Lake

Apache 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 streamlined deployment
- Latest framework versions available with AWS Glue 4.0

32



AWS Glue Sensitive Data Detection

Define sensitive data

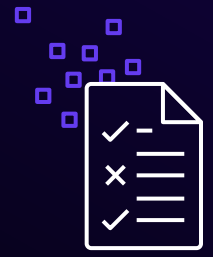
Detect sensitive data at scale

Remediate: delete, redact, replace, or report

Manage data quality in your data lake

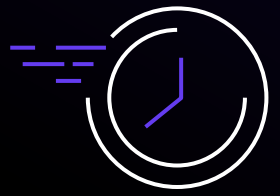
AWS Glue Data Quality

AUTOMATICALLY MEASURE, MONITOR, AND MANAGE DATA QUALITY IN YOUR DATA LAKE



Generate automatic data quality rules

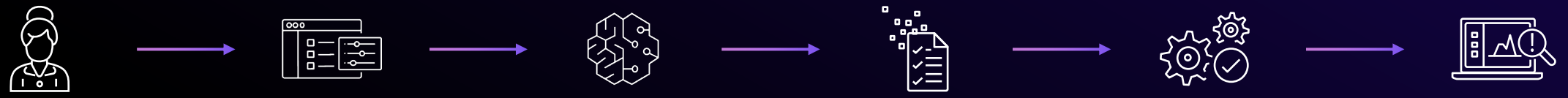
Enhance data quality for better decision-making



Reduce manual efforts from days to hours

AWS Glue Data Quality in the AWS Glue Data Catalog

HOW IT WORKS



Data steward

Selects a dataset in the AWS Glue Data Catalog

AWS Glue Data Quality analyzes data and recommends rules

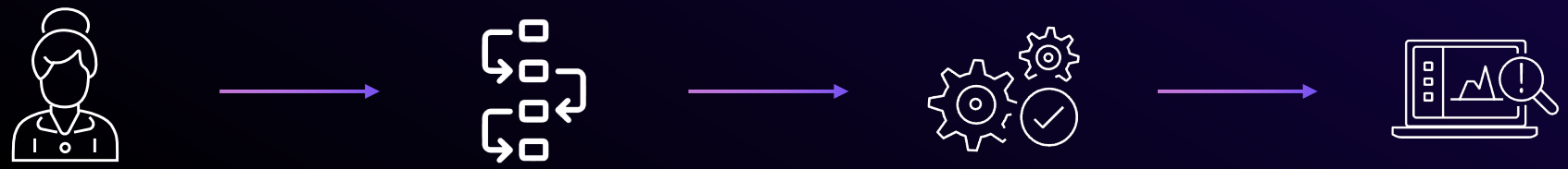
Data steward refines rules to create finalized rules

AWS Glue Data Quality evaluates rules

Data steward reviews results and alerts and takes appropriate action

AWS Glue Data Quality on AWS Glue ETL pipelines

HOW IT WORKS



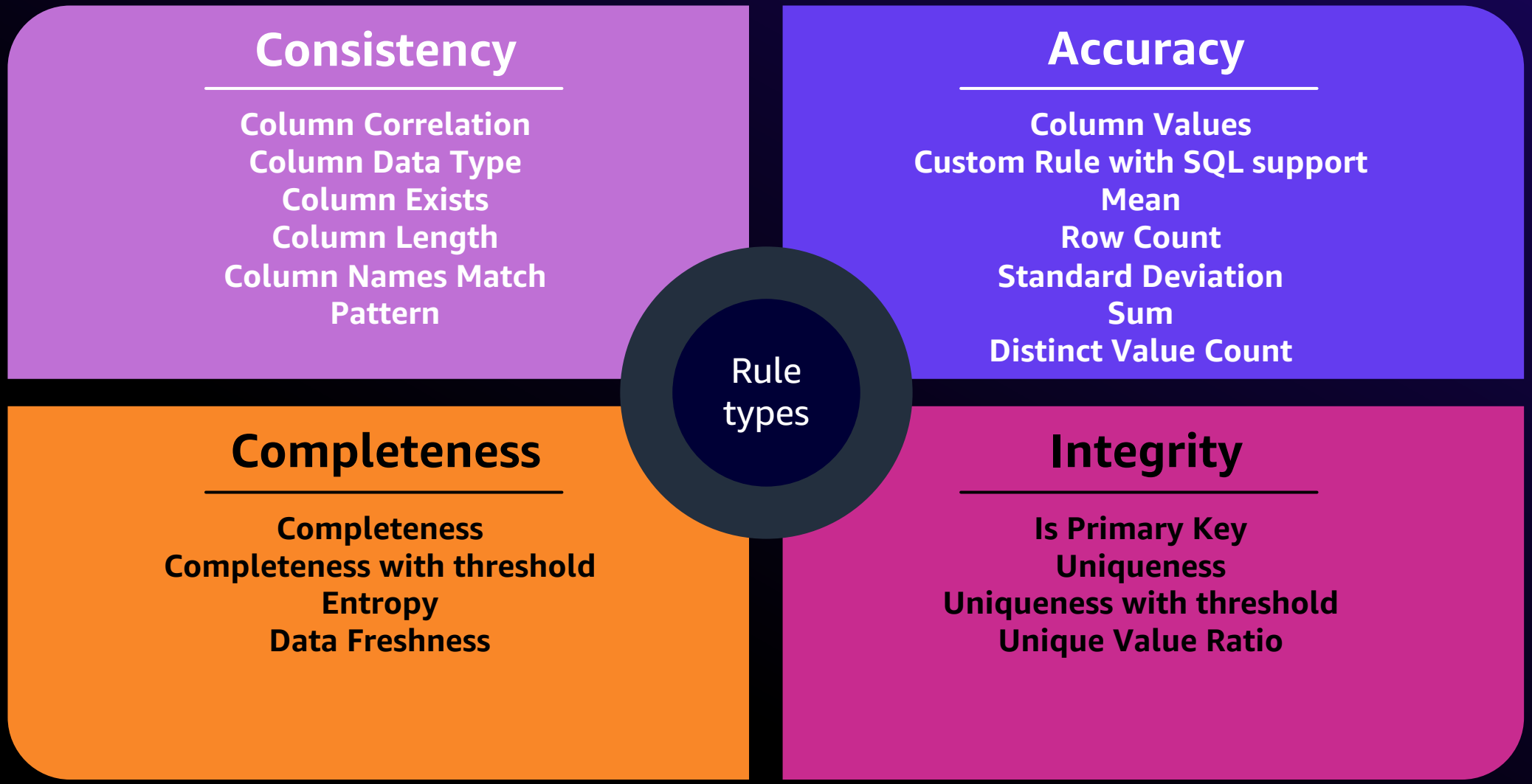
Data engineer

Selects an ETL job and adds AWS Glue Data Quality rules and actions

AWS Glue Data Quality evaluates rules

Engineer reviews results and alerts and takes appropriate action

Built-in rule types



Data integration engines

AWS Glue 4.0

NEW

Fast, predictable, and cost-effective

Upgrades AWS Glue engines

Apache Spark 3.3.0

Python 3.10

Scala 2.1

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

Performance-optimized Spark 3.3

Distributed Pandas API on Spark–improved data processing

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

Data integration engines options

AWS Glue for
Apache Spark

~5 sec startup

Performance-optimized Spark

Spark 3.3.0

Native data lake frameworks, Hudi, Iceberg, Delta Lake

AWS Glue for
Ray

Serverless Ray.io

Scale existing Python code on large datasets

Use familiar Python libraries

AWS Glue for
Python Shell

2x faster start times

Dozens of preloaded libraries

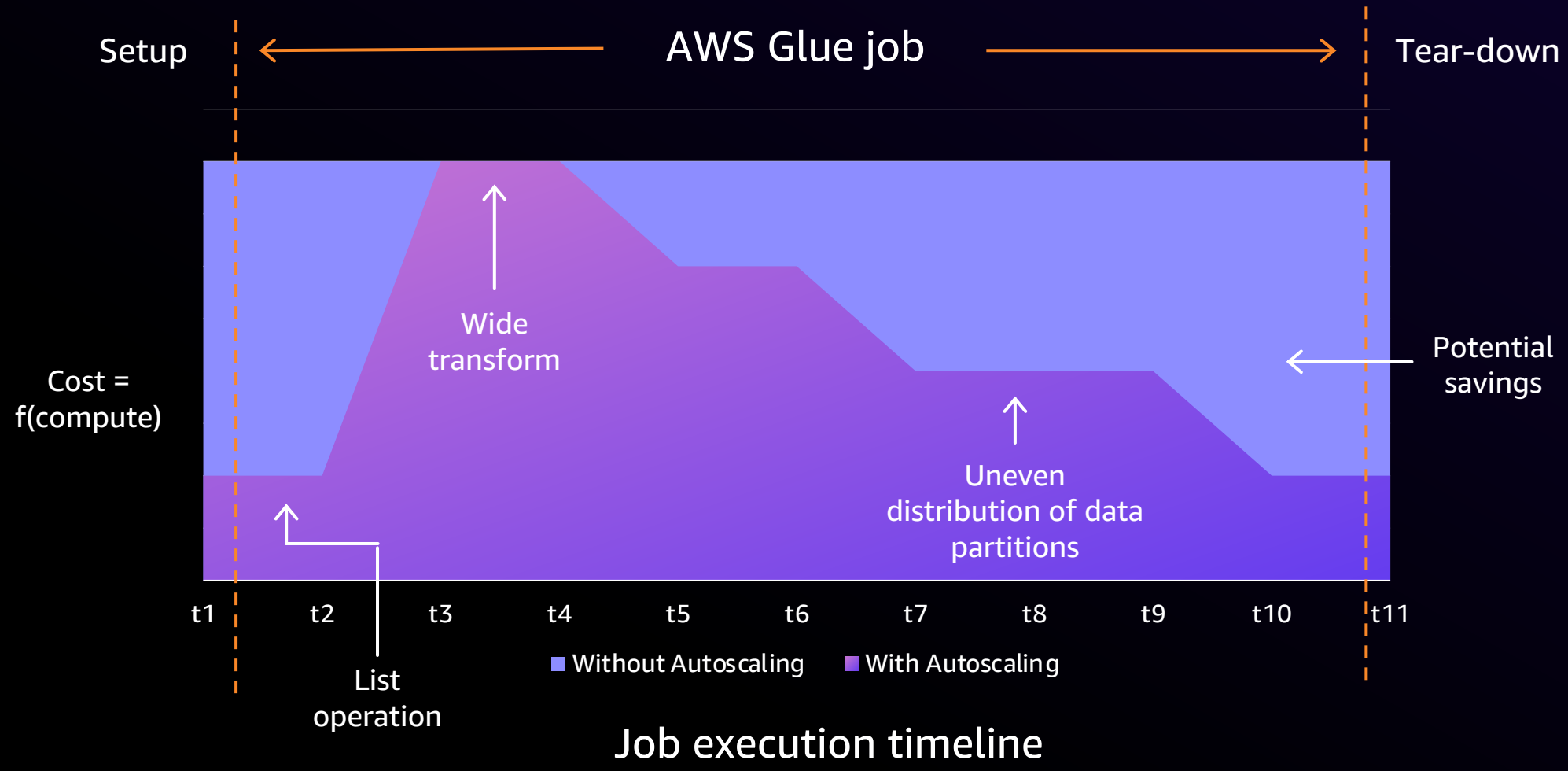
Python 3.10

AWS Glue
Flexible Execution

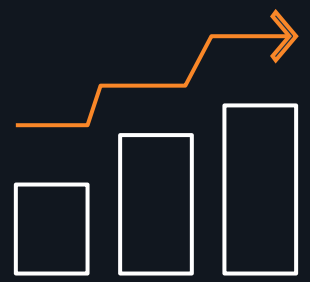
Save up to 35% on nonurgent workloads

Ideal for one-time data loads or nonurgent service-level agreements (SLAs)

AWS Glue Autoscaling



AWS Glue for Ray

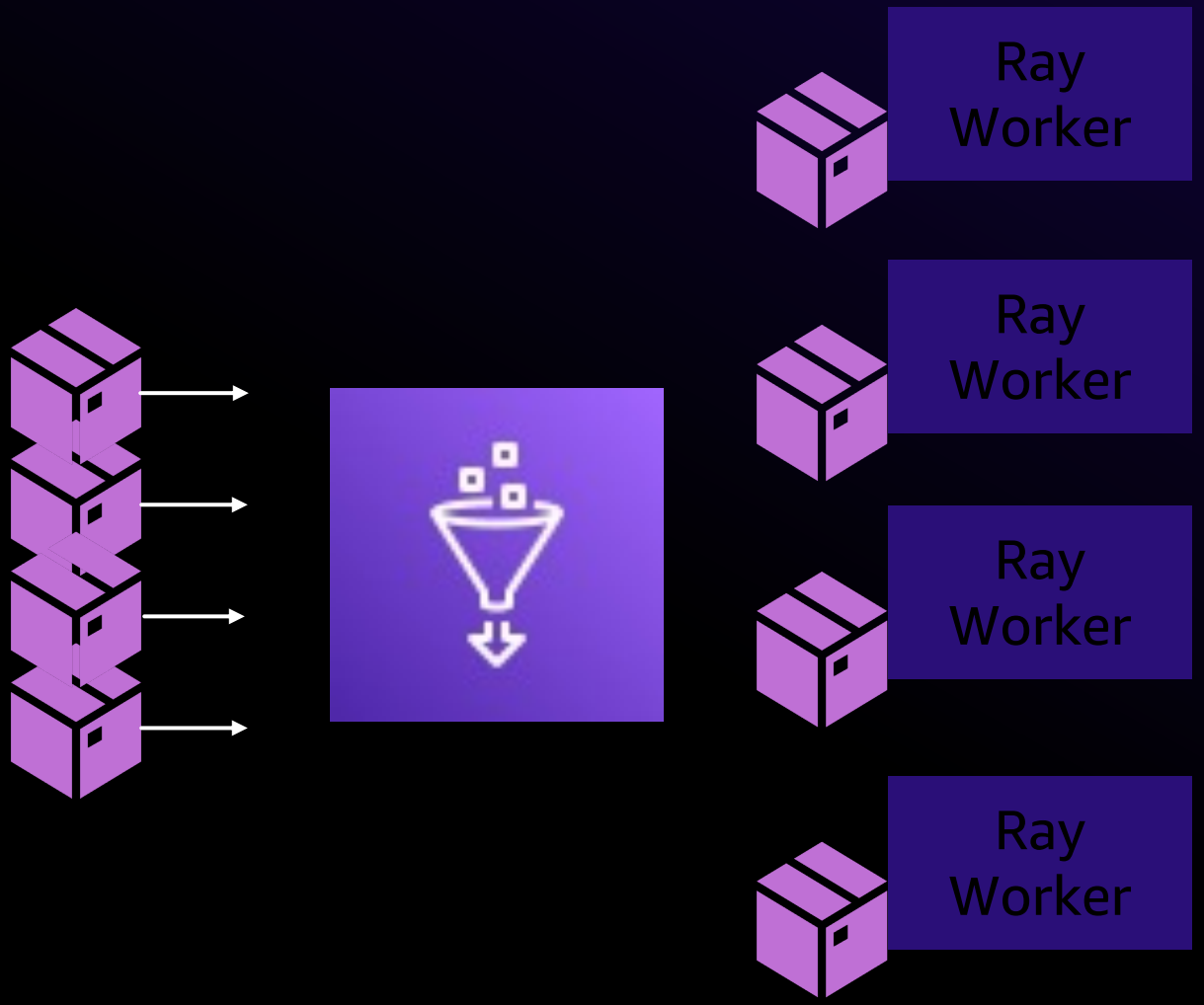


Scalable to hundreds of nodes

Superfast start and scale up/down time

Familiar Python primitives, built-in libraries

How it works: AWS Glue for Ray



- Serverless environment
- Automatic scale up and scale down
- Support batch and interactive workloads
- Jobs start in few seconds and scale up and down in few seconds.
- Customize the environment by using your own libraries and modules
- Author jobs using Sagemaker Studio Notebook or your own local notebook

Glue Flex

New execution option for AWS Glue that allows customers to reduce the costs by up to 35%

Standard
execution-class

10x faster job start times
Predictable job latencies

Enables micro-batching
Latency-sensitive workloads

Flex
execution-class

Up to
35% cost savings

Cost effective for non-time sensitive workloads

Operationalise

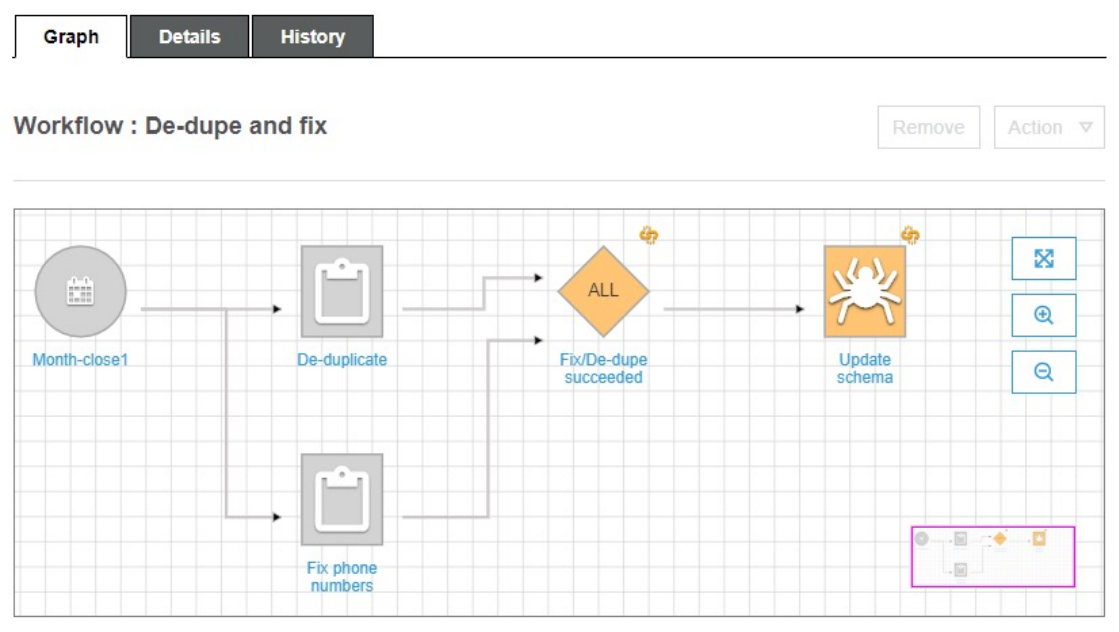
Operationalize and execute data pipelines at-scale

Glue Workflows
Orchestrate Glue jobs by schedule, on demand, or by event

Glue APIs
Programmatic control to build CI/CD pipelines

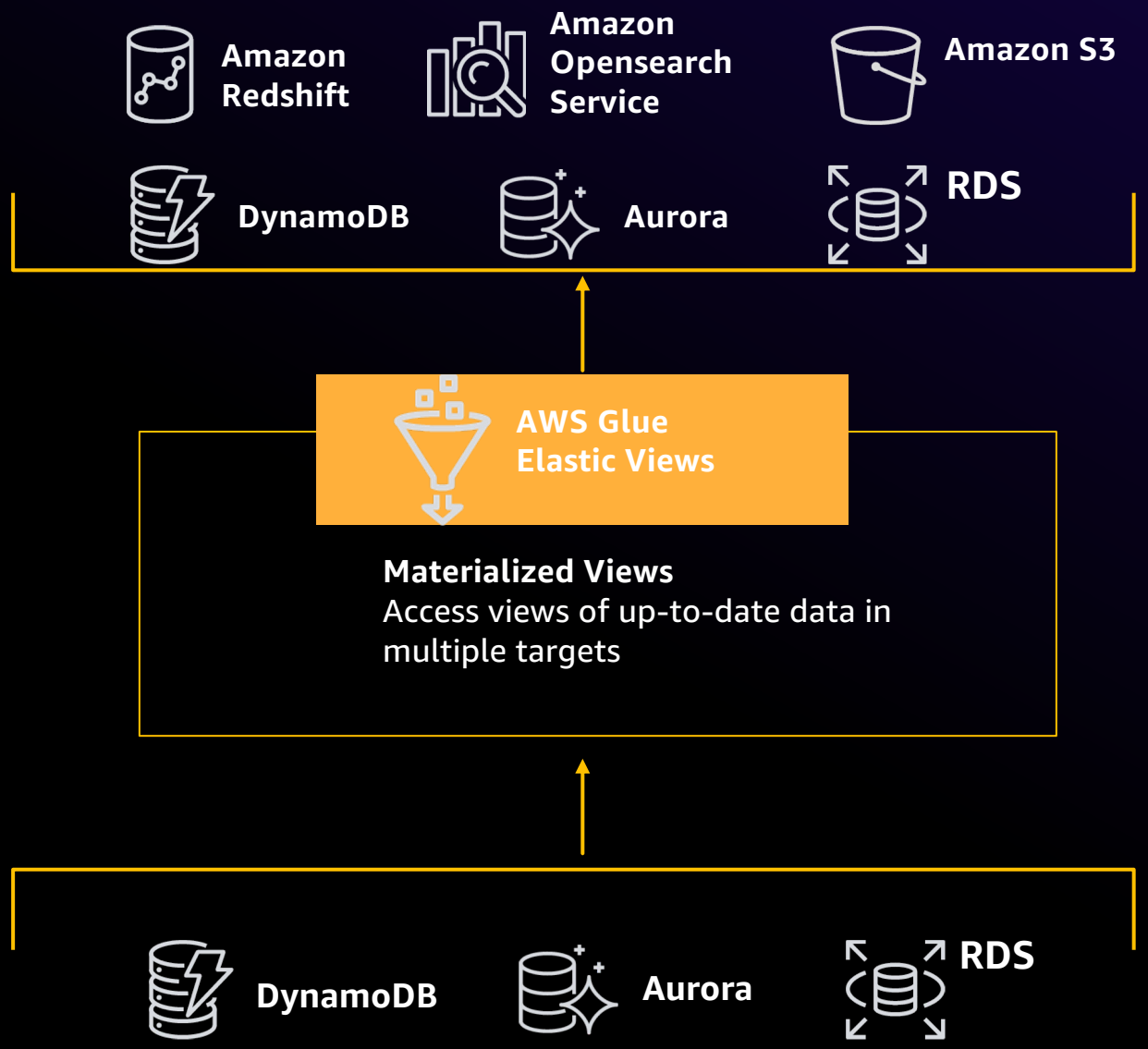
Glue Monitor
Monitor jobs easily

Orchestrate jobs easily with AWS Glue workflows



- Orchestrate Glue jobs and other AWS services
- Schedule jobs or trigger based on events
- Monitor execution of the workflows in one place

AWS Glue Elastic Views for real-time Data Integration



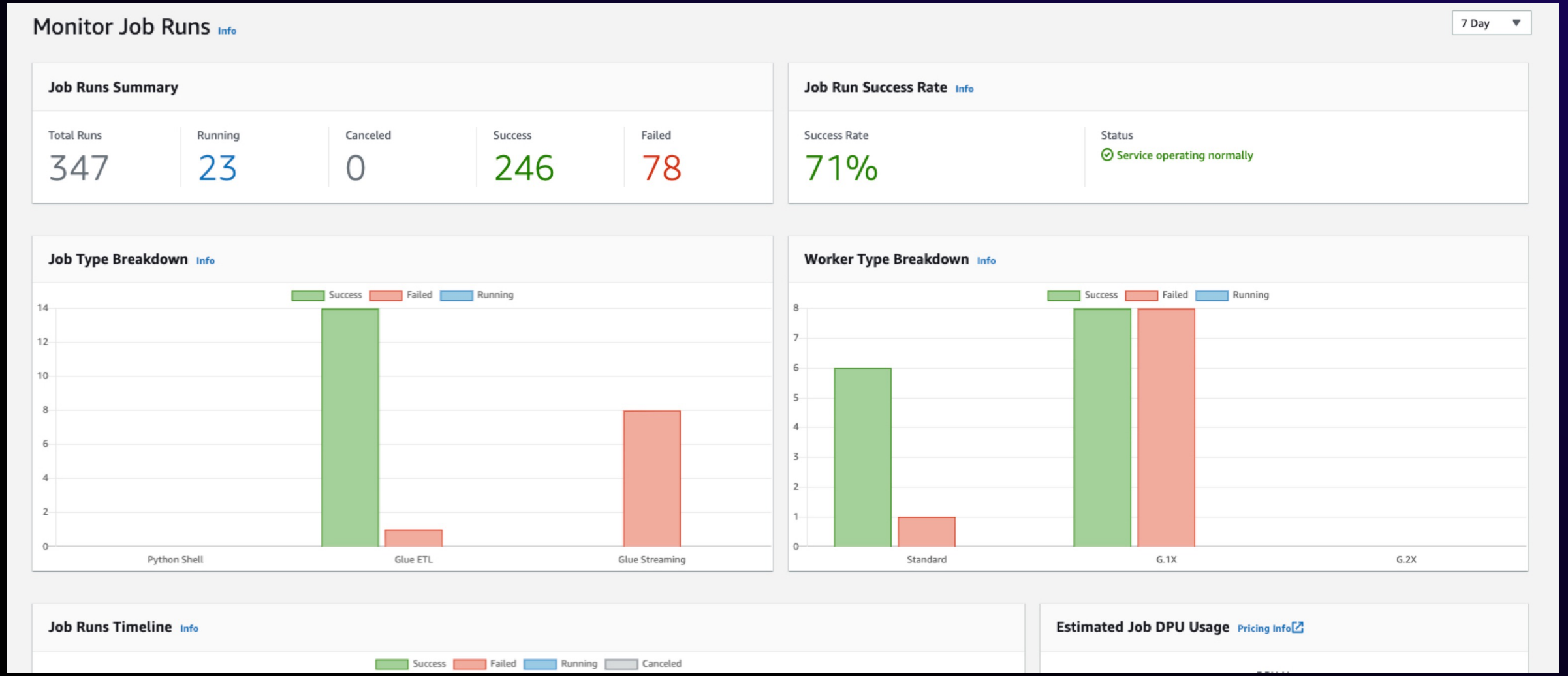
Create materialized views across a wide variety of databases and data stores using familiar SQL

Continually monitors source databases for changes and updates targets within seconds

Serverless and automatically scales capacity up and down to accommodate your workloads

Handles the heavy lifting of copying and combining data without requiring custom code

Monitoring dashboard to check job status



Better together: Faster time to value with Amazon Redshift and AWS Glue

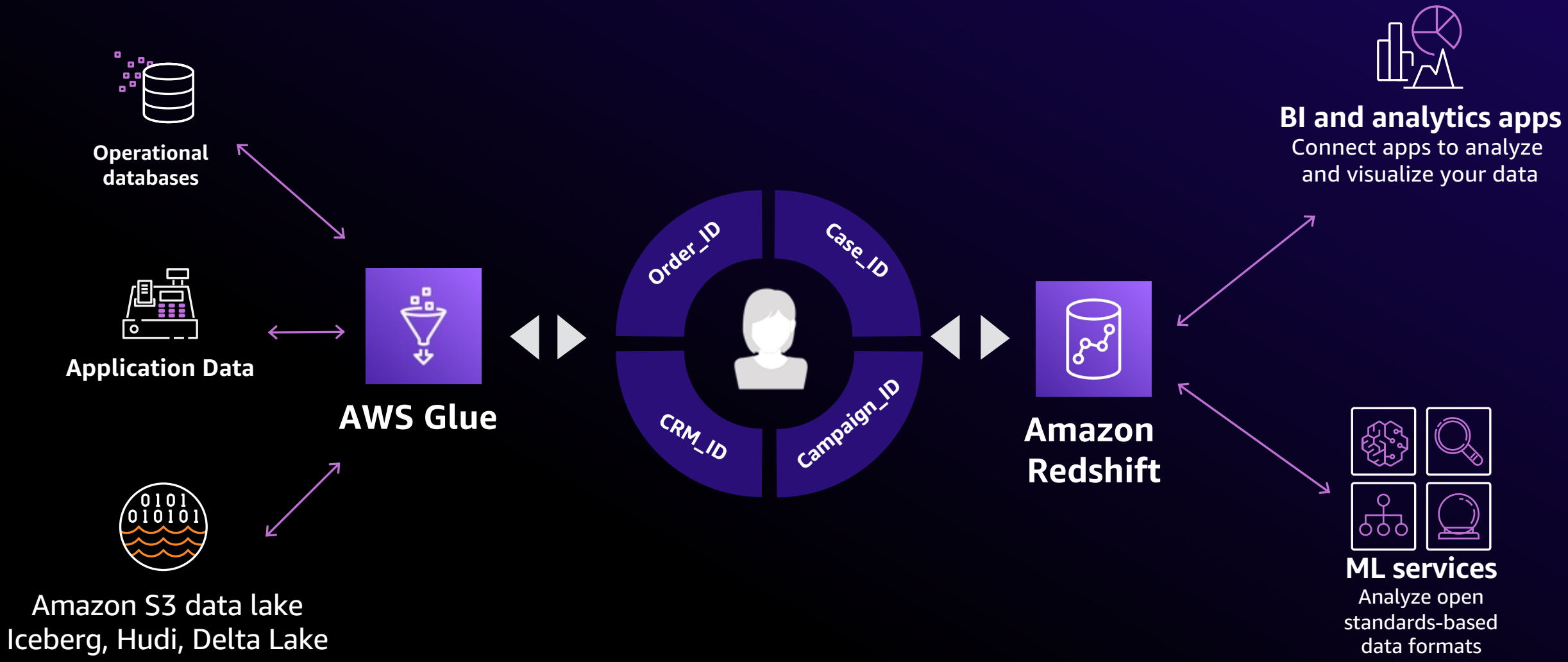
Redshift & Glue

Better together

- Access Redshift in seconds by directly browsing Redshift schemas and tables from the Glue Studio interface.
- Enable ELT workloads by reusing SQL queries to create a custom source
- Simplify common data loading operations into Redshift through new support for APPEND (insert), TRUNCATE, DROP and MERGE commands
- Native high performance Redshift-Spark connector – automatic ELT/ETL optimization

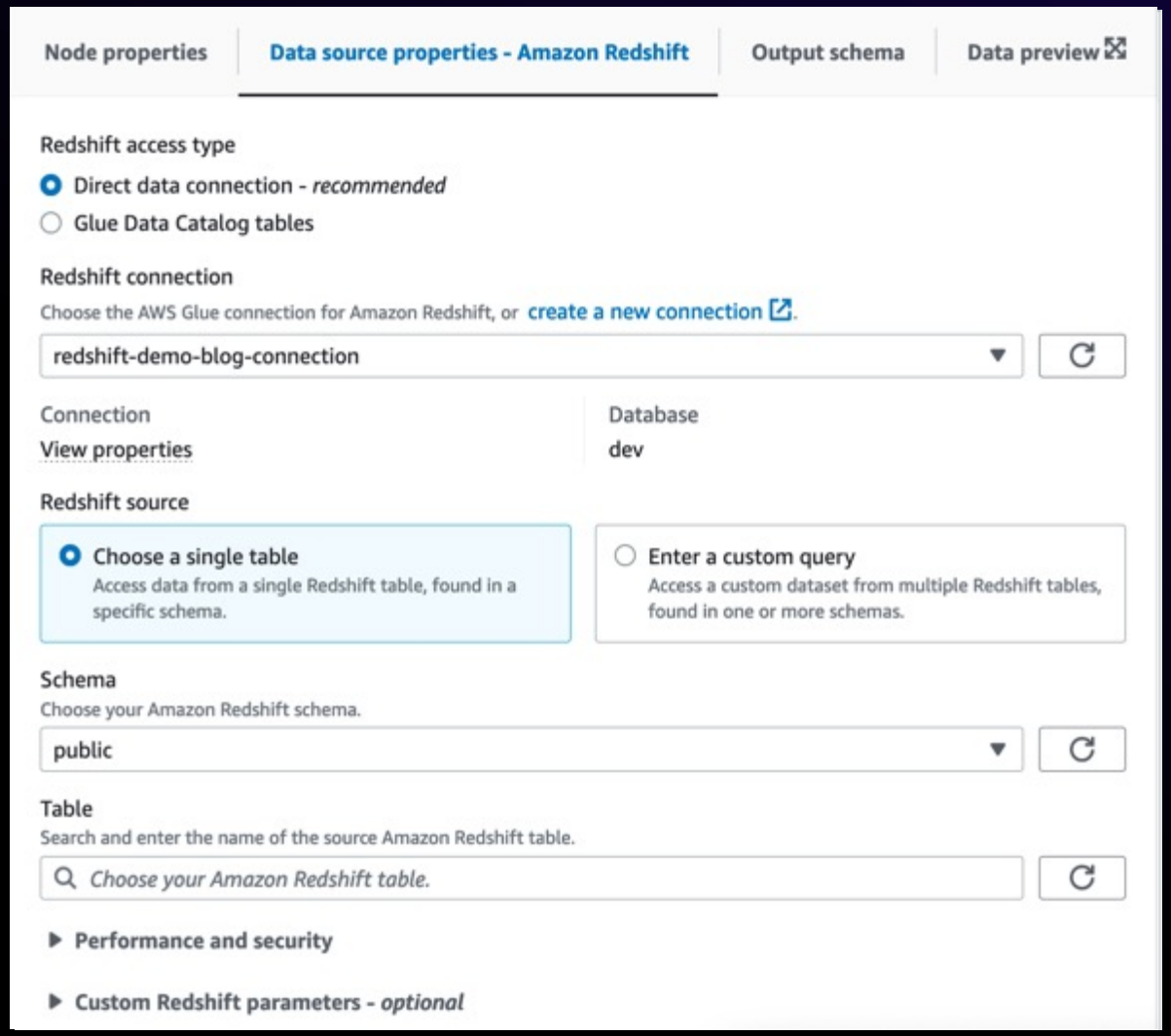
Customer 360: Unify and Enrich Customer Data

With a no-code, visual experience and modern data architecture



Redshift & Glue – Better Together

Source: Get started quickly through direct access



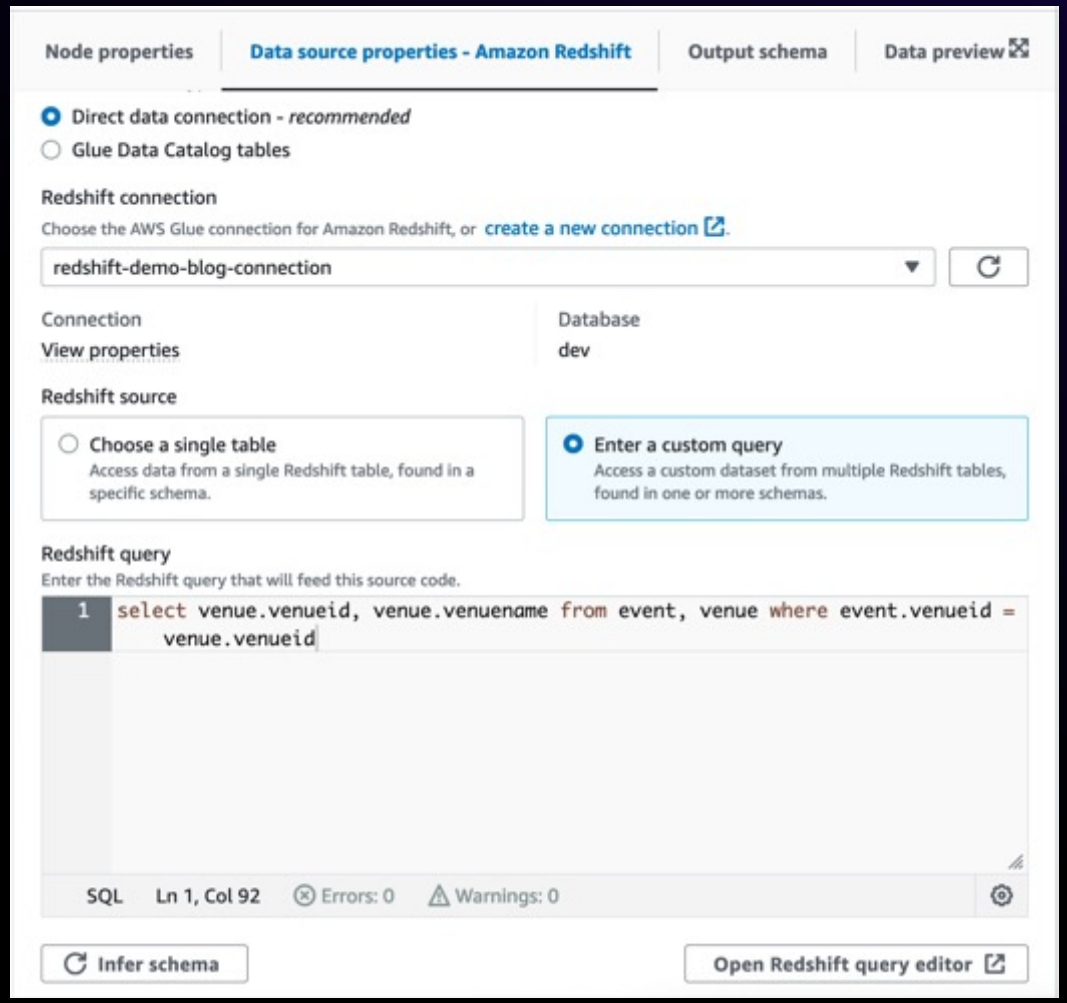
New: DIRECT DATA CONNECTION

Browse Redshift schema and tables from Glue Studio

Auto-populated defaults for performance and security

Redshift & Glue – Better Together

Source: Customize your source using a SQL query



Choose custom SQL query as a source

Access Redshift views or execute joins for a dynamic data source

ELT execution of complex Redshift SQL

Launch Redshift query editor from Glue Studio to copy & paste SQL

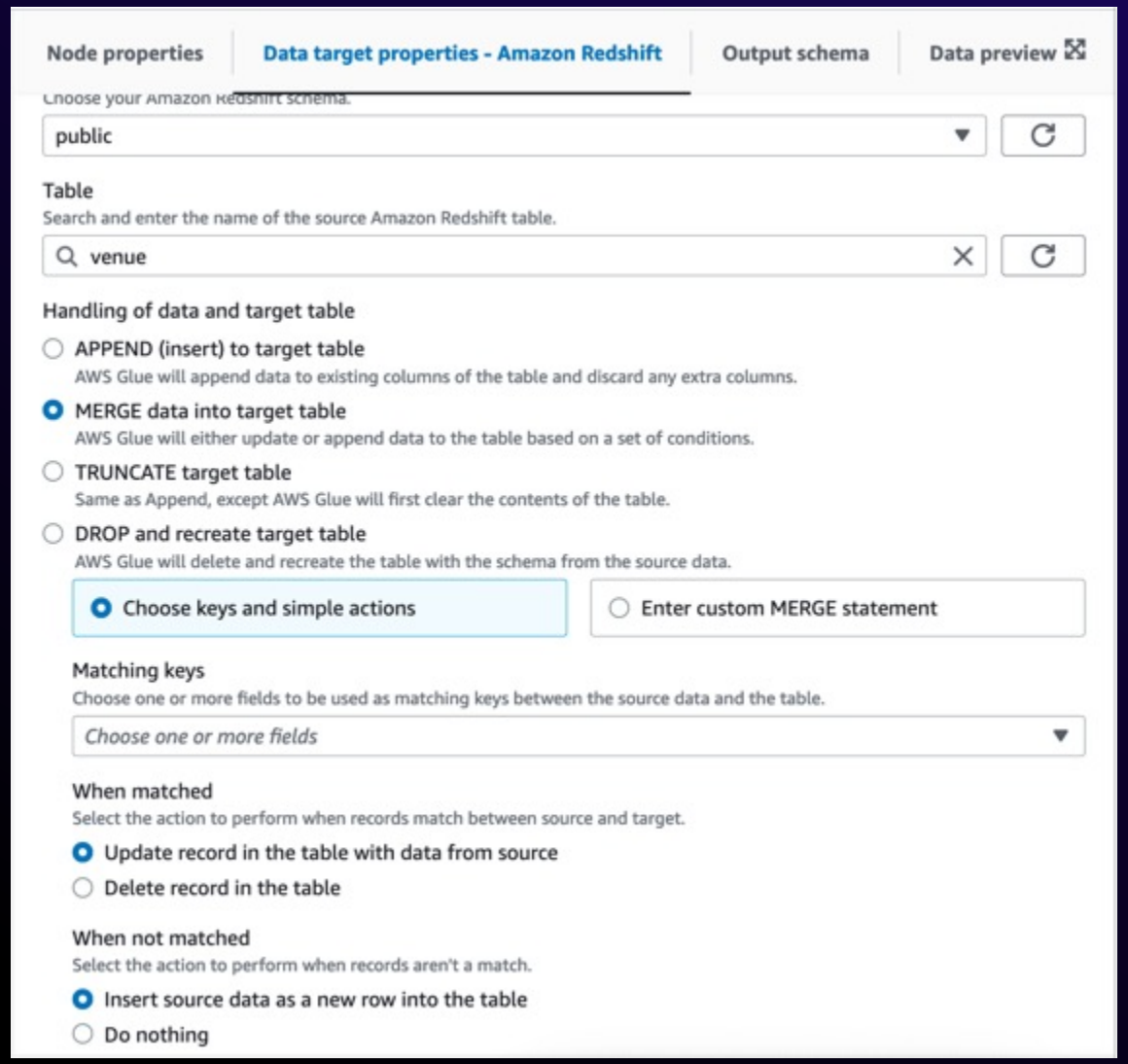
Redshift & Glue – Better Together

Target: Simplify common tasks

Select existing table or create new

Choose between common target actions like APPEND or MERGE * - executed as part of every job run

Use Glue Studio's built in options or enter your own custom



ANNOUNCING

Latest data lake performance enhancements

DELIVERING IMPROVED PERFORMANCE FOR DATA LAKE QUERIES IN AMAZON REDSHIFT

Amazon Redshift Serverless data lake performance (lower is better)

Benchmark derived from TPC-DS 3TB, all data in partitioned Parquet tables

- Up to 45% performance improvement for data lake queries on Redshift Serverless
- Leverages AWS Glue column statistics to further optimize data lake queries
- Incremental refresh support for materialized views on data lake tables (Preview) to eliminate the need for re-scanning already materialized data

New Features

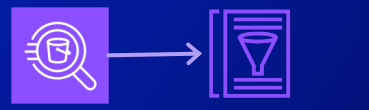
ANNOUNCING PREVIEW

AWS Glue Data Catalog views

CREATE ONE VIEW DEFINITION AND SECURE ONCE ACROSS ENGINES

1

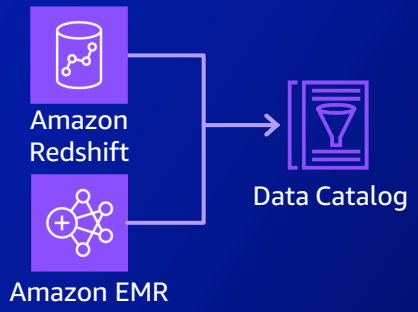
Create your view, stored in AWS Glue Data Catalog




```
CREATE VIEW lf_view AS
SELECT … FROM … WHERE
```

2

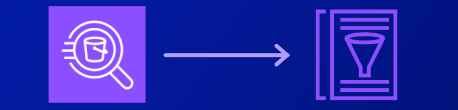
Add dialects




```
ALTER VIEW lf_view ADD
DIALECT
SELECT … FROM … WHERE
```

3

Query data




```
SELECT * FROM lf_view
```

- Query consistently across Spark on Amazon EMR on EC2, Amazon Redshift, and Amazon Athena
- No access required to underlying S3 base tables
- Views can be shared across account and linked across Region

GENERALLY AVAILABLE

Anomaly detection and dynamic rules in **AWS Glue Data Quality**

ML-powered anomaly detection algorithms to detect hard-to-find data quality issues and anomalies

Detect anomalies without data quality rules

Detection of anomalies that can be indicative of an unintended event, seasonality, or statistical abnormality

Catch potential problems in your data that data quality rules alone can't find

Automate tasks that evolve over time, such as limiting the number of rows ingested for data quality monitoring

GENERALLY AVAILABLE

Automatic compaction for Apache Iceberg tables in the **AWS Glue Data Catalog**

Built in maintenance operations to optimize query performance in your data lakes

Maintain a high performance data lake

Reduce metadata overhead and improve query performance

No need to build bespoke maintenance processes to optimise object size

GENERALLY AVAILABLE

AWS Glue Data Catalog now supports generating column-level statistics

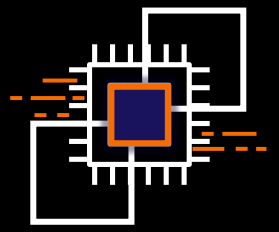
Improved performance when querying your data lake using Amazon Athena

Integrated with Amazon Redshift Spectrum for better performance when integrating with your data warehouse

Reduces resource utilisation and minimises cost

Available now

Amazon S3 Express One Zone storage class



For compute-intensive workloads

Single-digit millisecond latency

One zone S3 storage class that delivers the fastest data access speed and highest performance of any cloud object storage for customers' most latency-sensitive applications

Most frequently accessed data

Designed for **request intensive applications** – ML training and inference, interactive analytics, media content creation

10x faster + 50% lower request costs

Data access speeds up to **10x faster**, and request costs up to **50% lower** than S3 Standard. Fully elastic with no storage provisioning

AWS Glue: Summary of re:Invent 2023 announcements

IMPROVING DATA INTEGRATION AND DATA QUALITY

- AWS Glue Data Quality announces anomaly detection and dynamic rules
- AWS Glue Data Catalog supports multi engine views with AWS Analytics Engines
- AWS Glue announces entity-level actions to manage sensitive data
- AWS Glue Data Catalog now supports generating column-level statistics
- AWS Glue Data Catalog supports automatic compaction for Apache Iceberg tables
- AWS Glue launches native connectivity to 6 databases
- AWS Glue for Apache Spark announces native connectivity for Amazon OpenSearch Service
- Glue Studio Visual now supports interactive data previews

Simplifying ETL migration

Partner-built converters to migrate from legacy ETL tools

"BladeBridge ETL converters use AWS Glue Studio APIs to convert legacy ETL code to AWS Glue Studio objects that can be **easily managed and maintained by non-Spark developers**."

Troy Clemente, Executive VP & Co-founder, BladeBridge

"Impetus LeapLogic is the leading migration automation solution for ETL, EDW, and Hadoop. We provide a modern, efficient, and fast approach to **migrate to AWS Glue** and power up a unified data platform that **better enables data scientists to provide value through data**."

Chetan Kalanki, VP of Global Engineering Services, Impetus

Migrate your ETL logic

ETL Migration Program: Get help migrating your ETL processes to AWS Glue

Converters available for

Informatica

Talend

SAS

Microsoft SSIS

IBM DataStage

Dozens of SI Partners help migrate your ETL pipelines

AGILISIUM
BIG ON CLOUD. BIG ON DATA.

mission™
CLOUD SERVICES

quantiphi
Solving What Matters

slalom

wavicle
DATA SOLUTIONS

rackspace
technology

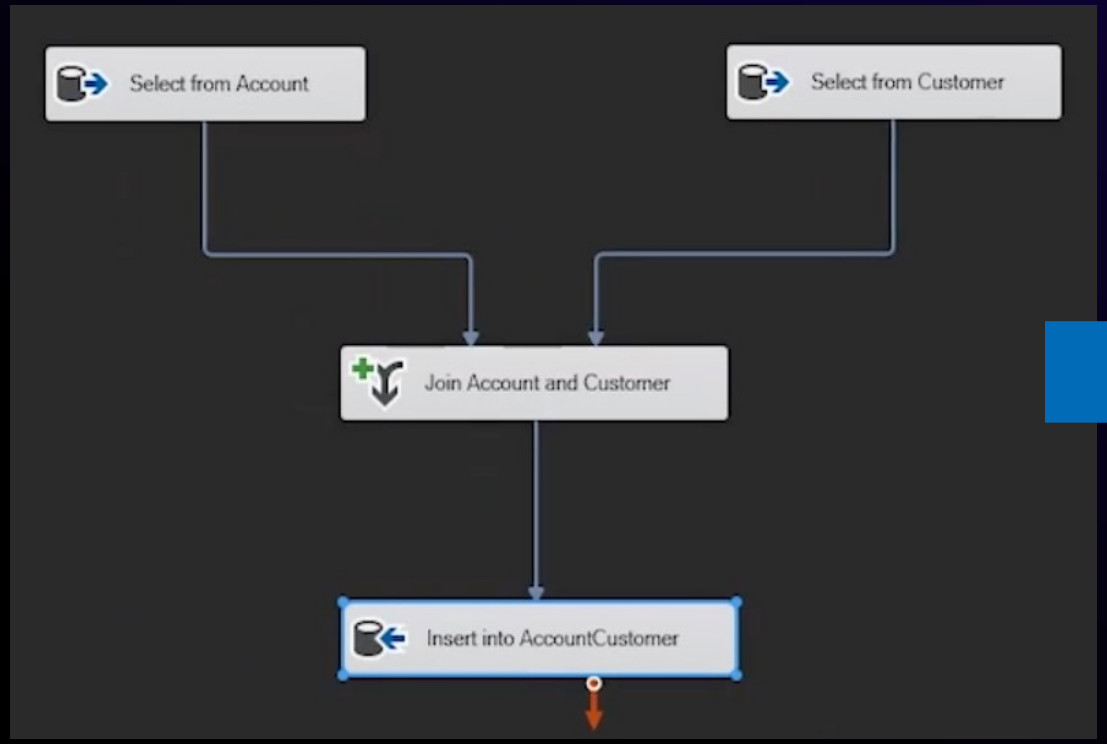
TEKsystems
Global Services

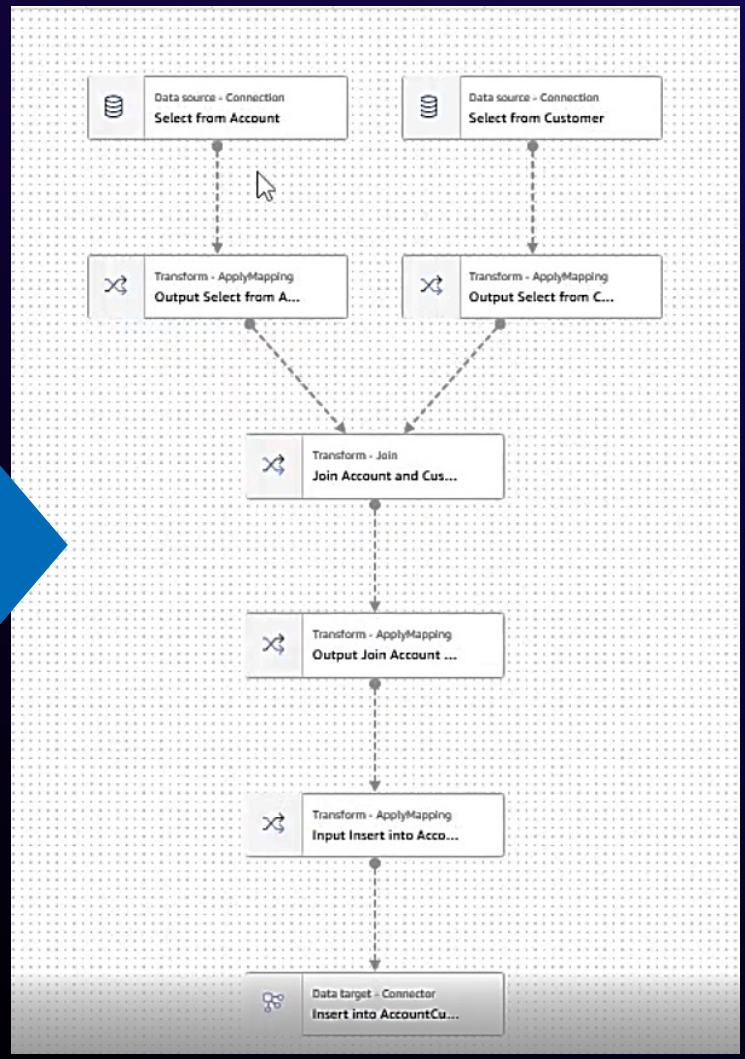
This is not a complete list. To view all AWS Partners for this category, visit AWS Partner Solutions Finder. This list of partners is current as of November 27, 2022.

Migrate your ETL logic

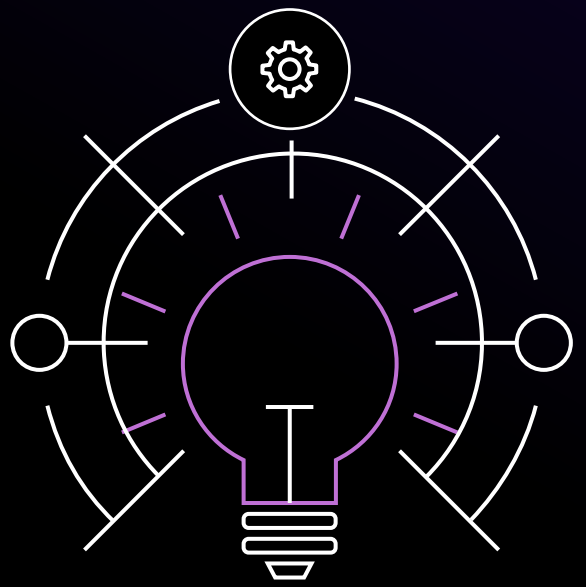
Automated translation tools convert 80% of your ETL jobs

Example: Conversion from SSIS using AWS Schema Conversion Tool





ETL Modernization Program



Goals

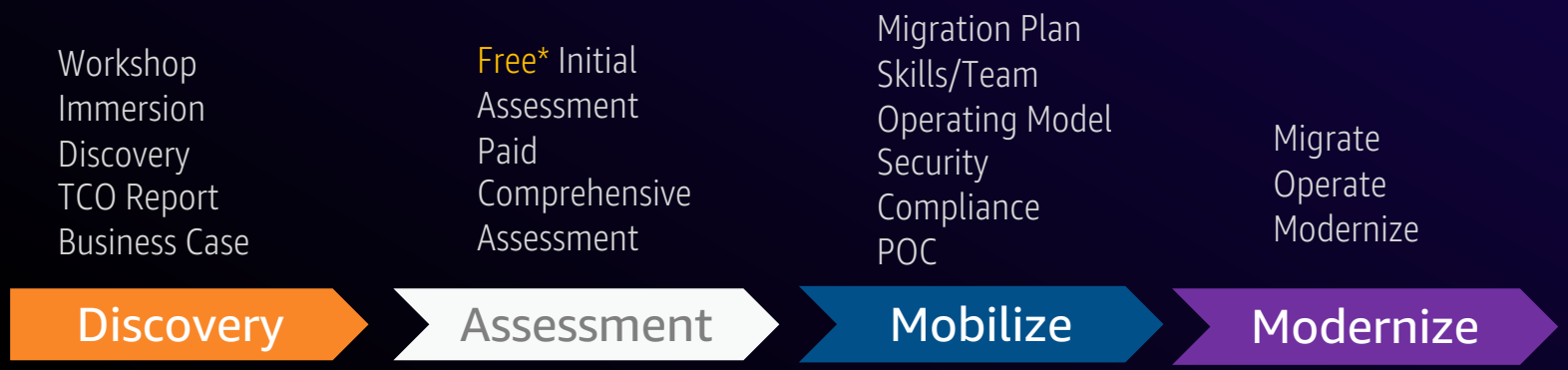
- Modernize legacy ETL workloads
- Reduce data integration costs by migrating to AWS Glue
- Accelerate migration

Benefits

- No-cost two day workshop on AWS Glue
- No-cost AWS partner delivered assessment and Proof-of-Concept (POC) for up to 5 ETL jobs using automated tools*
- No-cost ETL workload analysis and Total Cost of Ownership (TCO)
- High level architecture and roadmap for migration
- Funding options to offset migration to AWS
- Support through the migration

* Automation available for selected ETL tools like Informatica, Talend, SSIS, PL/SQL

ETL Migration Partner Program – Consulting Partners



Tooling ISV's

next PATHWAY

blade bridge

IMPETUS

Some of the participating SI's…

accenture

TATA CONSULTANCY SERVICES

rackspace

TEKsystems An Allegis Group Company

IBM

virtusa Accelerating Business Outcomes

Infosys

bitwise

epam

wipro

WAVICLE Data Solutions

Cognizant

quantiphi

softserve

compasso UOL

IMPETUS

AGILISIUM BIG ON CLOUD. BIG ON DATA.

mission AWS Premier Consulting Partner

Next Steps

AWS Glue Pricing

Only pay for the resources you use, with no lock in

Batch and streaming jobs

- $0.44 per DPU hour , billed per second, 1 minute minimum
- $0.29 per DPU hour for Flex
- Same price for Data Quality tasks

DataBrew

- Jobs - $0.48 per node hour (or portion)
- Interactive sessions - $1.00 per session

AWS Glue data catalog

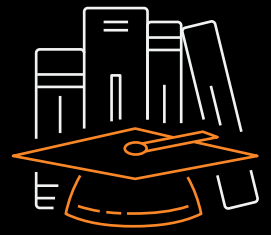
- Storage - first 1m objects free, then $1.00 per 100,000 objects
- Requests - first 1m free, then $1.00 per million
- Crawlers - $0.44 per DPU hour, billed per second, 10 minute minimum

Resources: Videos and tutorials

BLOGS

[AWS Glue Blogs]

DOCUMENTATION



[Product Documentation]

ONLINE WORKSHOPS

[Workshops]

AWS programs to help you get started

Want to build a data vision and strategy?

aws data-driven everything

- Joint engagements with business and technology stakeholders
- Create an organizational vision for innovation with data to drive business outcomes
- Define the first pilot, learn, and build

Gain business and IT strategic alignment

Need help from strategy to implementation?

aws ProServe

- Assess your needs; align data and business strategies
- Work closely with AWS experts to build your data governance framework
- Implement at scale to drive business outcomes

Build your data governance framework

Thank you