AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

ANT317

Serverless data integration with AWS Glue

Santosh Chandrachood
General Manager
AWS Glue

Josef Viehhauser
Product Owner
BMW Group

Simon Kern
Lead Engineer
BMW Group

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What to expect

01

Data integration

What is data integration, why it's important, and why it's hard

02

Key innovations

How AWS Glue helps with faster, easier, and cost-effective data integration

03

Customer journey

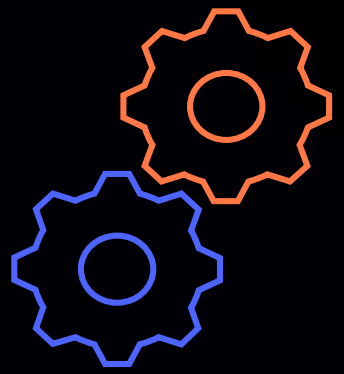
How BMW built a petabyte-scale data lake using AWS Glue

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you! You have helped us grow



Hundreds of thousands of customers



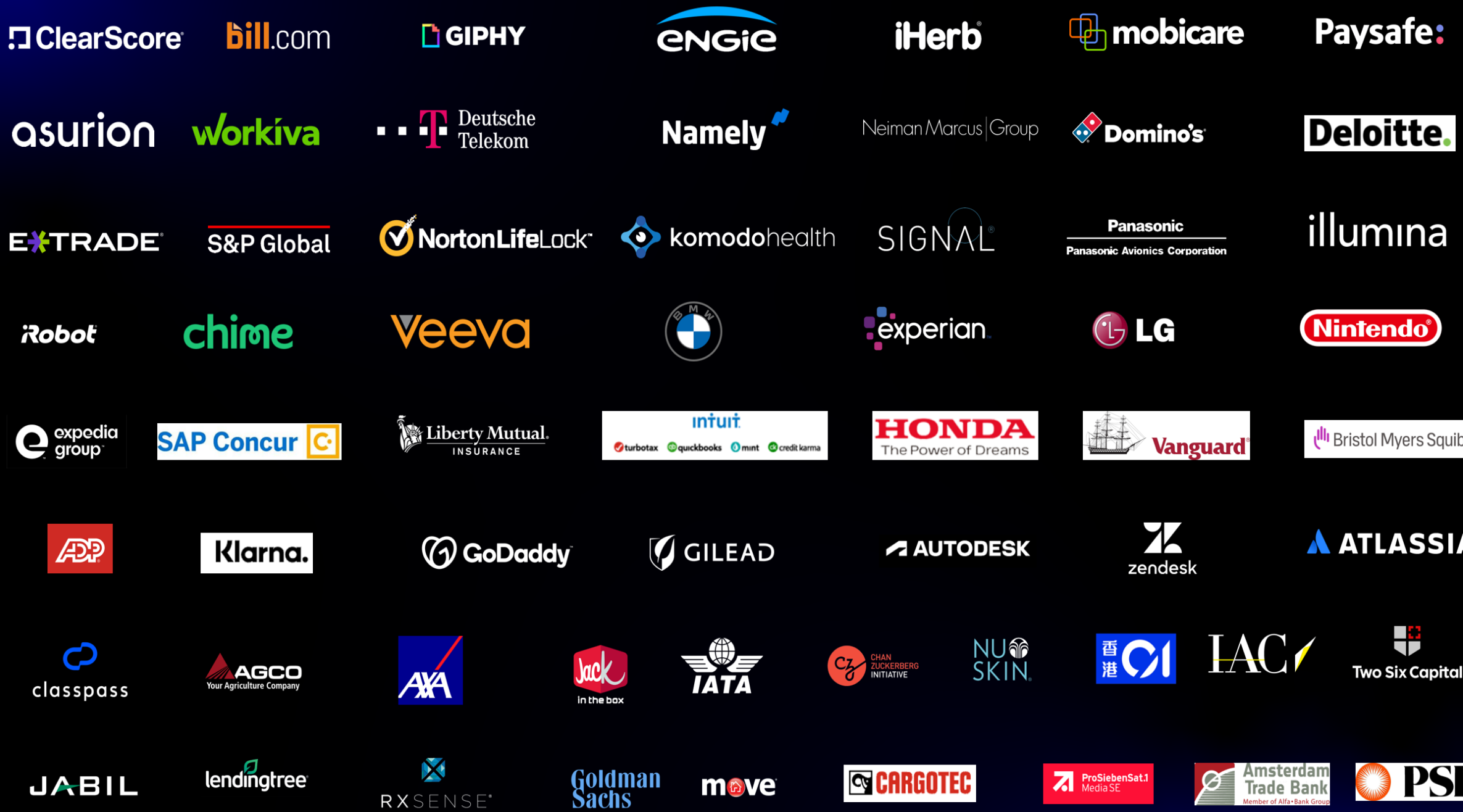
Hundreds of millions of integration jobs



25 AWS Regions

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.


ClearScore
bill.com
GIPHY
ENGIE
iHerb
mobicare
Paysafe:
asurion
Workiva
Deutsche Telekom
Namely
Neiman Marcus | Group
Domino's
Deloitte.
E*TRADE
S&P Global
NortonLifeLock
komodohealth
SIGNAL
Panasonic
Panasonic Avionics Corporation
illumina
iRobot
chime
Veeva
BMW
experian
LG
Nintendo
expedia group
SAP Concur
Liberty Mutual INSURANCE
intuit
turbotax
quickbooks
mint
credit karma
HONDA
The Power of Dreams
Vanguard
Bristol Myers Squibb
ADP
Klarna.
GoDaddy
GILEAD
AUTODESK
zendesk
ATLASSIAN
classpass
AGCO
Your Agriculture Company
AXA
Jack in the box
IATA
CHAN ZUCKERBERG INITIATIVE
NU SKIN
香港
IAC
Two Six Capital
JABIL
lendingtree
RXSENSE
Goldman Sachs
move
CARGOTEC
ProSiebenSat.1 Media SE
Amsterdam Trade Bank
Member of Alfa-Bank Group
PSEG

"[AWS Glue] gave us great advantage over what we were doing before, so we're able to ingest data faster and transform it and get faster insights."
Brian Zellner, Assoc. Director, R&D Data Lakes
Bristol Myers Squibb





"AWS Glue is a cost-effective serverless Data Integration service for Data Lake building that CargoTec trusts with its most valuable data – the data on how our customers' equipment performs in real-life installations around the globe."
Pekka Mikkola, Director, Data Driven Services
CargoTec

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Partners recommend and build on AWS Glue



"It was remarkably simple to build our product with AWS Glue as our backend and integrate with the Cloudwick Data Analytics Platform. Our customers build visual jobs for Advanced Analytics and complex Data Pipelines that run on AWS Glue. Every job gets its own cluster in a few seconds and is extremely cost-efficient for us."
Jayant Shekar, CEO
Sparkflows.io

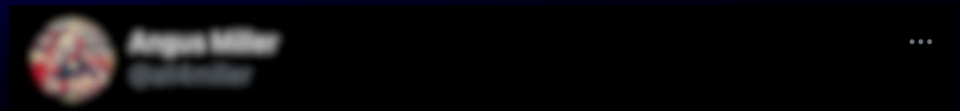
"We have used AWS Glue to help our Fortune 500 clients migrate from on-premises ETL and Spark workloads and build modern data platforms. This has helped enable our clients to reduce total cost of ownership and focus on delivering business value versus maintaining servers."
Sampath Parthasarathy
AVP Data and Intelligence Practice

Cognizant

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What customers have been saying



I'm super stoked... tabling a data processing and overall transactional and DWH upgrade path for a client in the SA finance space. Migrating to a stream based architecture ingesting data and using @AWS Glue for ETL and various other AWS bits and pieces like RDS etc fun times!



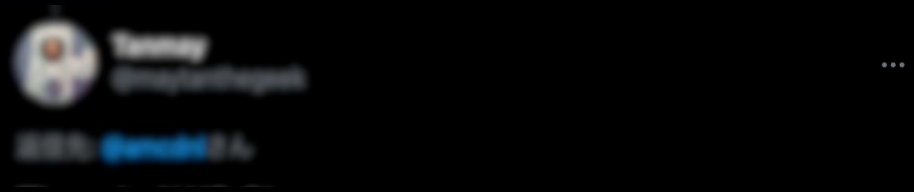
AWS Glue.. sounds pretty useful and powerful.



I love aws glue! Paired with athena and parquet data dumps into S3 it's a great way to externalise your data for other teams to easily pull in the materialised views they need.

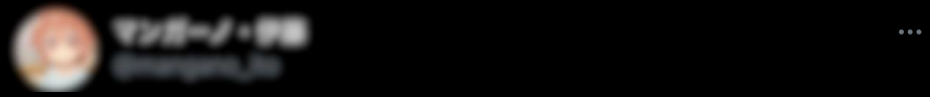


AWS Glue Custom Connectors いいね。SAP、Salesforce、Snowflake、BigQueryに接続して簡単にデータ連携できるコネクタの提供。マルチクラウドで欲しい機能だよね。

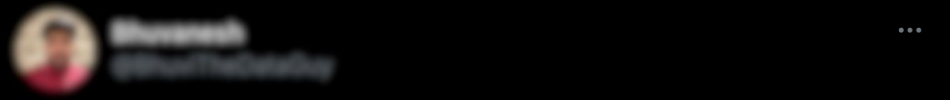


There is AWS Glue or better Lake Formation which helps you store data as parquet/avro in S3 and retrieve using SQL(Athena or spark).

I'm using it in production and it is solid.



AWS Glue Studio ポチポチやるだけで ETL できて超便利だ。かんたんなものなら Spark をチマチマ書く必要もないのだね。



Data Engineers spend most of the time on the data preparation side. After the arrival of #aws Glue DataBrew, its become easier now. Read one of the use cases of DataBrew to cleanup the Sensitive and PII data from the source dataset...lnkd.in/gfTnKTN

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What and why of data integration



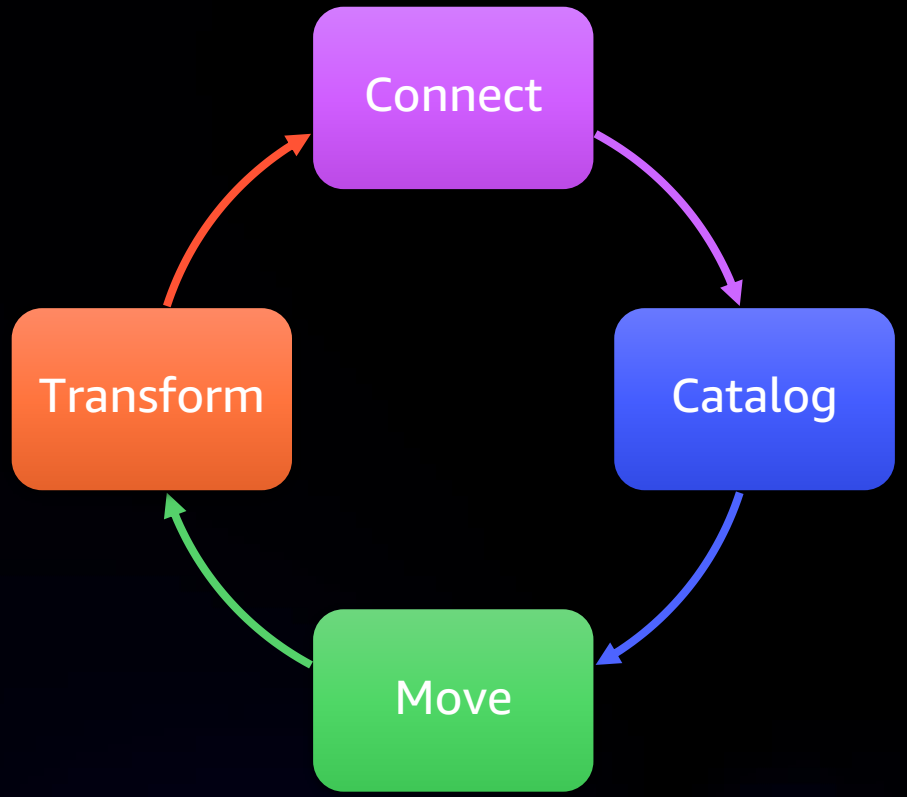
Meaningful insights



Increased collaboration



Faster decisions



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration is hard

Data



GROWING EXPONENTIALLY

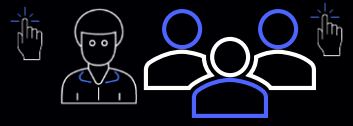


FROM NEW SOURCES



INCREASINGLY DIVERSE

Personas



NO OR LOW CODE

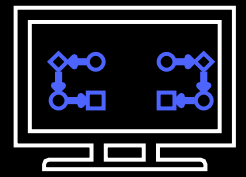


DEVELOPERS

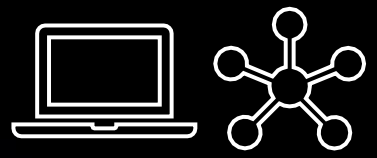


DATA ANALYSTS AND DATA SCIENTISTS

Applications



REAL-TIME/SLA SENSITIVE



HIGHLY SCALABLE



PRICE PERFORMANCE

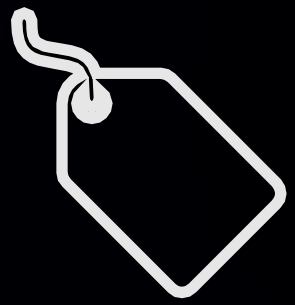
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Traditional solutions are not suited to handle the workloads



Non-scalable infrastructure

Traditional data integration tools are complex to install, manage, and scale



High cost

Advanced features such as centralized data catalog, transforms, and streaming data handlers need to be licensed separately

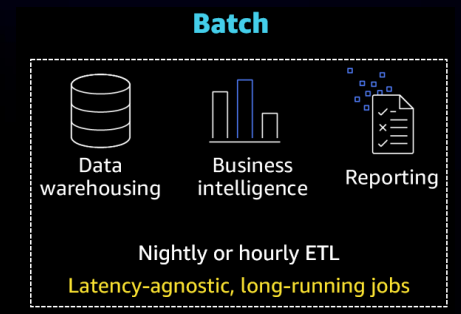


Lock-in

Jobs written in traditional tools are proprietary and cannot be easily ported

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

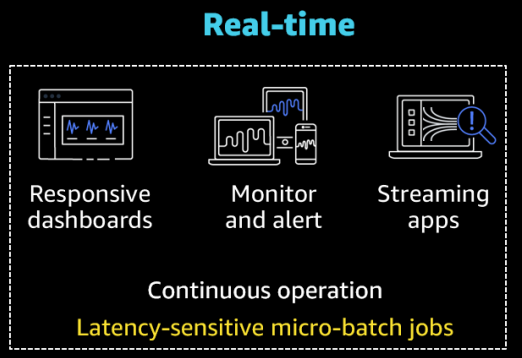
Evolution of AWS Glue



ETL service

Data preparation

Data integration



2017–2019

2020

2021

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue: Innovation pillars

SERVERLESS DATA INTEGRATION SERVICE

Connect & transform



Glue connectors



Built-in data transforms



Easy to migrate

Centralized & Unified Data Governance



Glue Data Catalog



Glue Crawlers

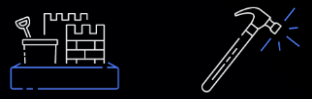


Lake Formation

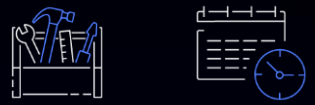
User Productivity and Data Ops



Variety of Persona



Productivity tools

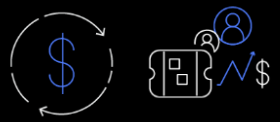


Data ops tools

Serverless Execution Engine



Scalable Execution engine



Flexible, cost-effective options



Execution monitoring

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Serverless execution engine



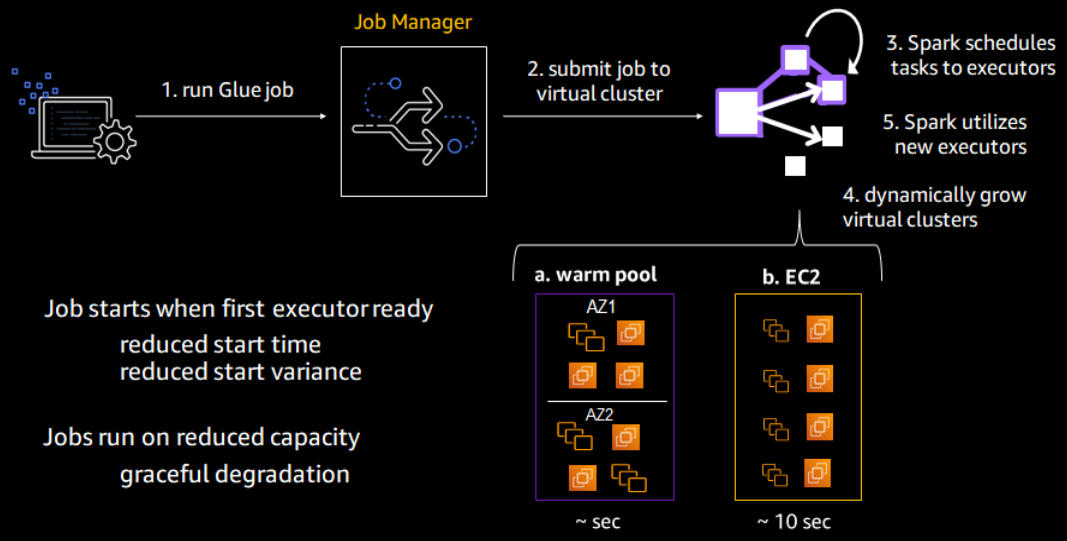
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue serverless execution engine recap

FASTER, PREDICTABLE, AND COST-EFFECTIVE

Fast | **Open** | **Scalable**

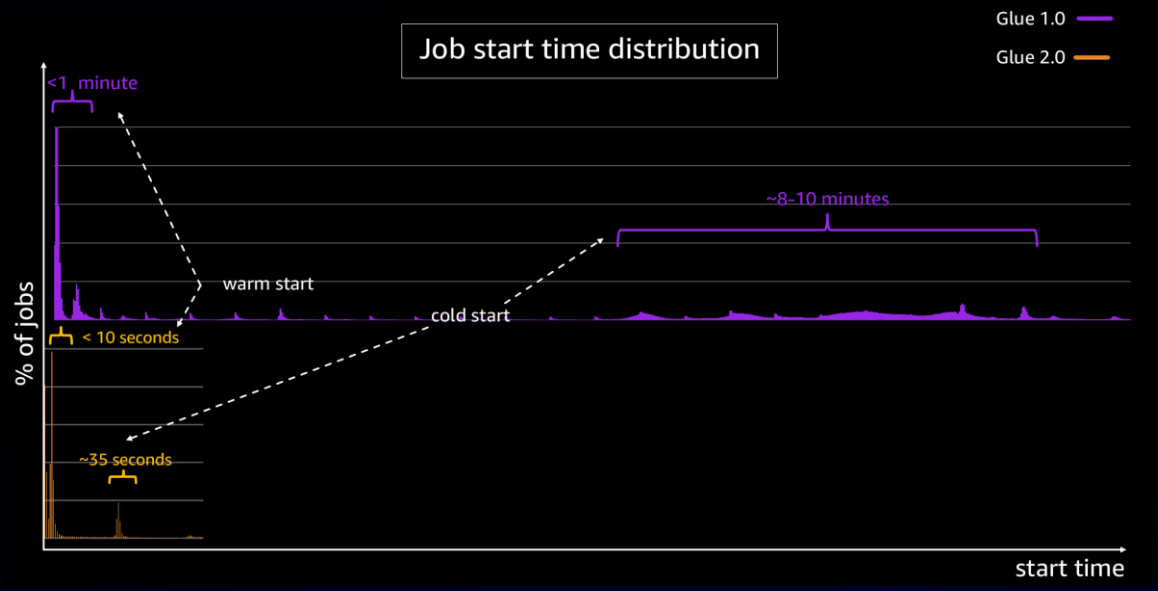
New job manager with AWS Glue 2.0



Job starts when first executor ready
- reduced start time
- reduced start variance

Jobs run on reduced capacity
- graceful degradation

Predictable start time Significant cost savings



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

New

Serverless execution engine – AWS Glue 3.0

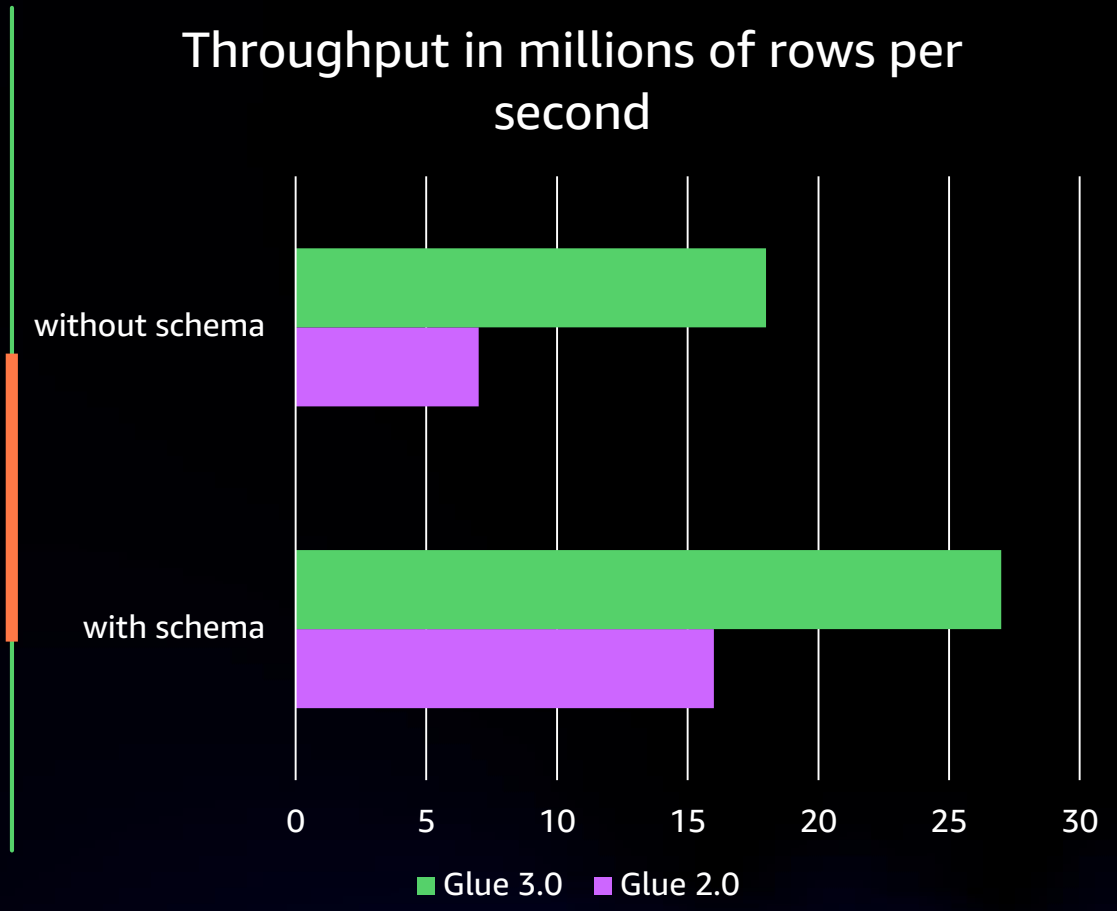
FASTER, PREDICTABLE, AND COST-EFFECTIVE

Optimized for reading row-based formats and writing to columnar formats

C++ SIMD-based vectorized readers

Columnar data structures for in-memory processing

As fast and as predictable startup times as AWS Glue 2.0



AWS Glue 3.0 is 2.5x faster than AWS Glue 2.0

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

GALE

"Speed is essential in our business. AWS Glue's serverless infrastructure enables our data scientists to quickly explore and transform data required for business use-cases in our Alchemy platform. The recent Glue 3.0 upgrades like faster startup times and optimized file IO have reduced our costs by 30%."

Vu Pham-Tran, Chief Technology Officer, GALE

nexi

"In Nexi, we've been using AWS Glue for a couple of years for large-scale processing. We recently tested Glue 3.0 and decided to adopt it in our production environment: its improved effectiveness in dealing with heavy workloads (especially when converting files to Apache Parquet format) has been key in this choice."

Alberto Danese, Head of Data Science, Nexi

Shutterfly

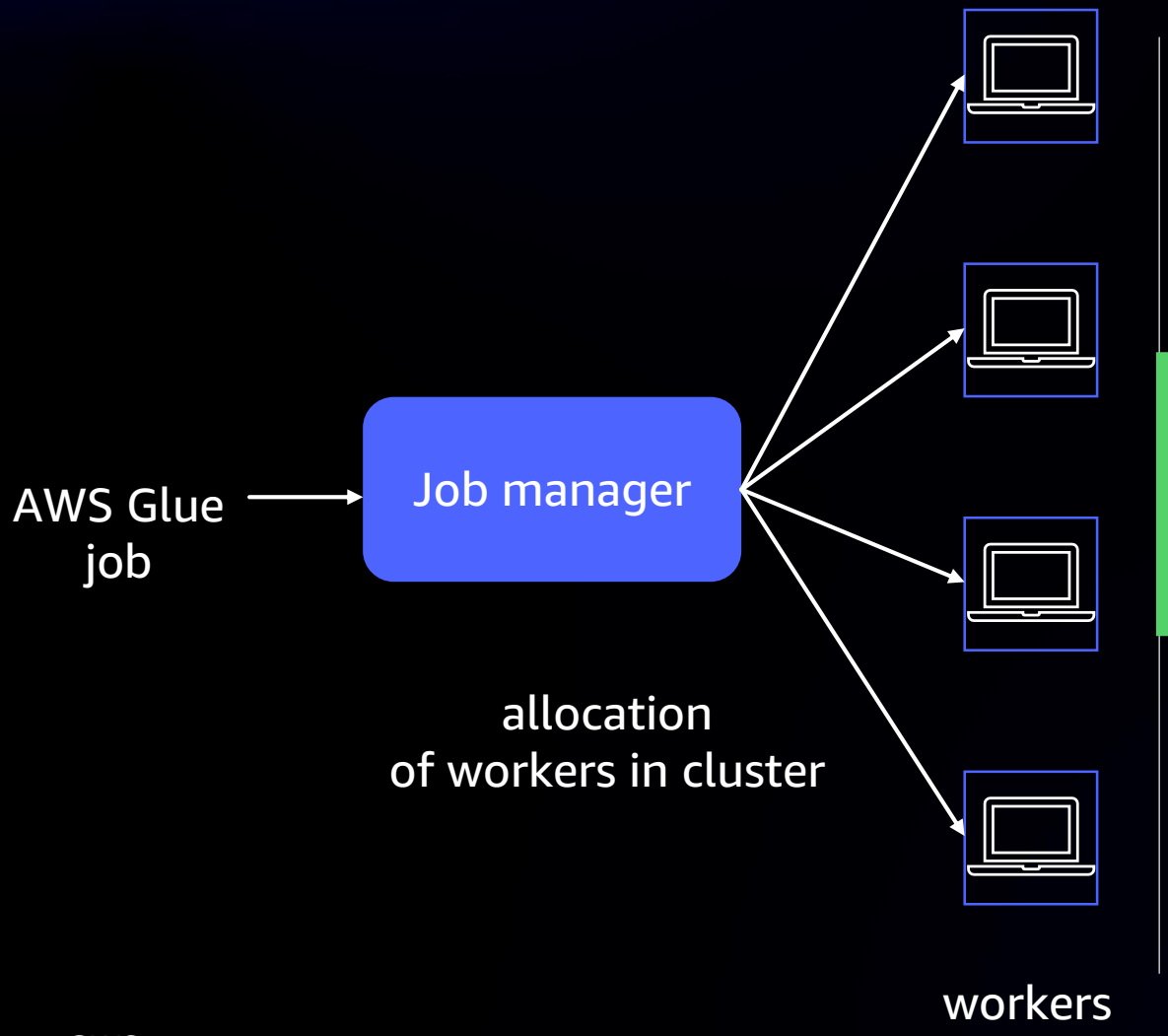
"We are using AWS Glue (v3.0) to ingress data from many varied sources and to run our ETL pipelines. The significantly reduced cost of Glue and shorter startup times, as compared to our legacy system, have allowed us to run more incremental loads per day, giving our analysts hourly refreshes instead of daily refreshes at a fraction of the cost."

Daron Scarborough, Data Engineer Director, Shutterfly

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scalable execution model



Jobs are divided into stages

Data is divided into partitions (shards) that are processed concurrently

1 stage x 1 partition = 1 task

Job manager schedules task on workers

Glue's execution scales to 1000s of workers

You pay for compute used

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Predicting capacity is difficult



Fluctuation in data

Time of day

Seasonality

Data skew

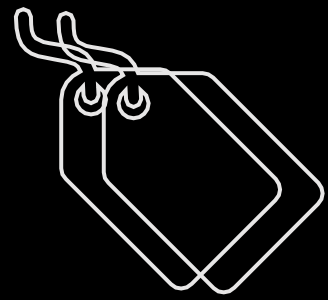


Needs tuning

Admin inputs

Monitoring

Cost controls



Provisioning

Over-provisioning

Extra cost

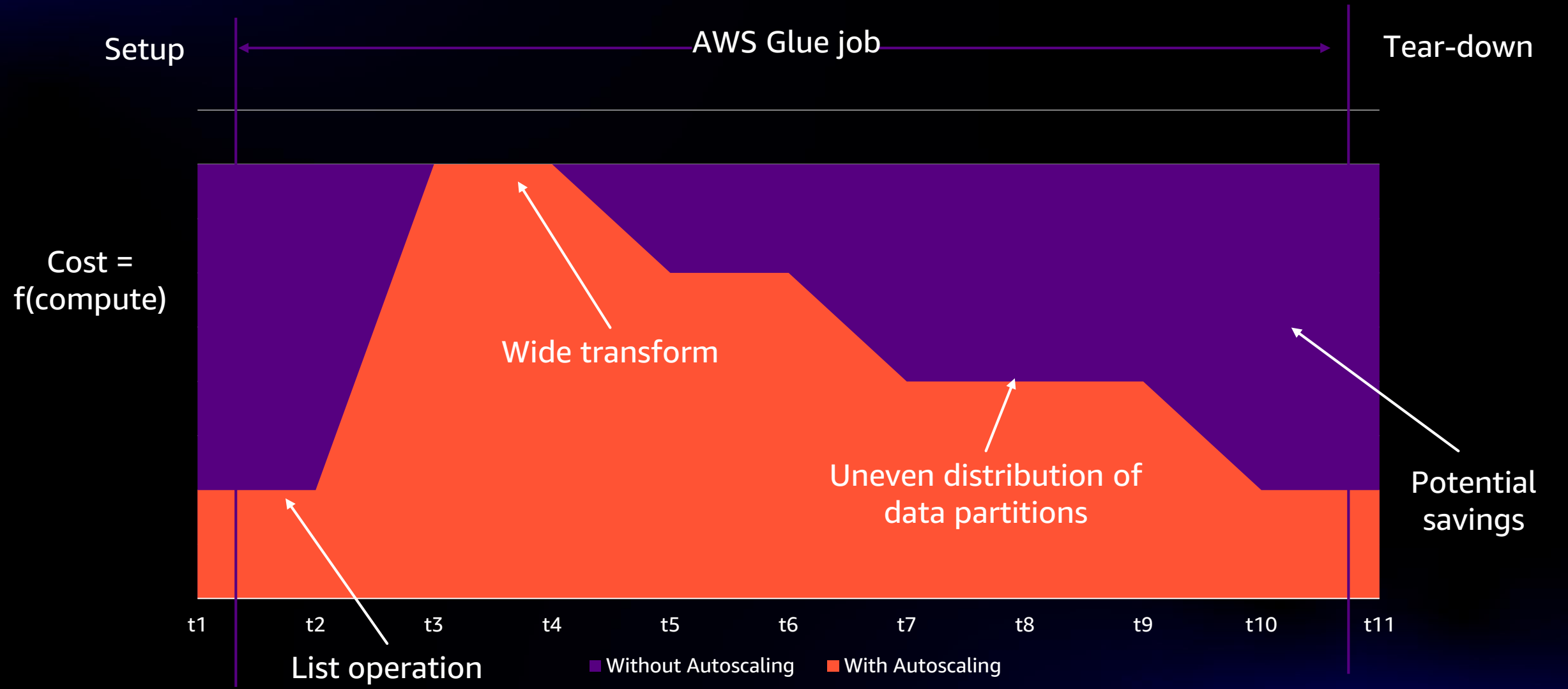
Performance variation

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

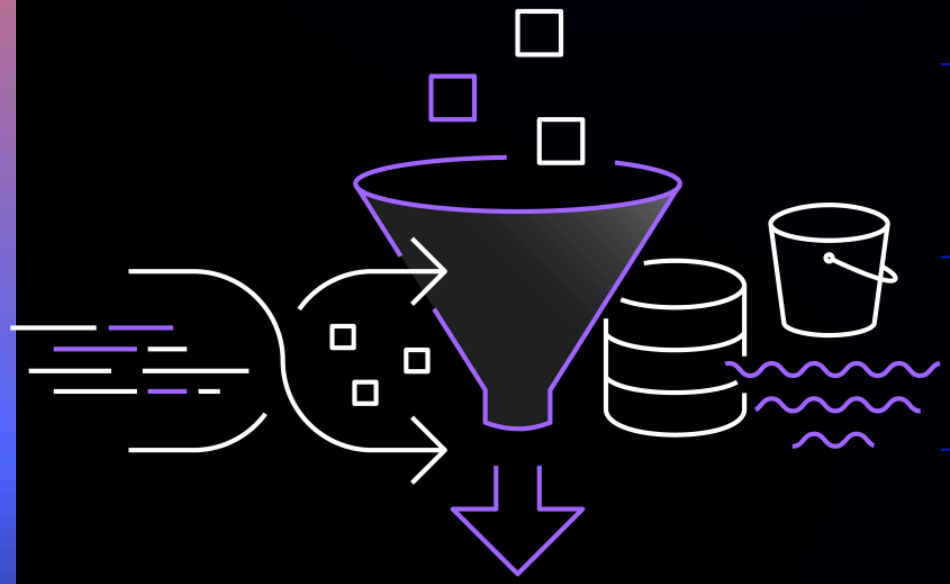
Preview

Announcement: AWS Glue Auto Scaling



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Streaming ETL



Process stream data & make it queryable in seconds

Join streams against each other or static data

Automatic updates to the AWS Glue Data Catalog

Dozens of supported data targets

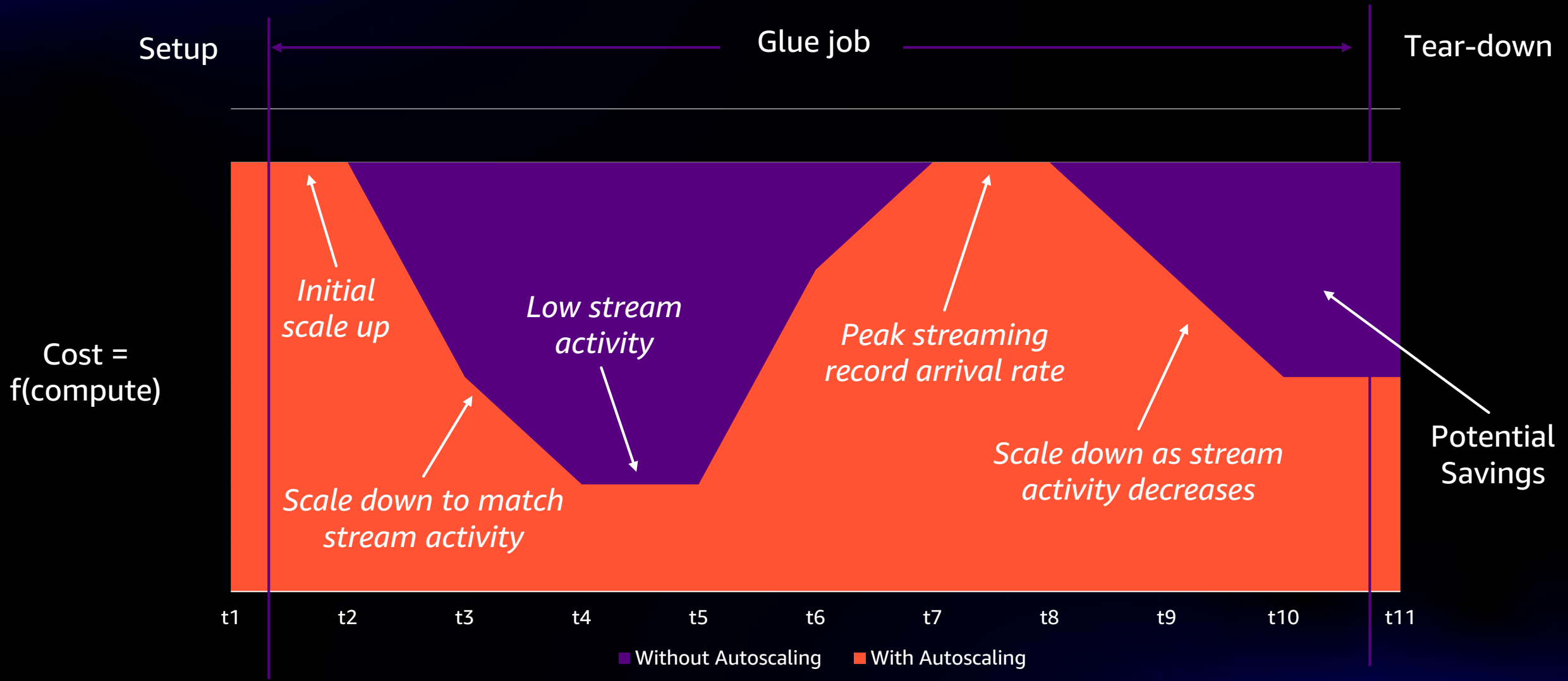
Simplify your architecture with one service for streaming and batch data integration

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Announcement: Streaming Auto Scaling

Preview



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

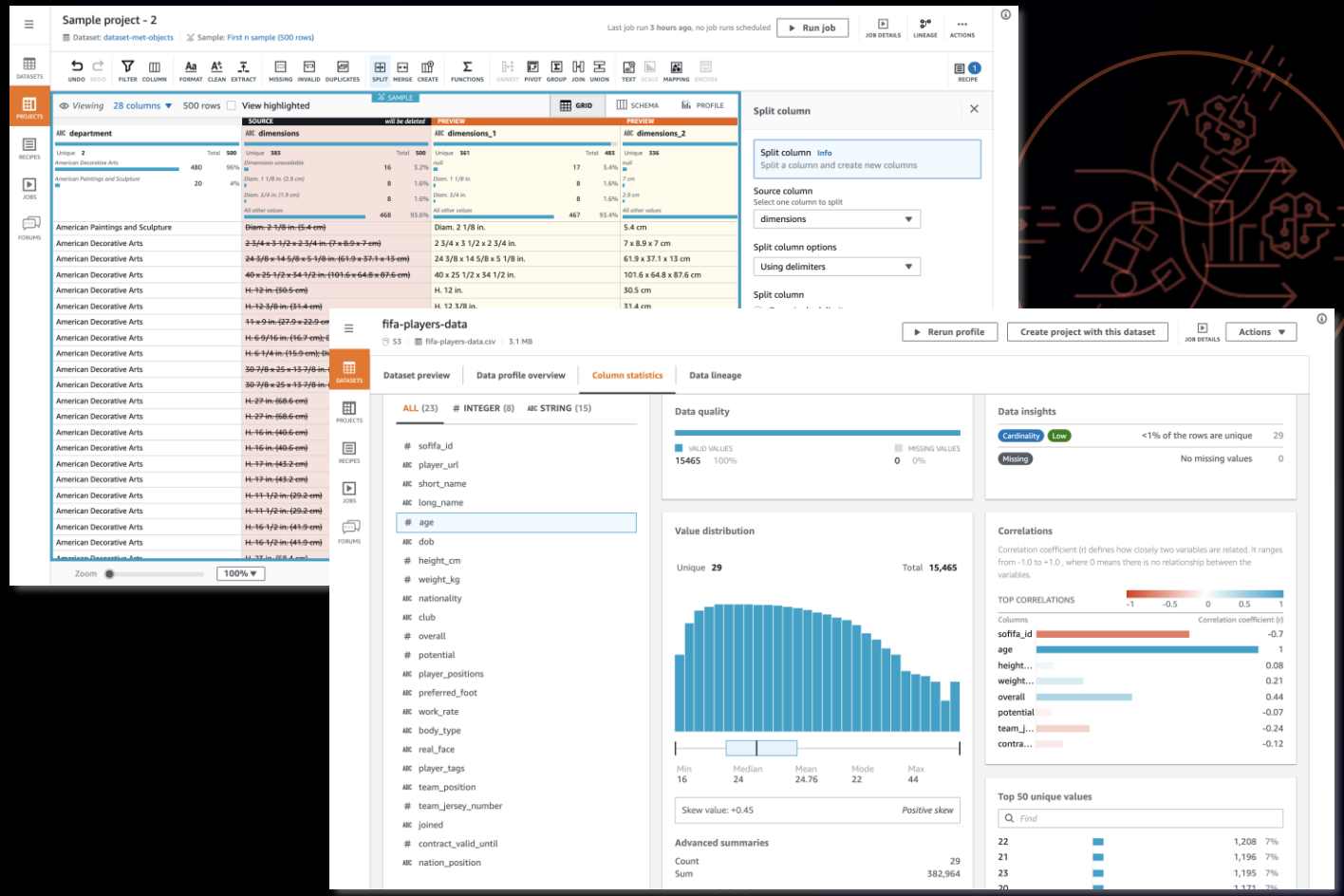
User productivity and data ops



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue DataBrew

VISUAL DATA PREPARATION AND DATA QUALITY CHECKS



Chose from 350+ transformations to prepare data

Identify dataset quality with 50+ statistics

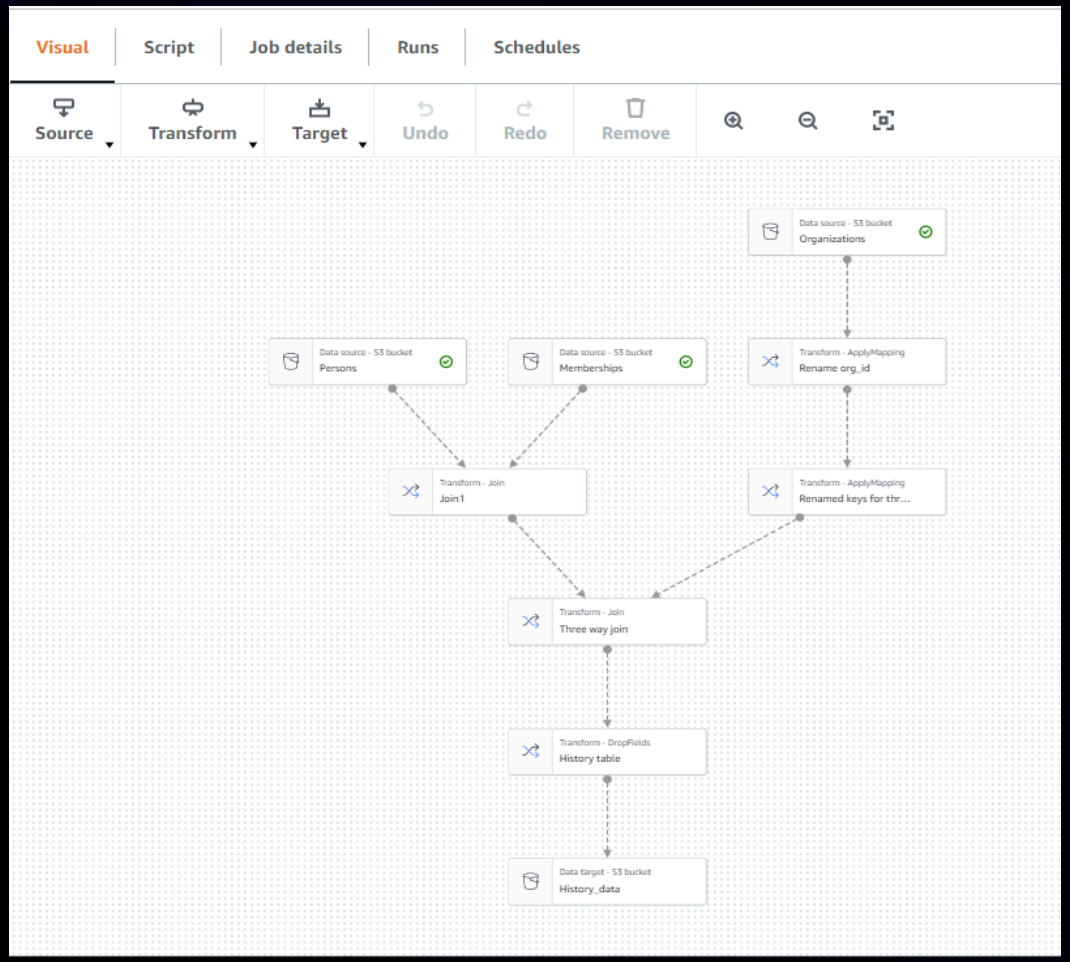
Share, reuse, schedule transforms across users and teams

Run serverless data preparation and at scale

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Studio

VISUAL JOB AUTHORING AND MONITORING

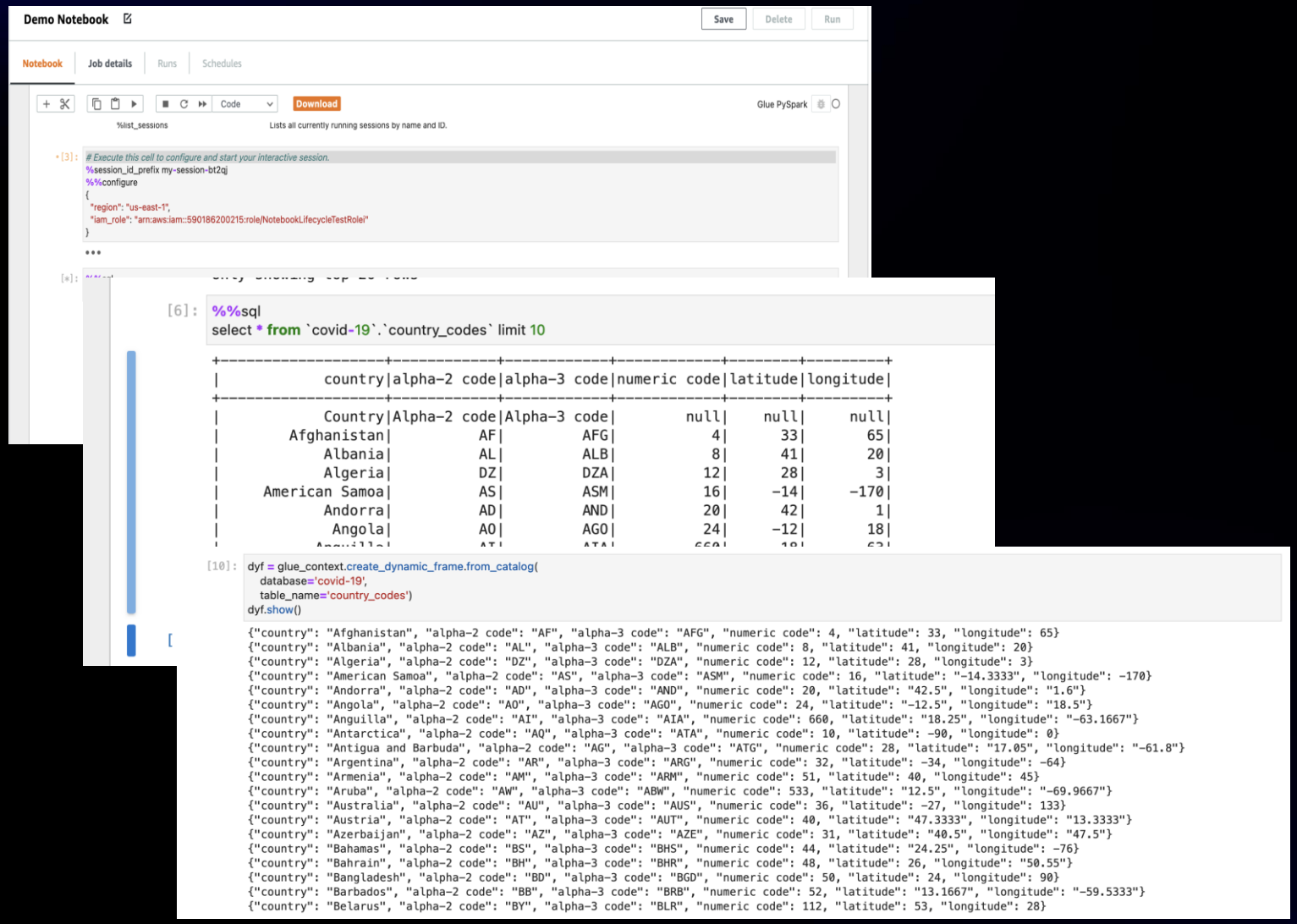


- Monitor thousands of jobs through a single pane of glass
- Advanced transforms though code snippets
- Support for AWS Marketplace and custom connectors
- Preview your data at each step of the visual job authoring process
- Real-time schema inference without having to catalog

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Preview

Announcing: AWS Glue Studio Notebook



Interactive AWS Glue jobs development

Submit AWS Glue jobs from the AWS Glue Studio notebook

Use notebook magic to define transforms in SQL and control cost

Built-in monitoring support

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Announcing: AWS Glue Interactive Sessions

Preview

Existing options

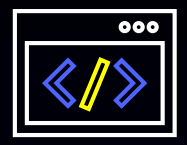
Time to first query = 10-15 minutes

High cost of a long-running cluster

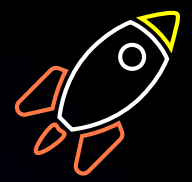
"Noisy Neighbor" problem

AWS Glue Interactive Sessions

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



Development tool of your choice



Rapid development

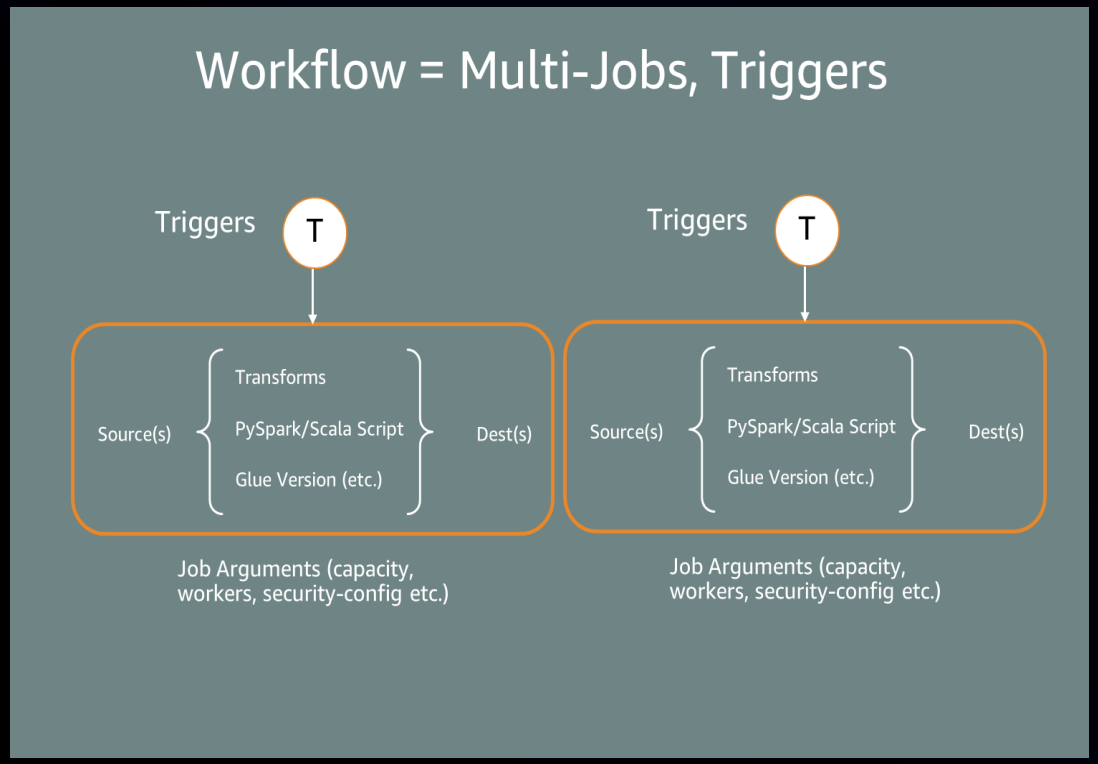


Built-in cost control

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Orchestrate jobs easily with AWS Glue Workflows



Orchestrate AWS Glue Jobs and other AWS services

Monitor execution of the workflows in one place

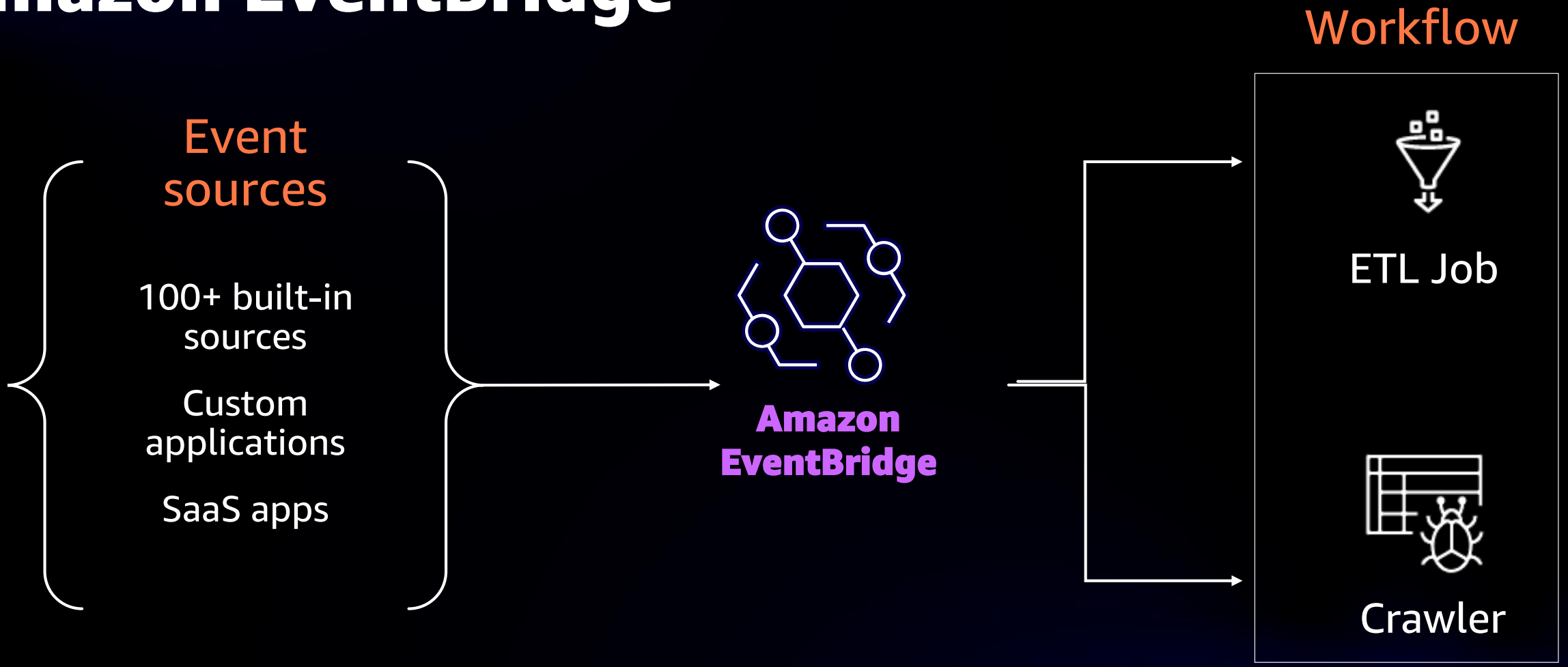
Schedule-based and on-demand triggers

Easy to access logs and monitor progress

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

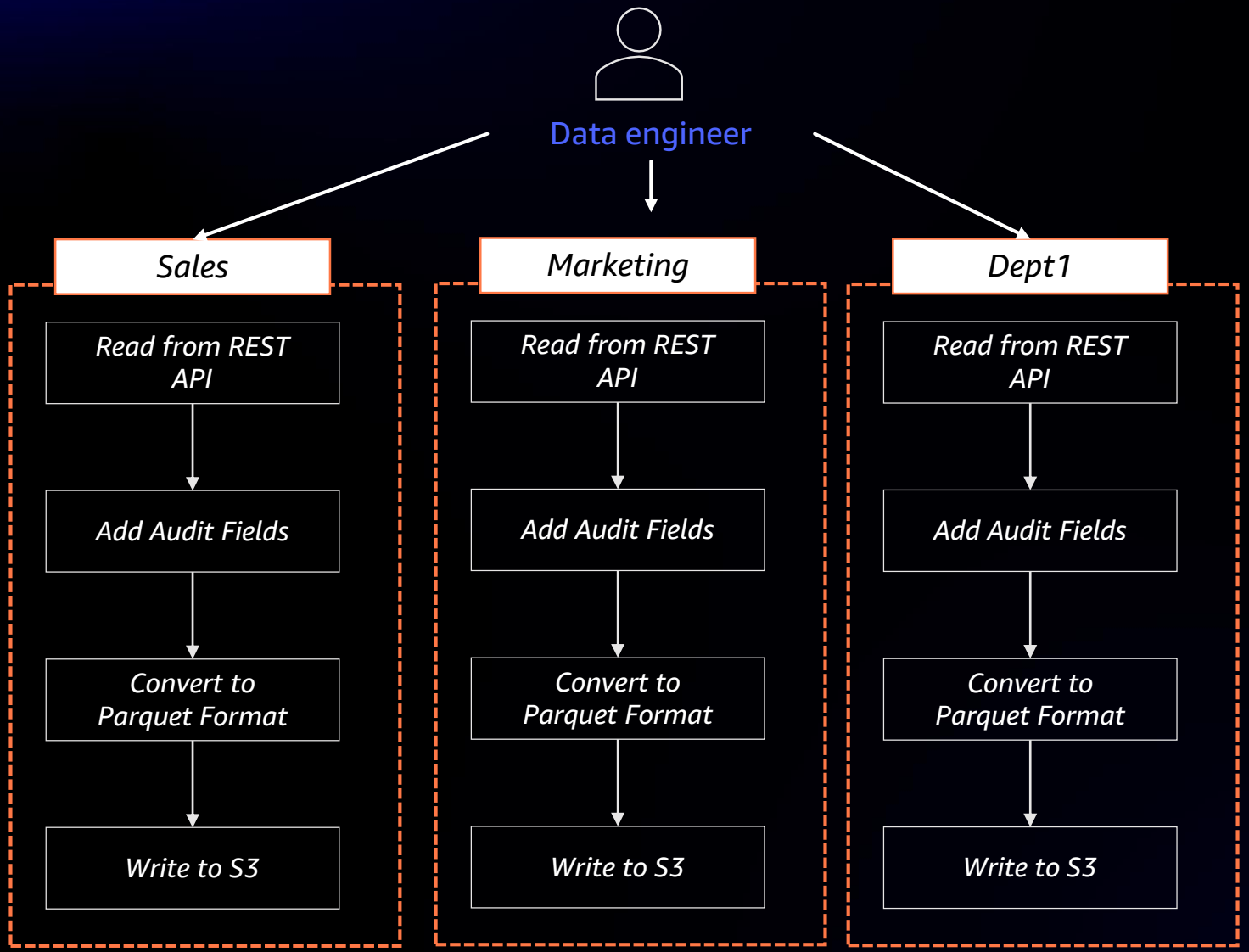
New

Event-driven workflows using Amazon EventBridge



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Many ETL processes are repeatable… Developers spend valuable time in creating redundant code



Challenges

Manual Coding

Error prone

Repetitious

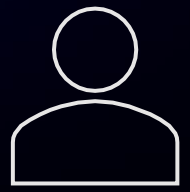
Non-scalable

aws

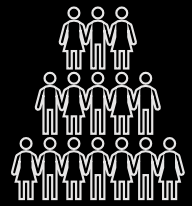
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue custom blueprints

New



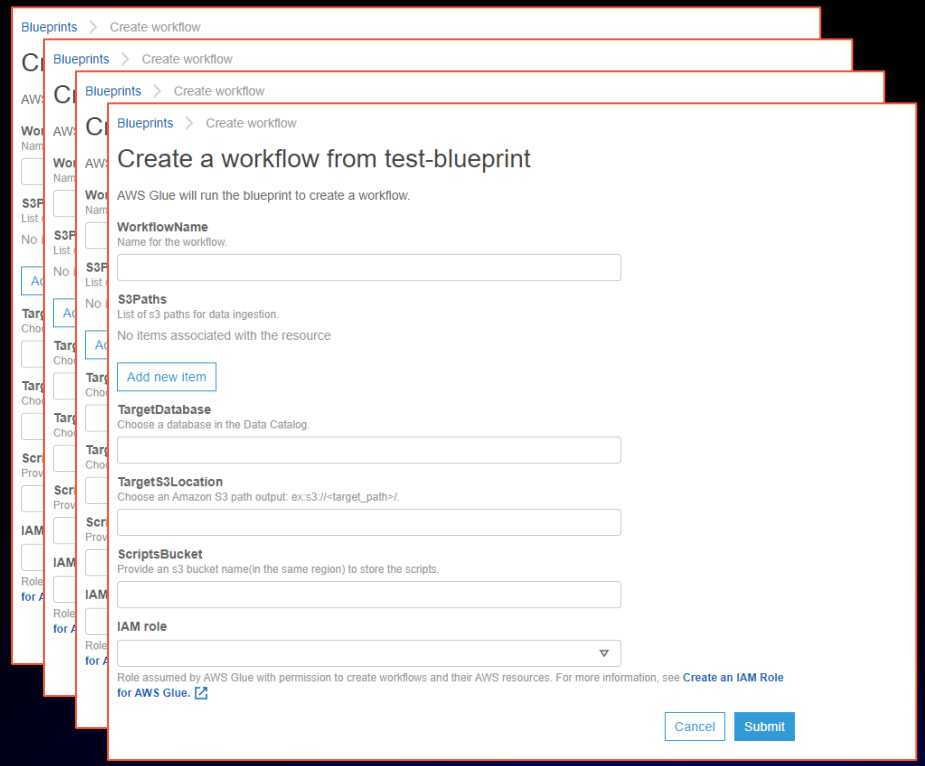
Data engineer



Analysts

Glue blueprint

- Glue Job script
- Configuration File
- Layout File



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Connect and transform data

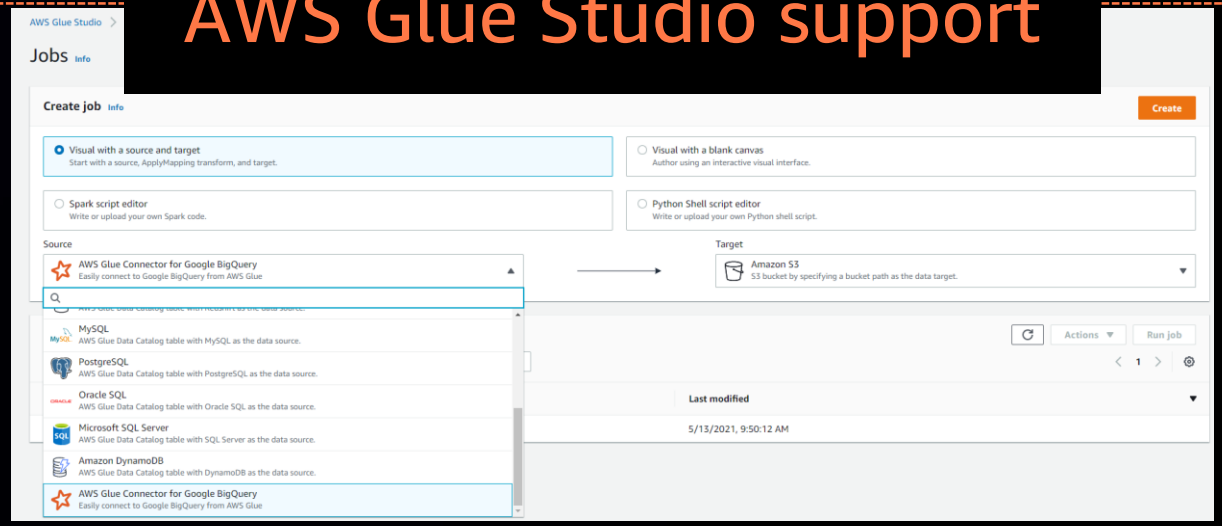


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue connector ecosystem



AWS Glue Studio support



Scalable ingestion

Pushdown predicate for efficient processing

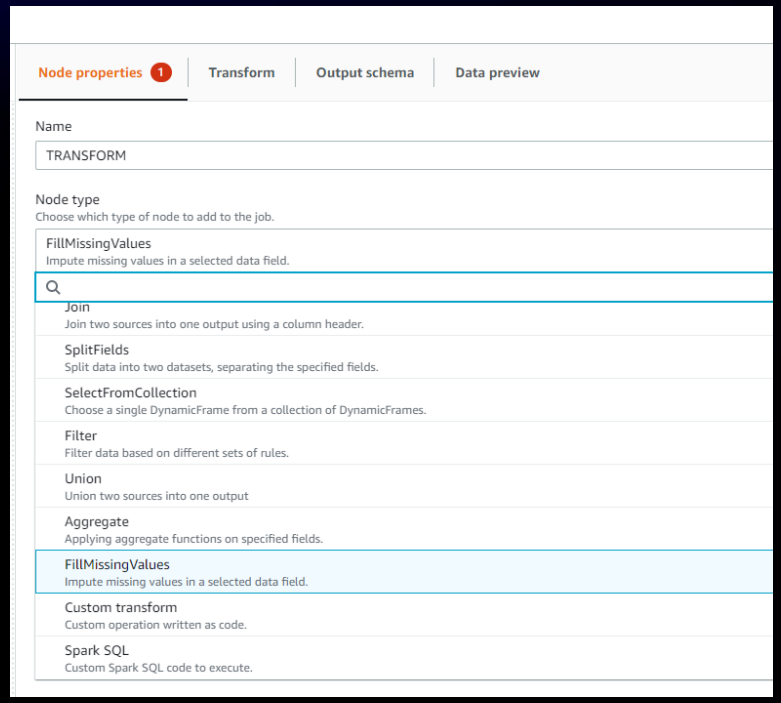
AWS Glue Data Catalog integration

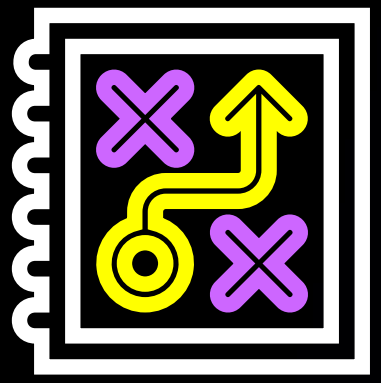
AWS Secrets Manager integration

Job bookmark for incremental processing

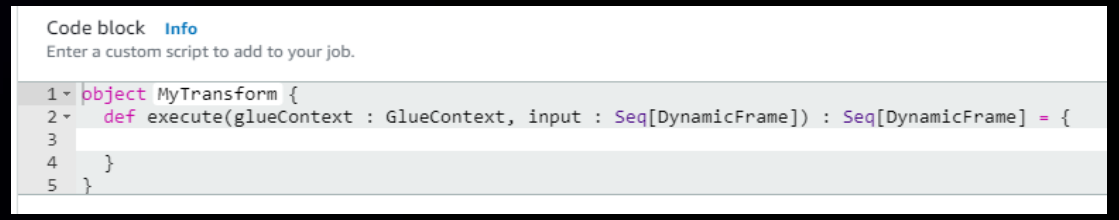
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

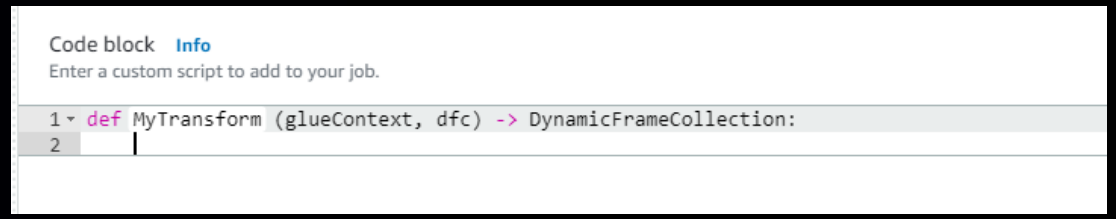
Support for built-in and custom transforms

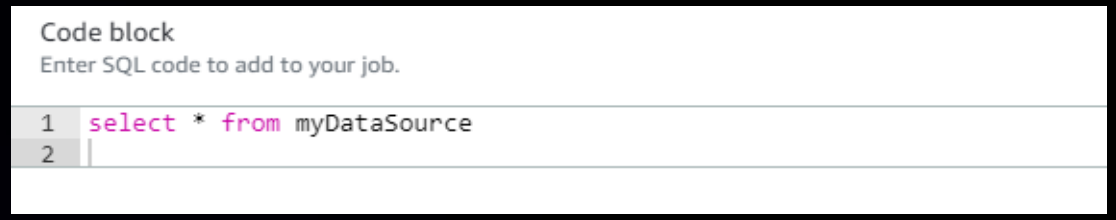


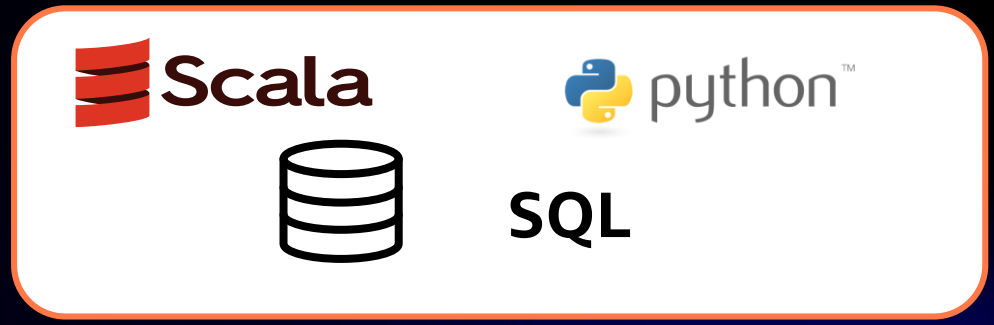


built-in transform



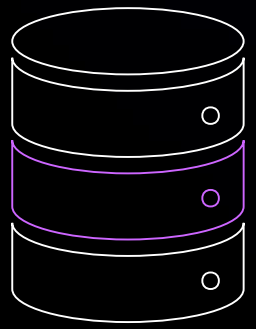






© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

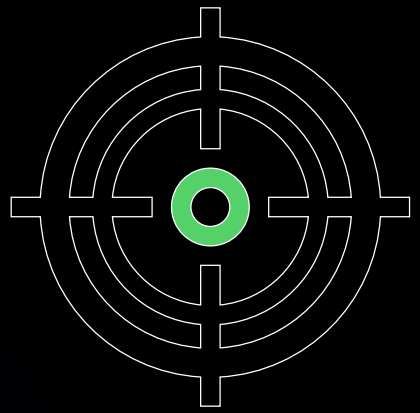
Challenges with PII detection and remediation



Size of the dataset



Location of PII



Accuracy

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

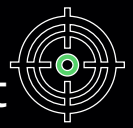
Preview

Announcement: AWS Glue PII Detection and remediation

THREE SIMPLE STEPS

1 Type of Scan 

- Full Scan
- Sample Scan

2 Entities to detect 

- Built-in Entities (e.g. SSN, passport)
- Custom Entities

3 Remediation 

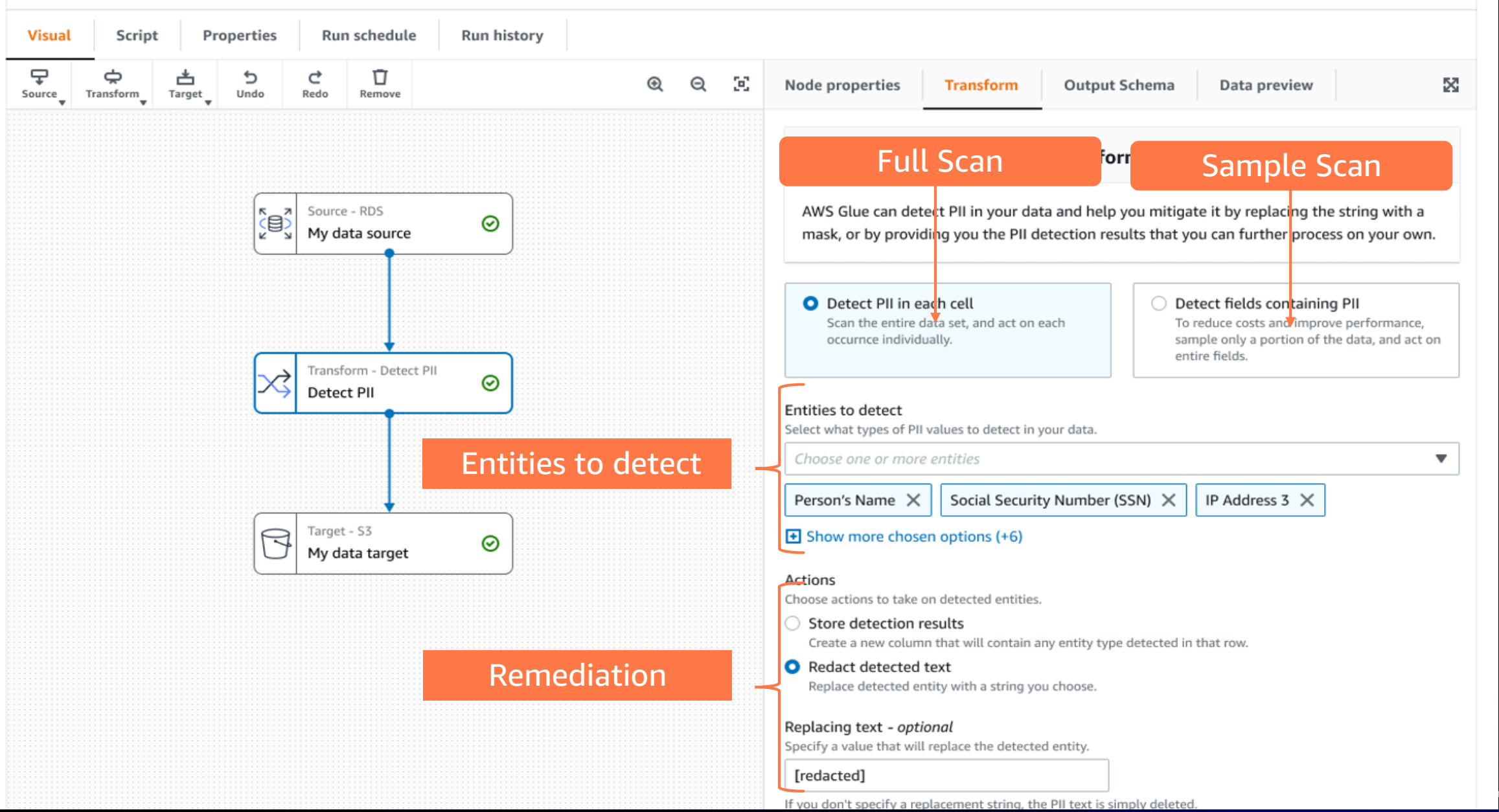
- Store results
- Redact/mask results

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

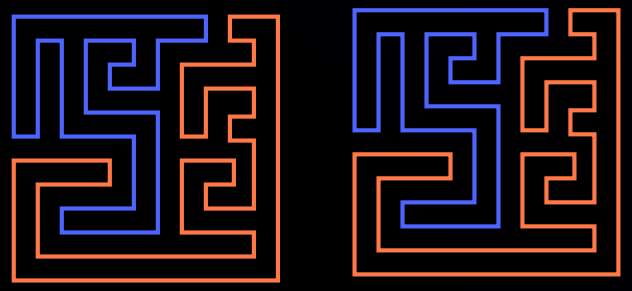
How does PII Detection work

Preview



Challenges with migrating from traditional ETL tools

Traditional ETL solution

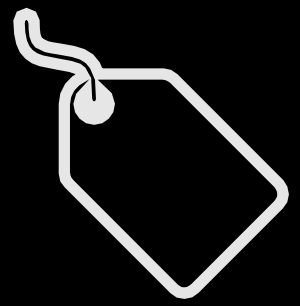


Proprietary format

Migration challenges

Hand coding

Rewriting



Time consuming, complex, and expensive

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Preview

Announcing: Migrate your traditional ETL jobs

Legacy ETL solution

Source

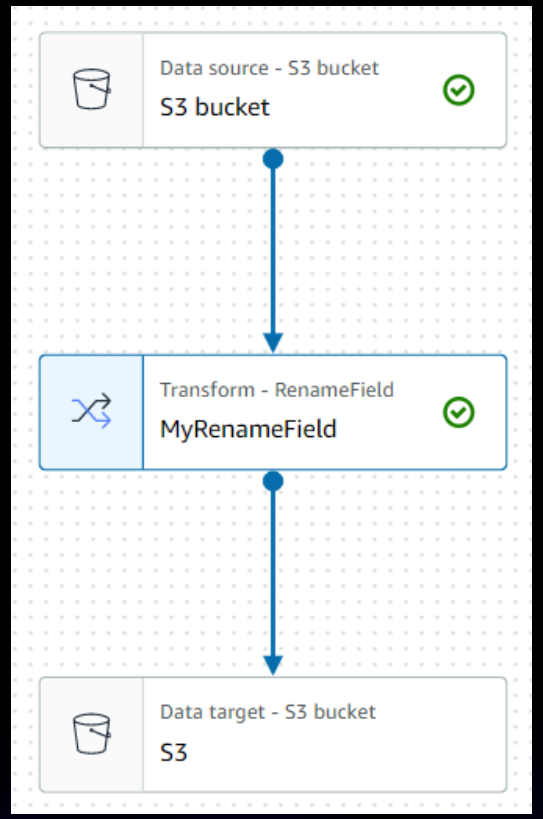
Transforms

Sink

Representative JSON



```
{
  "classification": "Transform",
  "type": "RenameField",
  "inputs": [
    "node-1"
  ],
  "name": "MyRenameField",
  "sourcePath": "col3",
  "targetPath": "name"
}
```



aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ETL migration: AWS Partner quotes



Bladebridge ETL convertors uses 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 & Cofounder, BladeBridge



Over 80% of large enterprises need to modernize their legacy ETL as they move analytics to the cloud. Next Pathway is excited to integrate with AWS Glue Studio APIs to automate ETL migration to Glue and Glue Studio, making it faster to move analytics to AWS.

Chetan Mathur, CEO and Founder, Next Pathway

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Centralized and unified data governance



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Data Catalog is among the most popular metadata stores for data lakes

High **durability, availability,** and **scalability**

Serverless and **cost effective**

Most ways to **get metadata in**



Kafka **Schema Registry**

Broadest portfolio of integrated tools

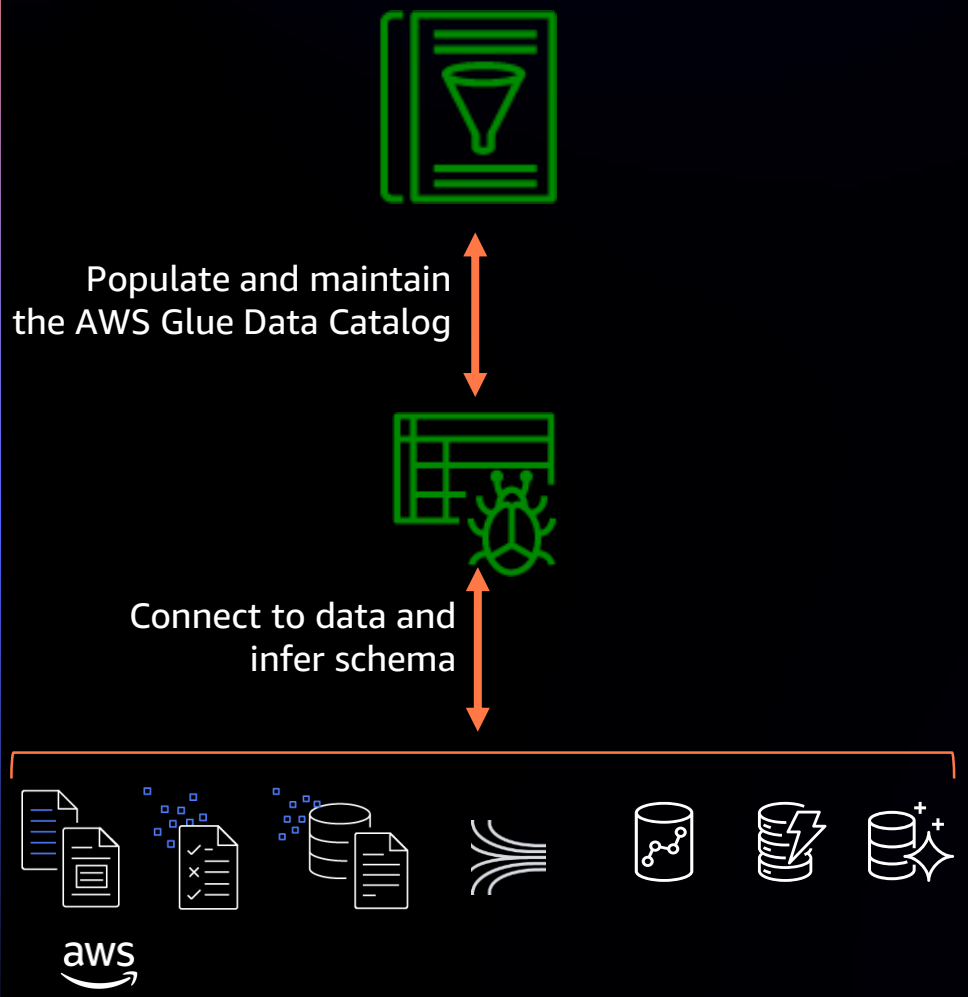
Best **security, compliance,** and **audit** capabilities

Hive metastore API compatible

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Crawlers: populate and maintain schemas



Automatically discover new data and extract schema definitions

Detect schema changes and maintain tables detect Apache Hive-style partitions on Amazon S3

Built-in classifiers for popular data types create your own custom classifier using Grok expressions

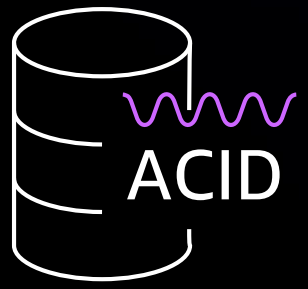
Run on demand, on a schedule, on an event, or as part of AWS Glue workflows

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation integration

NEW TYPE OF AMAZON S3 TABLE

UPDATE DATA

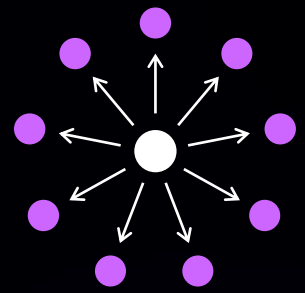


ACID TRANSACTIONS

CONSISTENT ACROSS TASKS
INSERT, UPDATE, DELETE
CONVERGE BATCH & REAL-TIME

RELIABLE

OPTIMIZATION ACCELERATION



STORAGE OPTIMIZATION

AUTO-COMPACT SMALL FILES
PUSH-DOWN FILTERS
REDUCE DATA SCAN

PERFORMANT

TIME TRAVEL



DATA VERSIONS

DATA HISTORY
REPRODUCE EXPERIMENTS
AUDIT CHANGED DATA

VERSIONED

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How BMW is using AWS Glue

Simon Kern
Lead Engineer
BMW Group

Josef Viehhauser
Product Owner
BMW Group

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

1. Overview
2. Architecture
3. Deeper dive
    - AWS Lake Formation
    - Lineage
4. Wrap-up

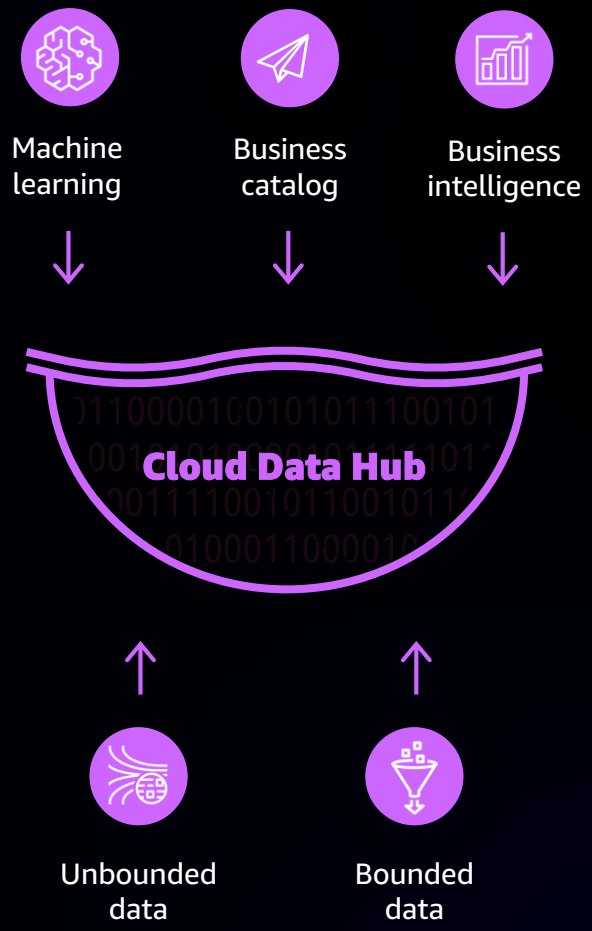
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building a platform that makes a difference



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building a platform that makes a difference



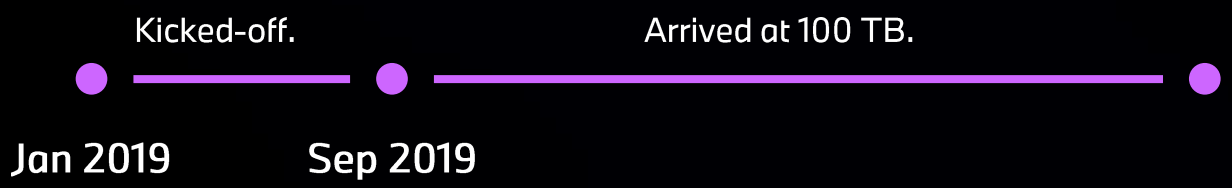
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

This is where we come from.



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

This is where we come from.



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

This is where we come from.



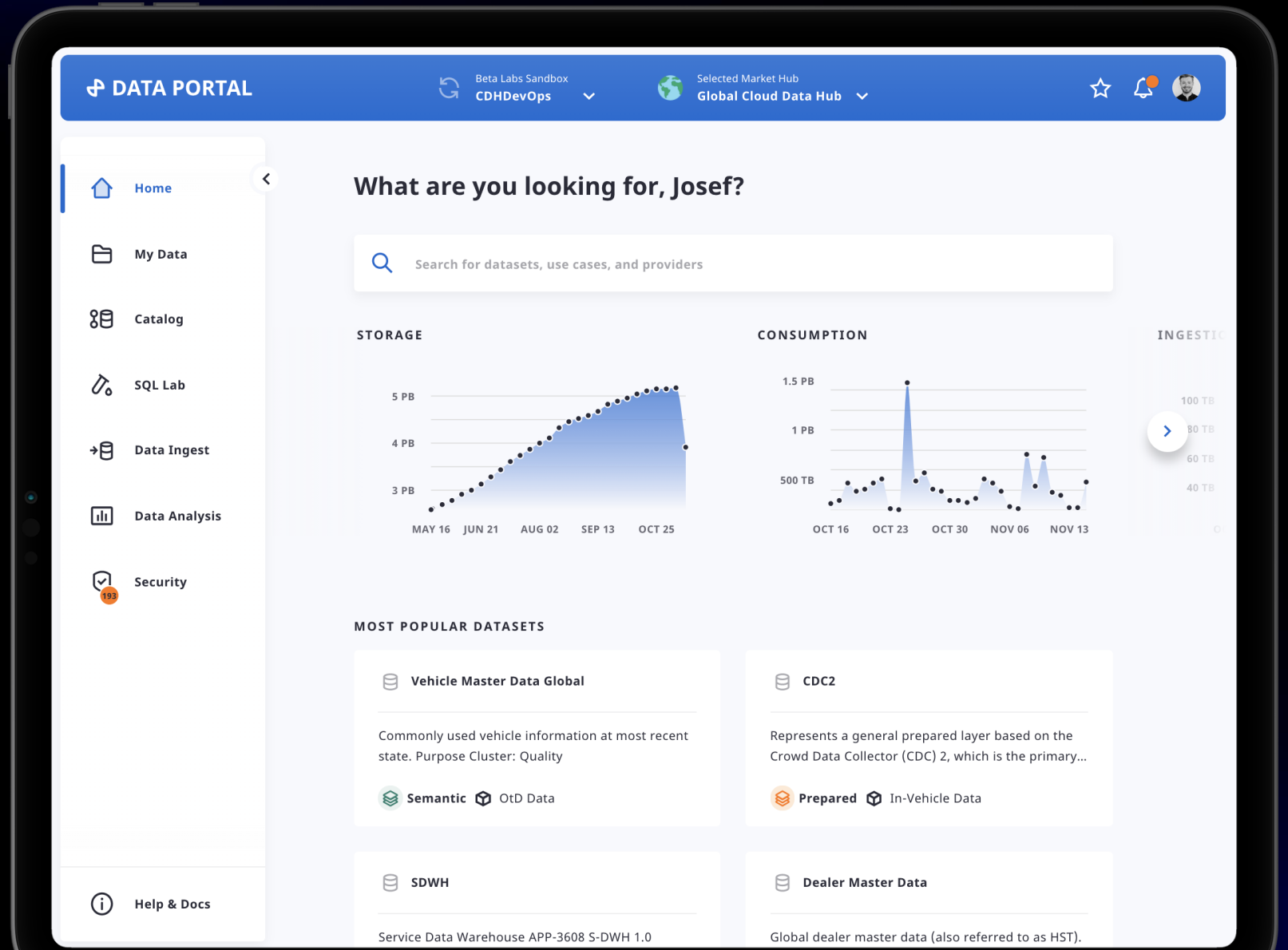
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

This is where we come from.



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our portal makes data universally actionable



© 2021, Amazo

Our portal makes data universally actionable

Discover

Browse through data and use cases.

Explore

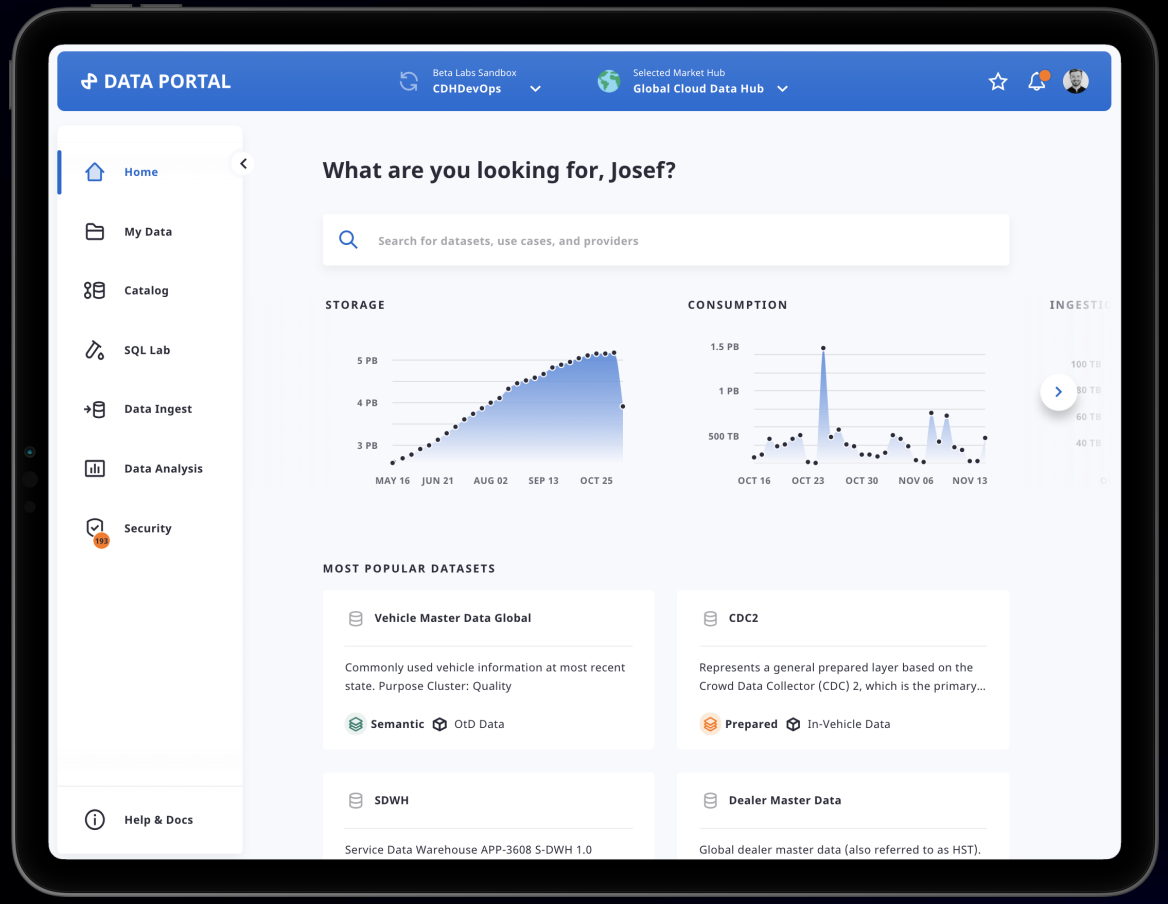
Understand the structure of datasets via SQL.

Publish

Use pre-baked IaC modules and no-code tools.

Consume

Transform and analyze data with code workbooks.



Manage

Curate metadata for data products.

Request

Obtain access to data in a governed fashion.

Monitor

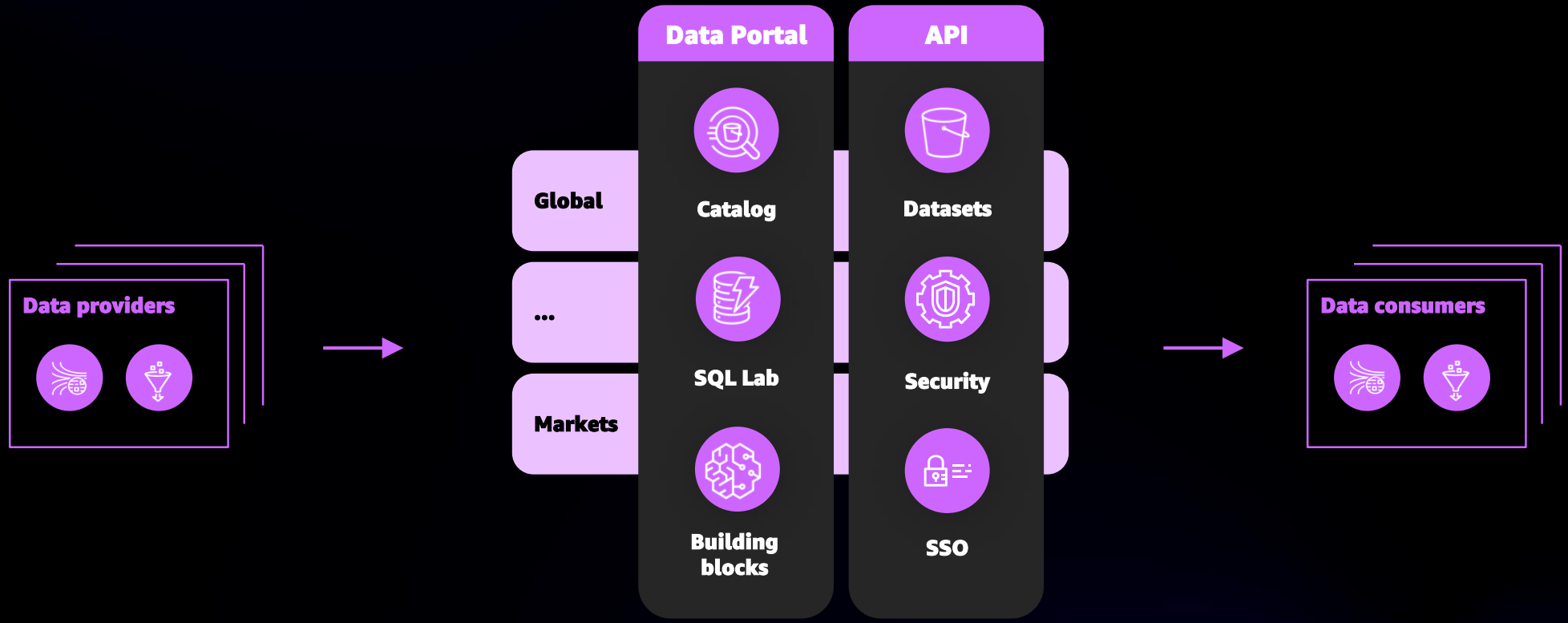
Comprehend and inspect lineage of data pipelines.

Connect

Integrate with systems outside of the platform.

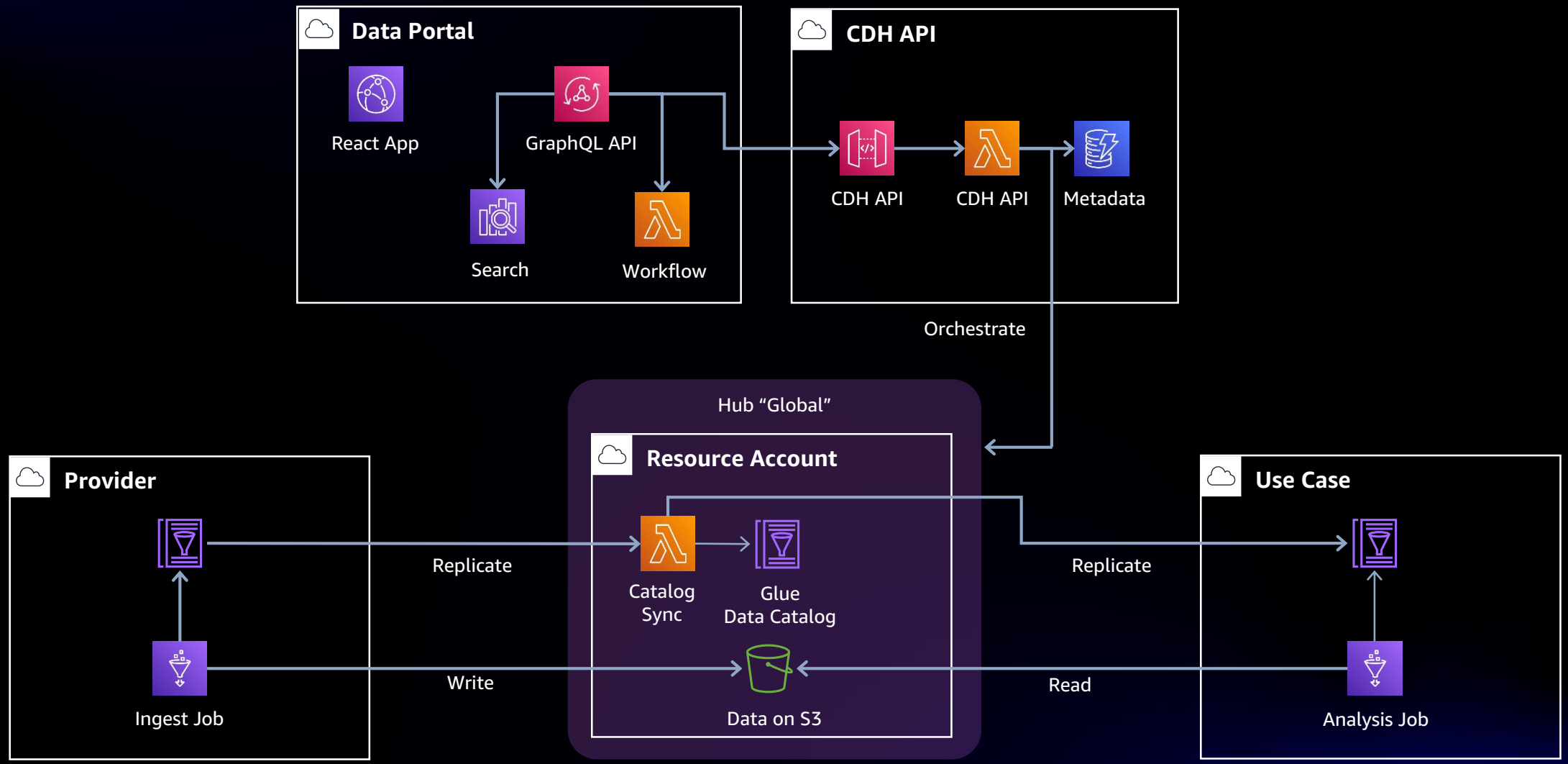
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our architecture scales horizontally



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

We are building around Glue at the core



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Moving towards AWS Lake Formation

Why

- Future-ready service aiming to make data lake setups easier and more powerful
- Leverage new features as part of that
  - Fine-grained access control (table-, column-, row-level)
  - Transactions
  - Automated compaction via governed tables
- Remove our custom-built data catalog synchronization

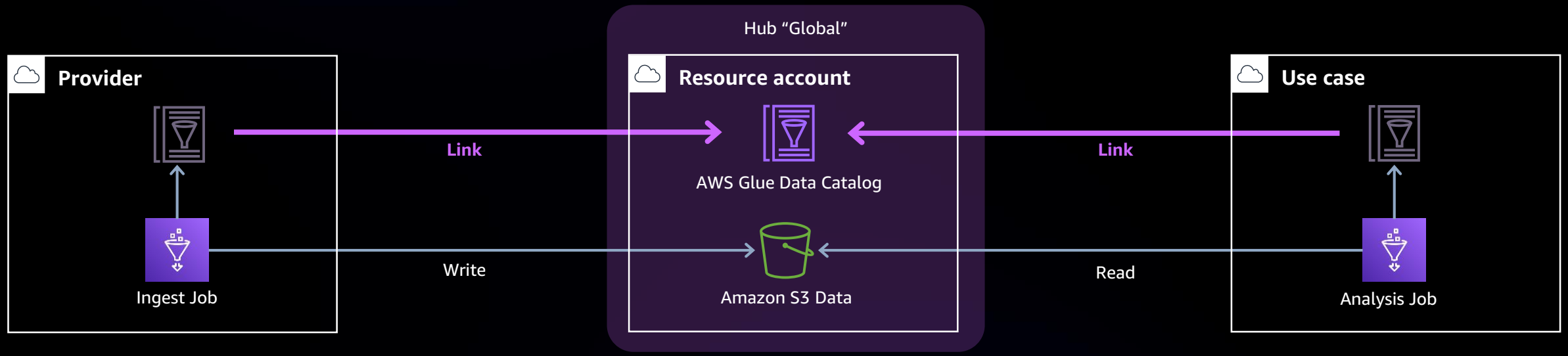
How

Migration to be conducted in two steps

- Step 1: Introduce Glue resource links for sharing metadata
- Step 2: Complete lake formation rollout

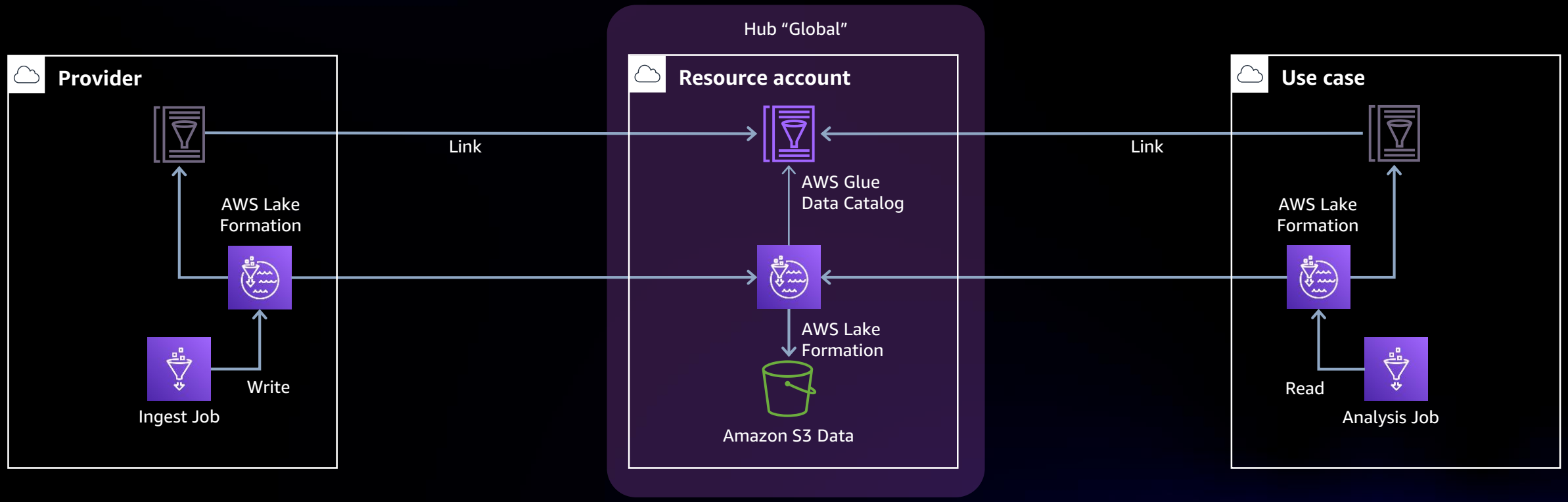
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Step 1: Introduce resource links



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Step 2: Roll out AWS Lake Formation



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

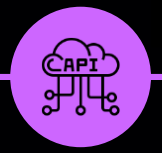
We have worked on integrating lineage E2E



Automated collection of data lineage for AWS Glue Jobs



Knowledge graph for coarse- and fine-grained data lineage



APIs to universally extend and extract data lineage

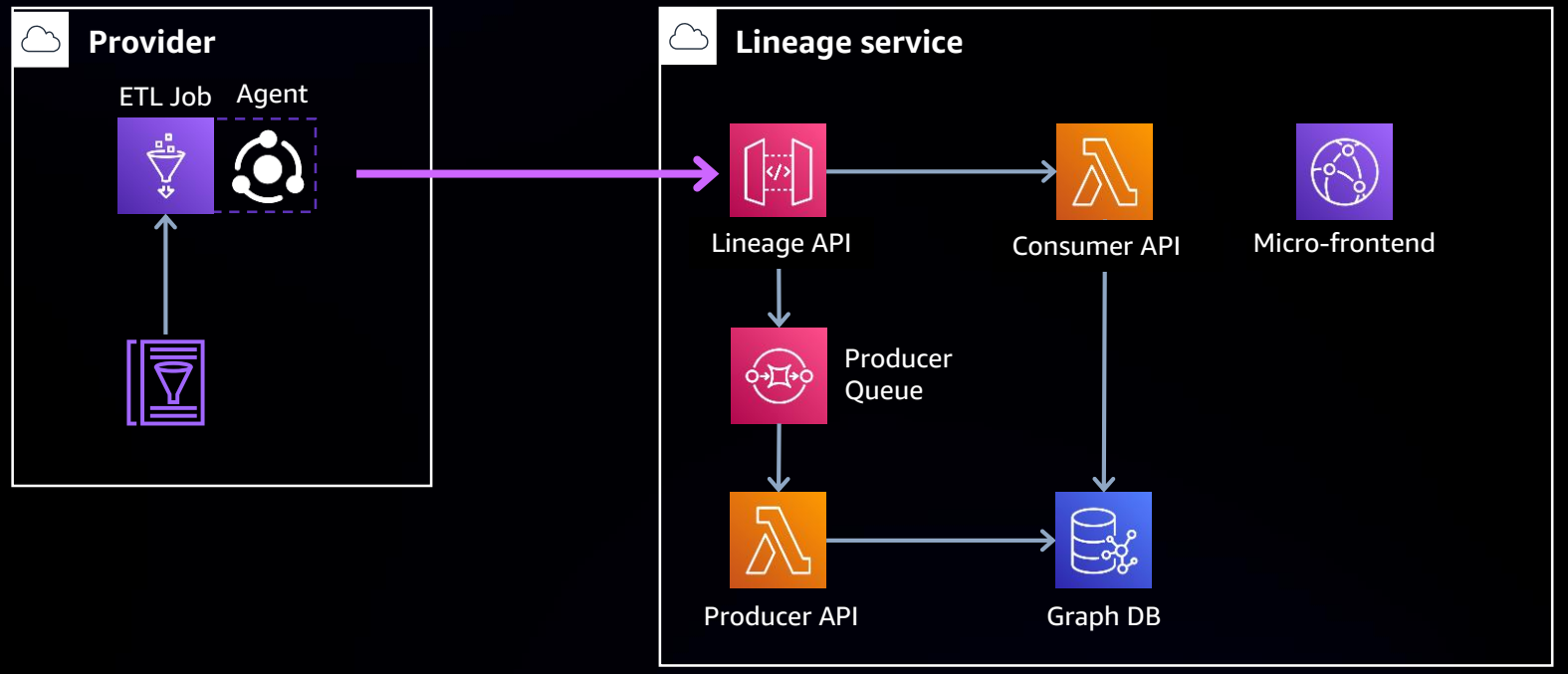


Holistic and interactive visualization

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Easy and pluggable architecture

SOON AVAILABLE AS OPEN SOURCE



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our platform at a glance

SOON AVAILABLE AS OPEN SOURCE

5 Petabyte	**1.2 billion API calls** per month	**Data Portal** as central gateway	
2,600 datasets	**1,200 AWS accounts**	**Built-in SSO** compatible with SAML/OIDC	**10 APIs** API GW and GraphQL/AppSync
Amazon ECS/AWS Fargate	**Amazon OpenSearch Service**	**35 Amazon DynamoDB tables**	**30+ AWS Lambdas**

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Making data universally actionable.

Discover

Search and discover data and use cases.

Explore

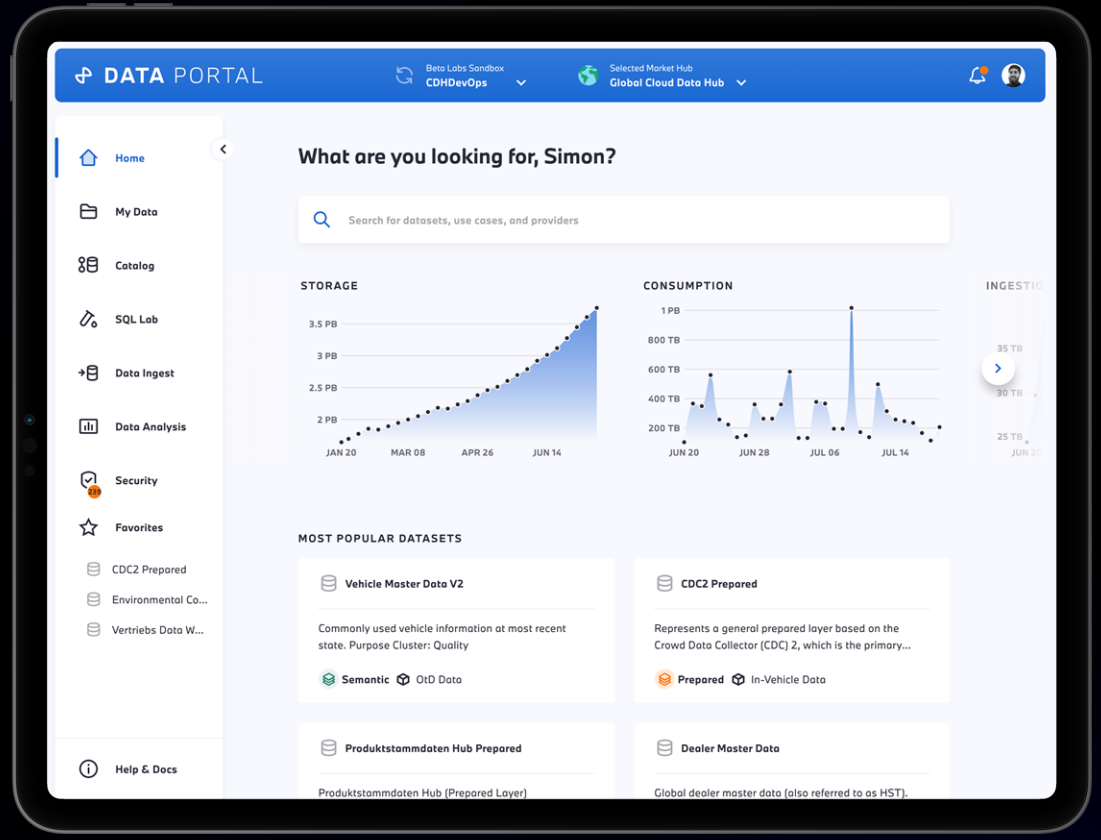
Query datasets via SQL.

Publish

Provide data with modules and UI-based ingests.

Consume

Analyze data with code workbooks.



Manage

Manage metadata and permissions for data products.

Request

Obtain access to data seamlessly.

Monitor

Comprehend data pipelines.

Connect

Integrate systems outside of the platform.

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue: Recap

SERVERLESS DATA INTEGRATION SERVICE

Connect and transform



Built-in data transforms
100s of built-in data transforms



Glue Connectors
connect to and access data where it resides



Easy to migrate
Tools for migration

Centralized & Unified Data Governance



Glue Data Catalog
Scalable, available, and integrated Hive metastore and schema registry



Glue Crawlers
infer the schema, detect drift, and manage versioned metadata

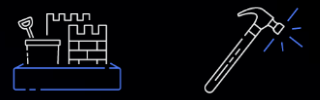


Lake Formation
unified fine-grained permissions and governance

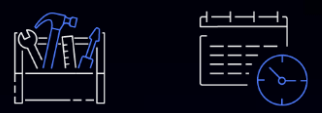
User Productivity and Data Ops



Variety of Persona
to explore data and build jobs and end-to-end workflows

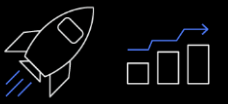


Productivity tools
code generation, templates, and ML tools for data cleaning and optimizations

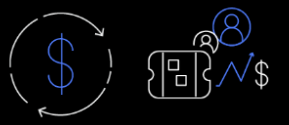


Data ops tools
integrate jobs and workflows into existing processes

Serverless Execution Engine



Scalable Execution engine
scalable, high-performance serverless engine



Cost-effective options
Auto-scaling for batch and streaming



Monitor
critical systems and processes to identify issues and trouble shoot

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.