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

ANT335

How Disney use AWS Glue as a data integration and ETL framework

Ralph Peterkin

Principal Technical Architect
Disney

Alona Nadler

Head of Product
AWS Glue

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

Agenda

What is data integration?

Challenges with data integrations

AWS Glue overview

How Disney chose AWS Glue

How Disney built ETL framework using AWS Glue

Conclusions

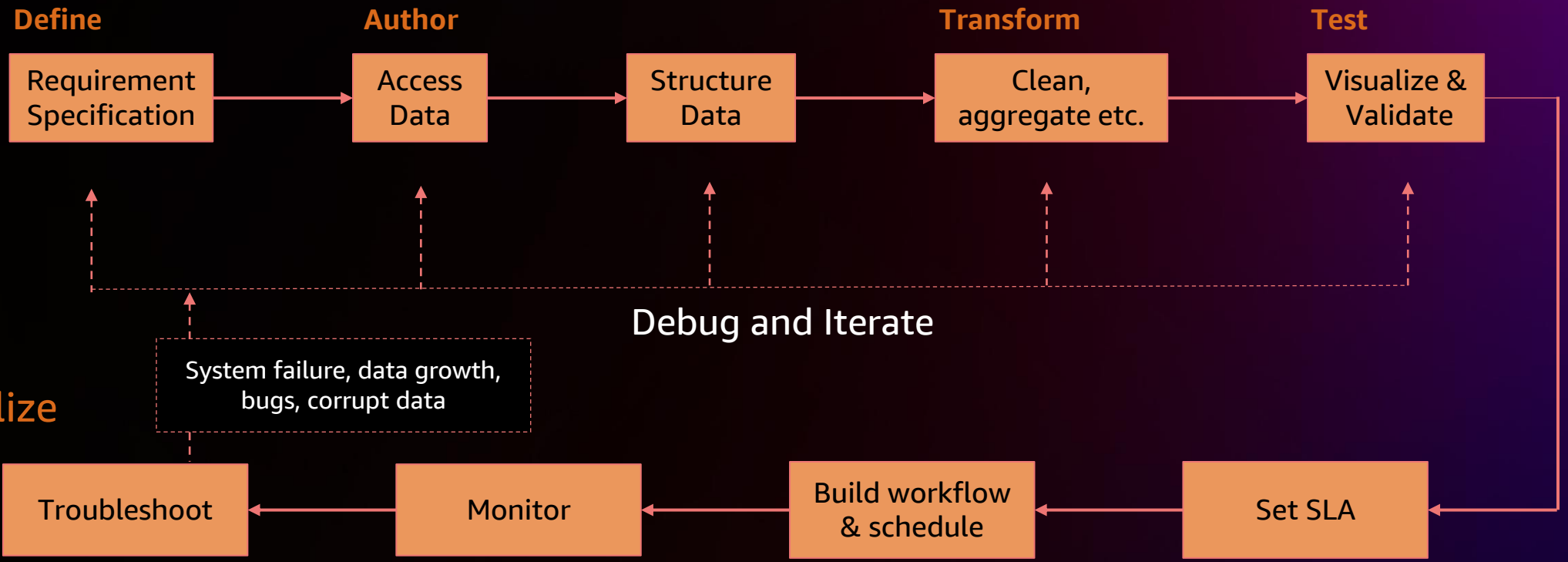
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration process

1) Motivation (use case examples)

New report request **Or** Data pipeline between teams **Or** Research using data from several sources

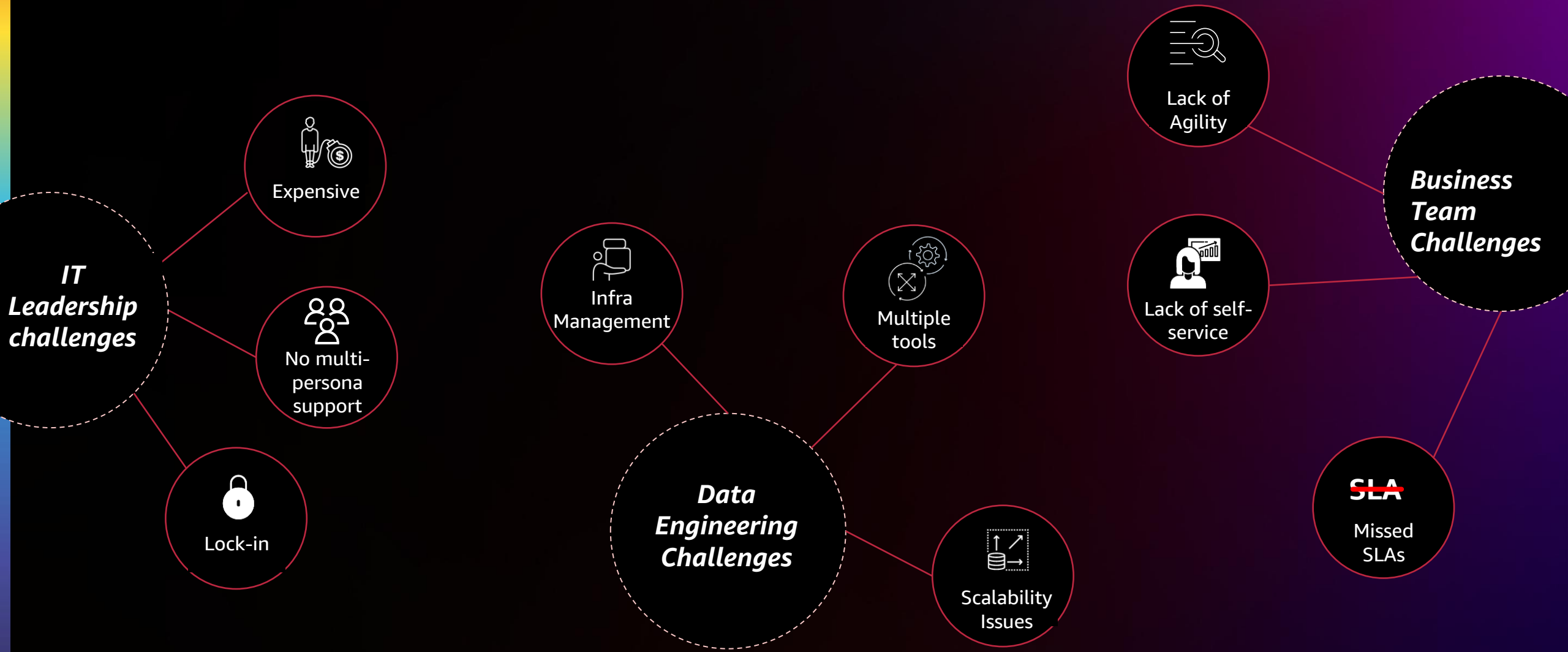
2) Build



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

aws

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

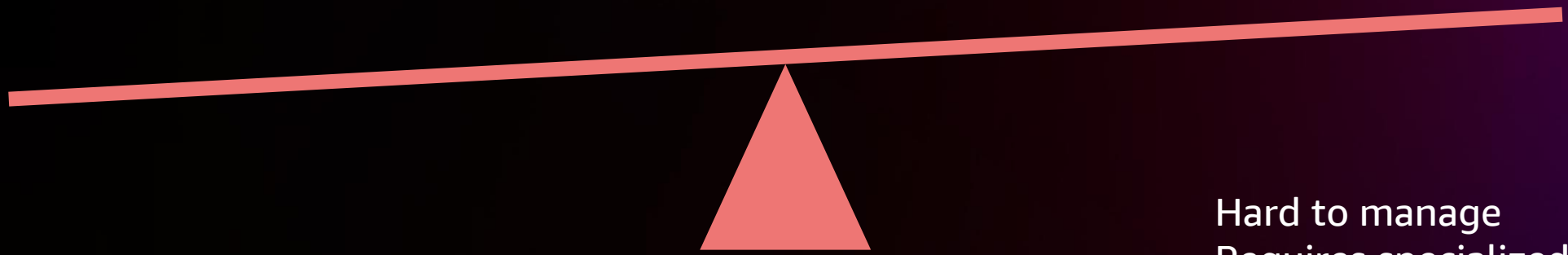
What are the options?

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





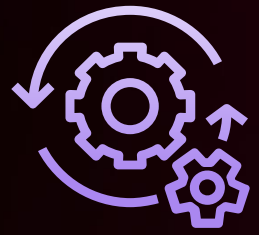
Focus on data

Low maintenance serverless solution



Pay ONLY per use

Avoid licensing cost and infrastructure idle time



Powerful open source engines

No lock in, support from wide innovative ecosystem



All in one

Support all your users 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



Sensitive data detection

Engines



Serverless infrastructure



Choice of data integration engines



Compliance and security

© 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



Sensitive data detection

Engines



Serverless infrastructure



Choice of data integration engines



Compliance and security

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

AWS Glue: Serverless data integration



Serverless infrastructure
No infrastructure to manage



Fast cluster 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.

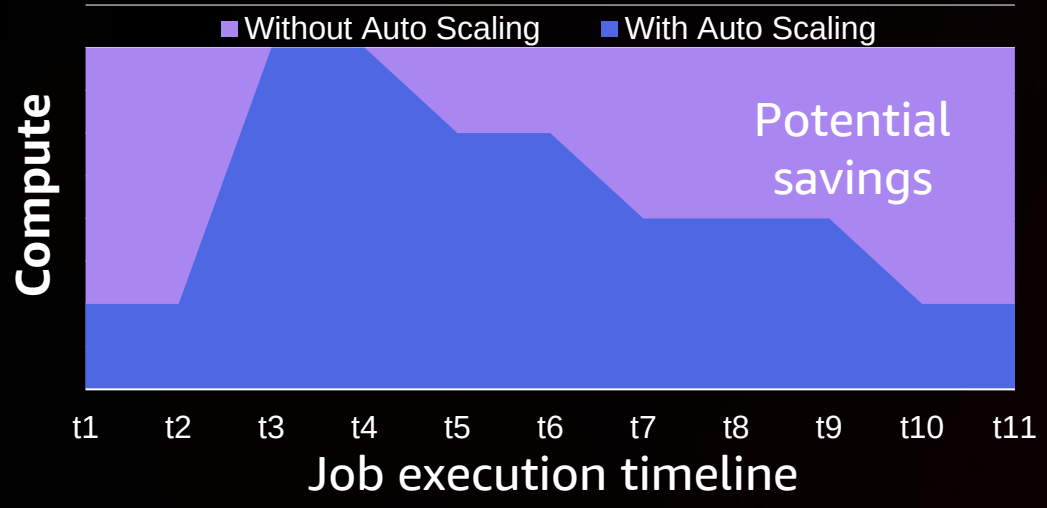
AWS Glue – Cost-effective

AWS GLUE JOB RUNS AT LOWER COST

Recently Released

Auto scaling

Automatically resize compute for lower cost



Reduce cost by 20%-40%

Simplify capacity planning

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.

A Glue engine for every workload

Spark	Python shell	Ray (New)
AWS Glue for Spark	AWS Glue for Python	AWS Glue for Ray

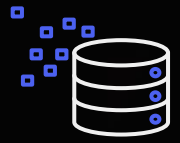
ANT343: AWS Glue for Ray session
12/1 Thursday 4:15 MGM – Grand 121

© 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



Sensitive data detection

Engines



Serverless infrastructure



Choice of data integration engines



Compliance and security

© 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.

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



Sensitive data detection

Engines



Serverless infrastructure



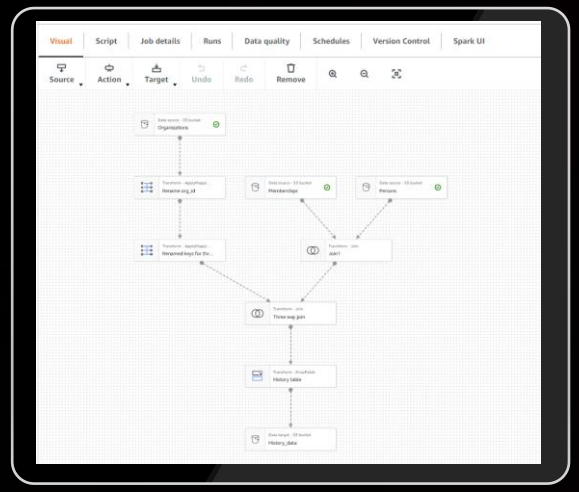
Choice of data integration engines



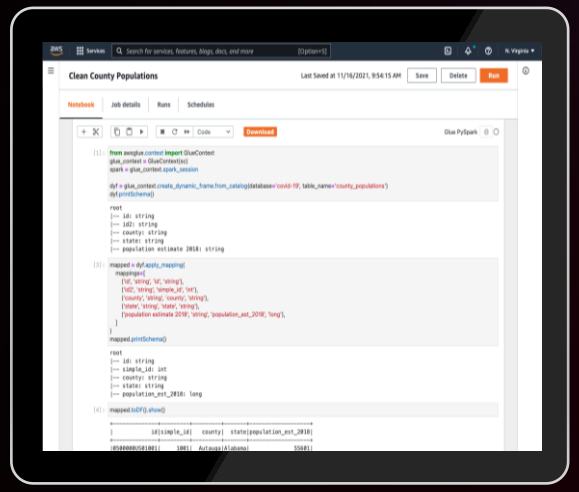
Compliance and security

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

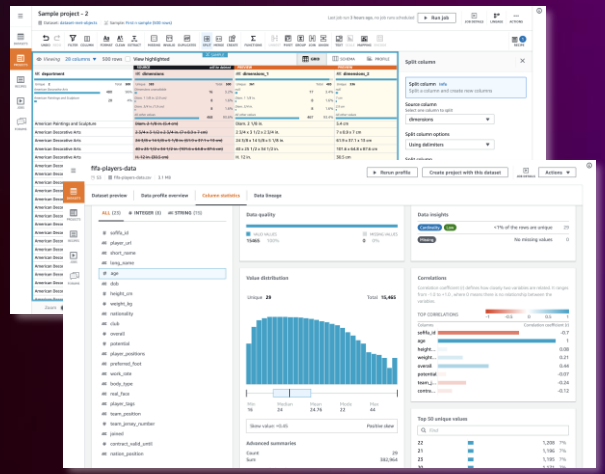
Interface for every persona



Glue Studio



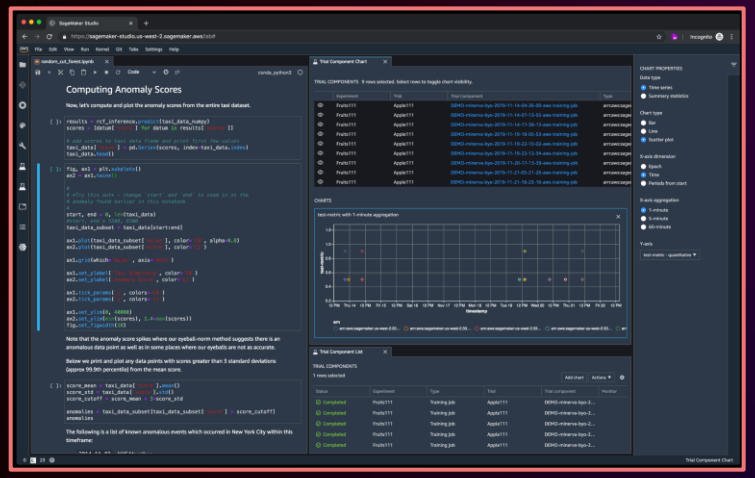
Glue Studio Notebook



Glue DataBrew



APIs, SDK, and Local Notebook

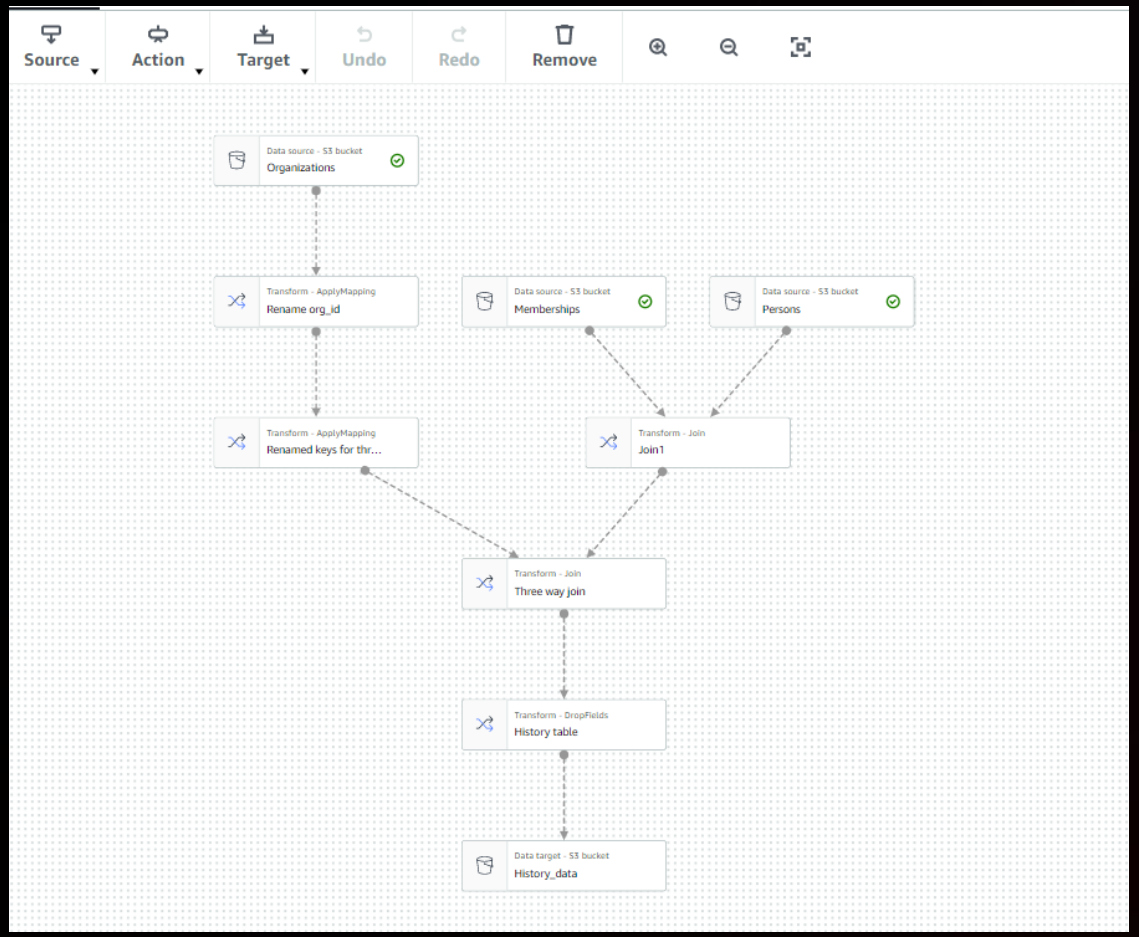


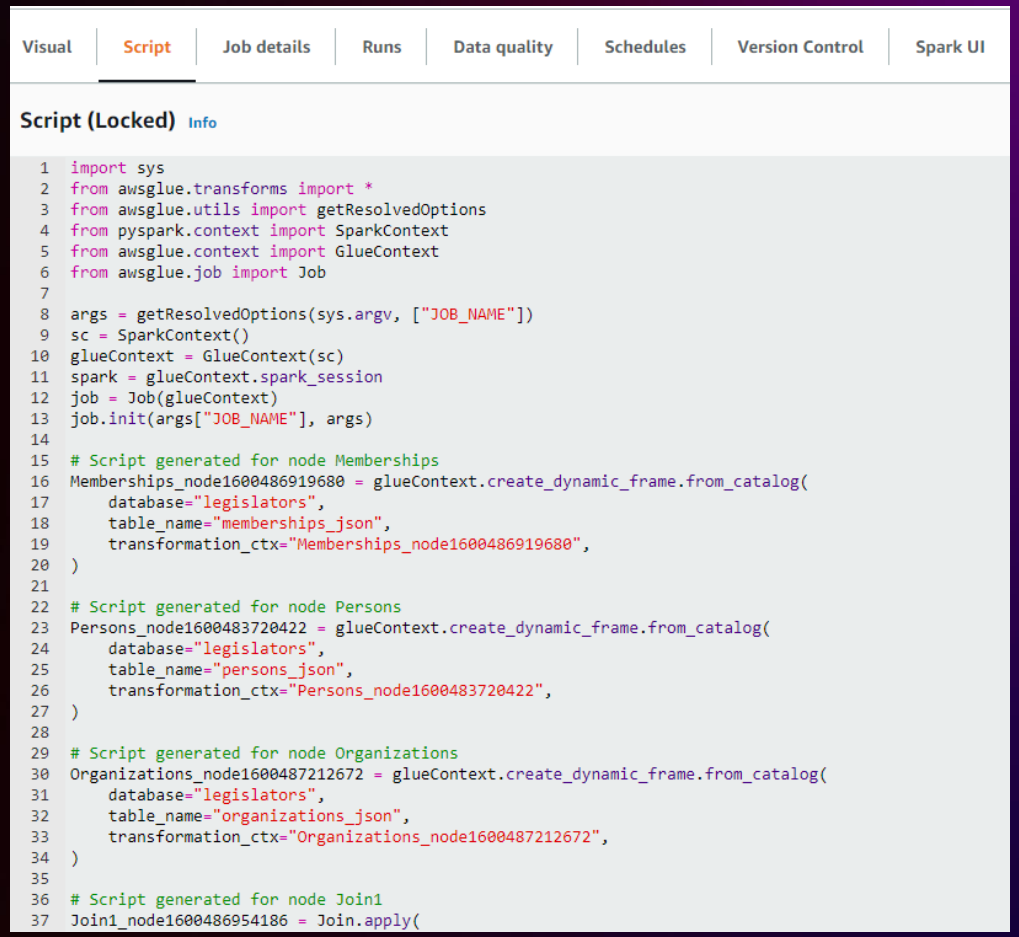
Amazon SageMaker Studio notebook

aws

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

Glue Studio generates a script





© 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



Sensitive data detection

Engines



Serverless infrastructure



Choice of data integration engines

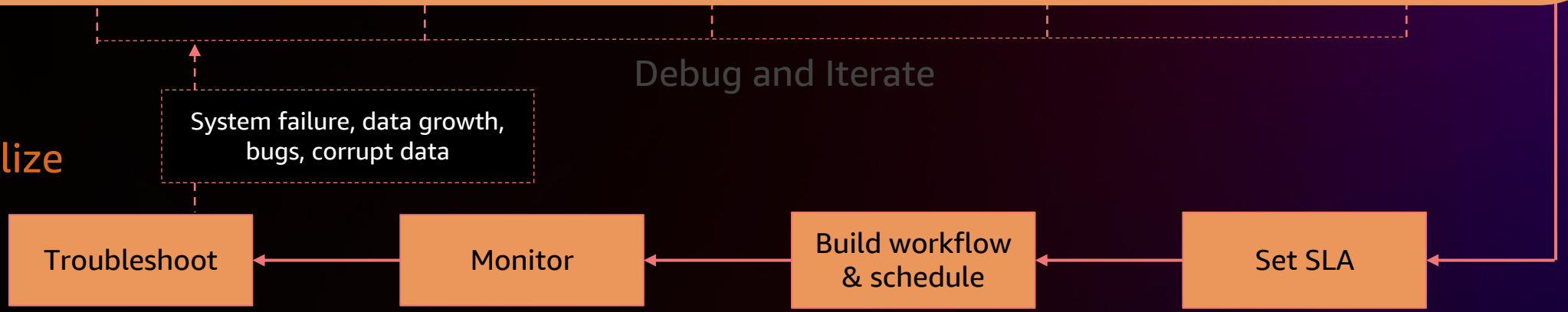


Compliance and security

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

Data integration process

Data engineers spend 60% - 80% of their time operationalizing data pipelines



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

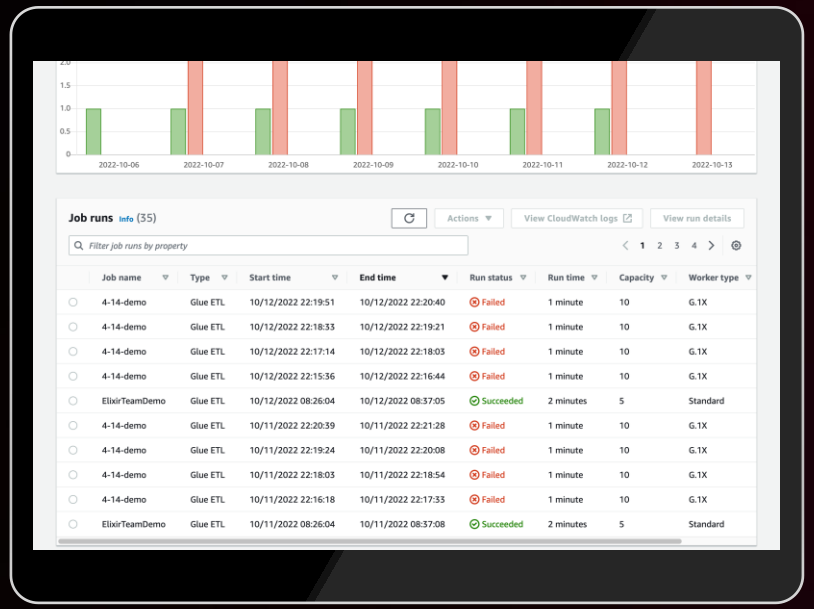
AWS Glue Operation

GitHub

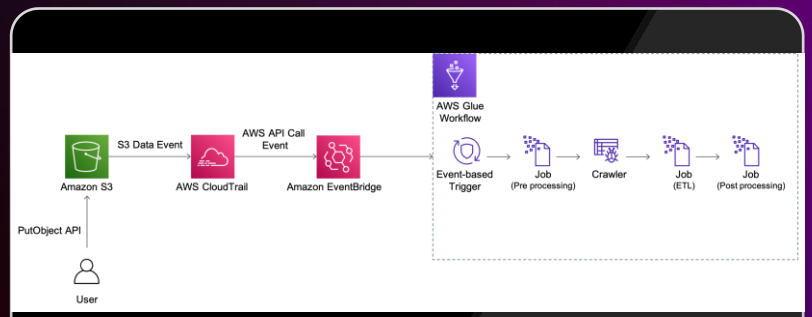
AWS Glue

AWS CodeCommit

Git integration



Job monitoring



Workflow & Orchestration

© 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



Sensitive data detection

Engines



Serverless infrastructure



Choice of data integration engines

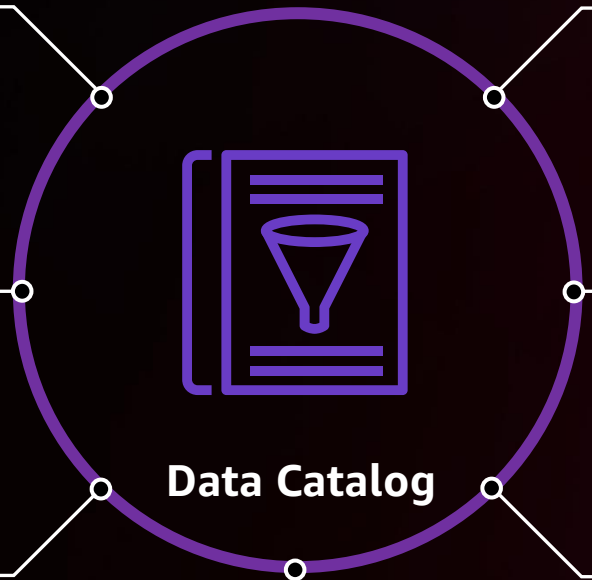


Compliance and security

aws

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

Data Catalog provides metadata store for data lakes



Great **durability, availability,** and **scalability**

Serverless and **cost-effective**

Many ways to **get metadata in**

Hive metastore API-compatible

Schema Registry Kafka integration

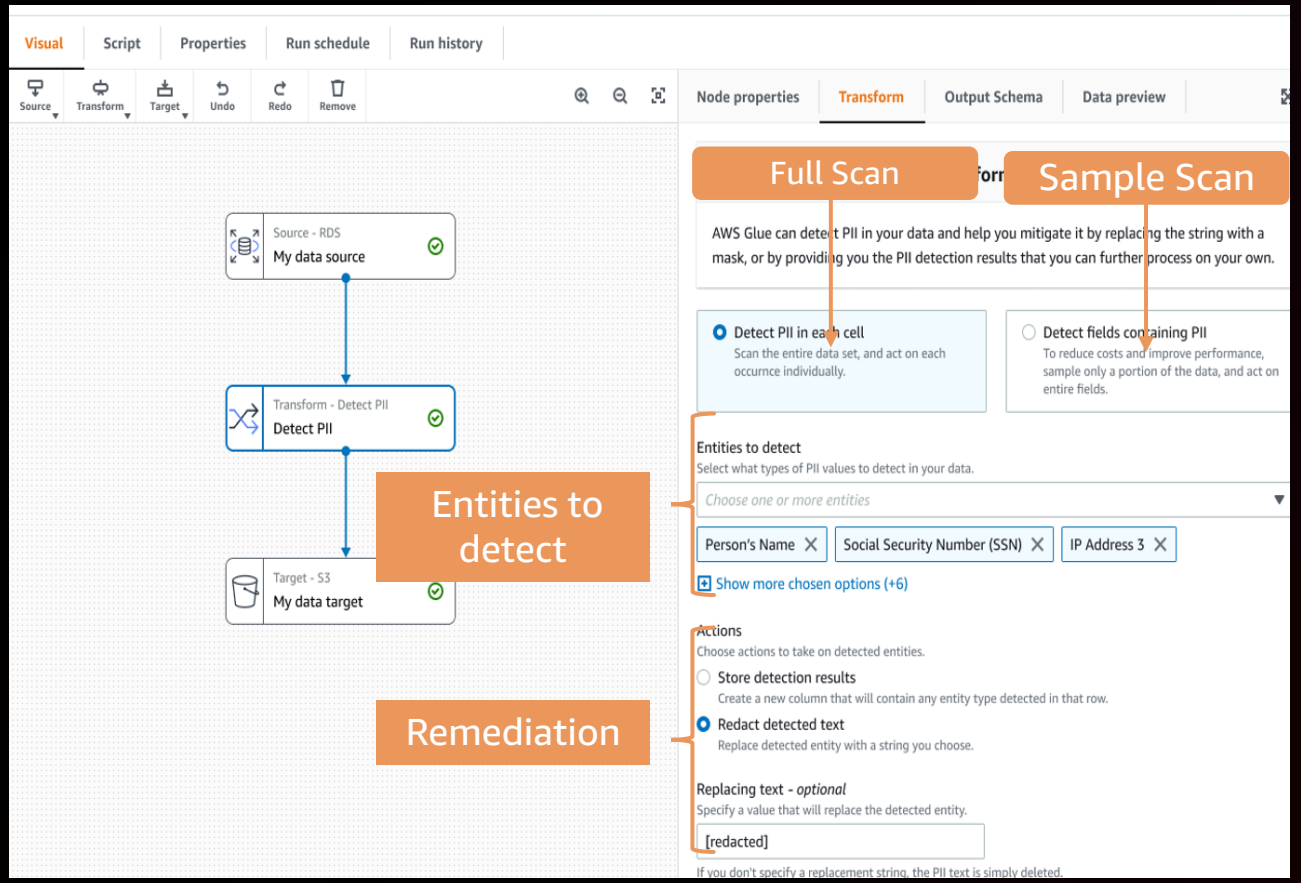
Broad portfolio of integrated tools

Top-tier **security, compliance,** and **audit** capabilities

aws

© 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, and mask sensitive data

Detect across TB of data in scalable way

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

How customers use AWS Glue

COMMON USE CASES

Migrate from traditional ETL

Gain flexibility and reduce cost

Data lake data integration

Easiest way to clean and transform data

Self service data integration

Allow various users in the company to have access to data

ETL / ELT from data warehouse

Extract structure and unstructured data, transform and load to get insights

Leverage powerful Apache Spark engine

Process PB of data without the maintenance burden

Data preparation for ML

Interactive tools for data preparation

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

Agenda

Why data integration?

Challenges with data integrations

AWS Glue overview

How Disney chose AWS Glue

How Disney built ELT framework using AWS Glue

Conclusion

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

Disney Parks, Experiences and Products



Disney Parks and Experiences :

One of the world's leading providers of family travel and leisure experiences

- 6 resort destinations across the globe
- 12 theme parks
- 53 resort hotels
- 5 cruise ship
- Guided tours

Disney Consumer Products, Games, and Publishing:

The division products brings beloved brands and franchises into the daily lives of families and fans through products

- Retail
- Licensees
- Publishing
- Games
- Global Products

aws

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

Covid 2020

Disney Parks began managing capacity of guests in parks due to social distancing requirements.

New data requirements to generate insight into current and future bookings in order to optimize park occupancy

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

What Existed Prior to AWS Glue

Hadoop cluster running on Amazon EC2

- HDFS
- YARN
- HIVE
- SPARK
- Sqoop
- MapReduce

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

Existing challenges with Hadoop



Securing a cluster is hard



Cluster was close to capacity



Scaling was not easy - it took a lot of planning and coordination

aws

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

Questions we needed to answer

1. Given the challenges, should we continue with existing systems?

2. What is the future of the new projects?

3. Will data engineers be able to use existing skills with Spark?



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

Why Choose AWS Glue?

CRITERIA FOR NEW SYSTEM

SERVERLESS



FAST TO MARKET



SPARK-BASED



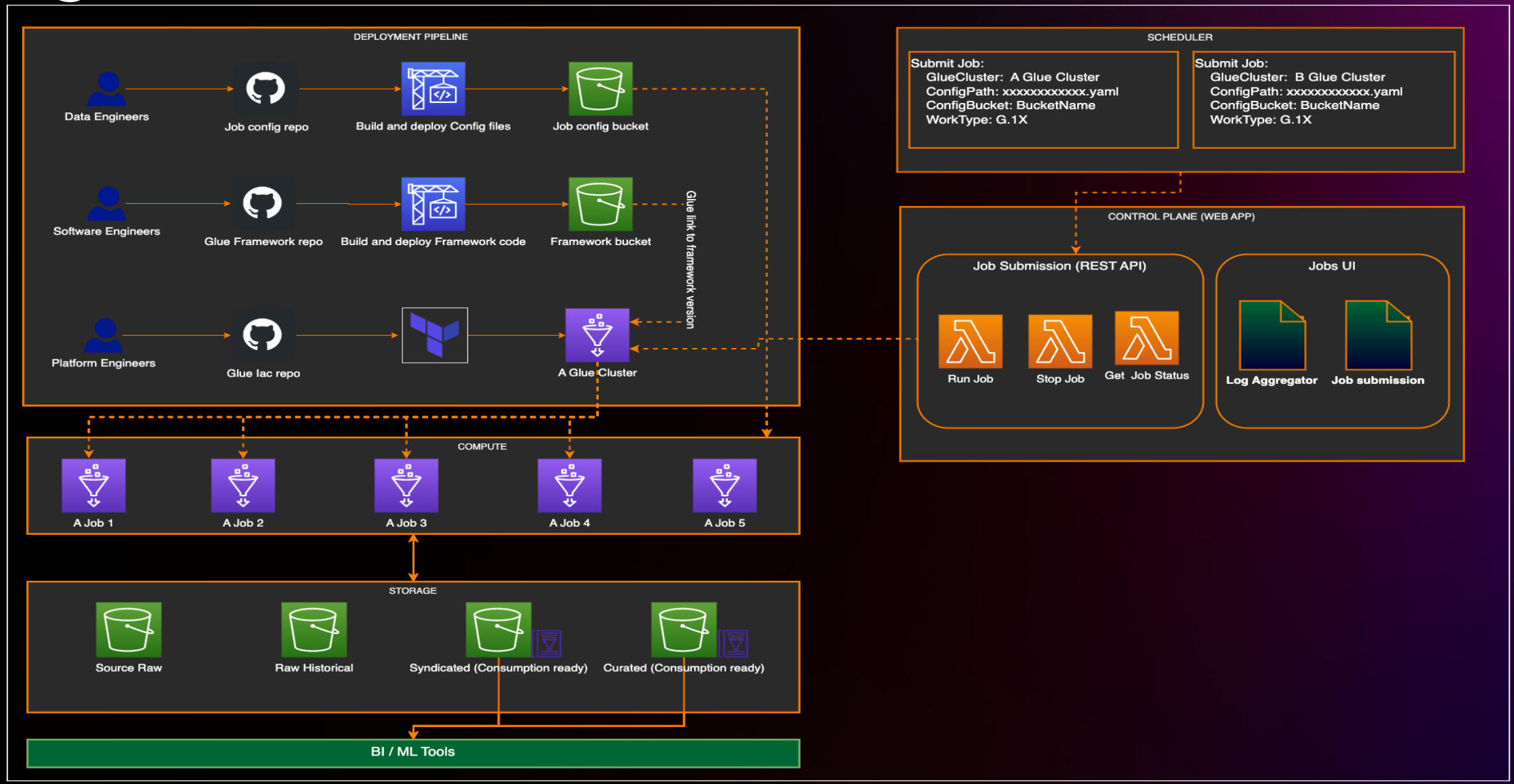
COST EFFICIENT



aws

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

High-Level Architecture



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

Motivation and goals

- Required a standard way to handle compliance and data with different sensitivity levels.
- Simplicity for users: As simple as applying a tag to a dataset.
- Provide data engineers end to end tools without needing to understand every low-level implementation
- Existing users included many SQL developers that had no prior Python or Spark experience.

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

Internal Glue Framework Components

Internal Glue framework

- A single pyspark codebase that uses YAML as a declivity language to, run, and track AWS Glue jobs without having the need to write new spark code.

Job runner

- This is the entry point for our AWS Glue jobs. It sets up the Spark sessions and runs jobs.

Job Tracker

- Used to keep track of files that are successfully processed.

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

Glue Framework

Who uses the AWS Glue framework?

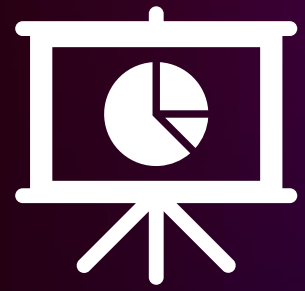
Data Engineer



Data Scientist



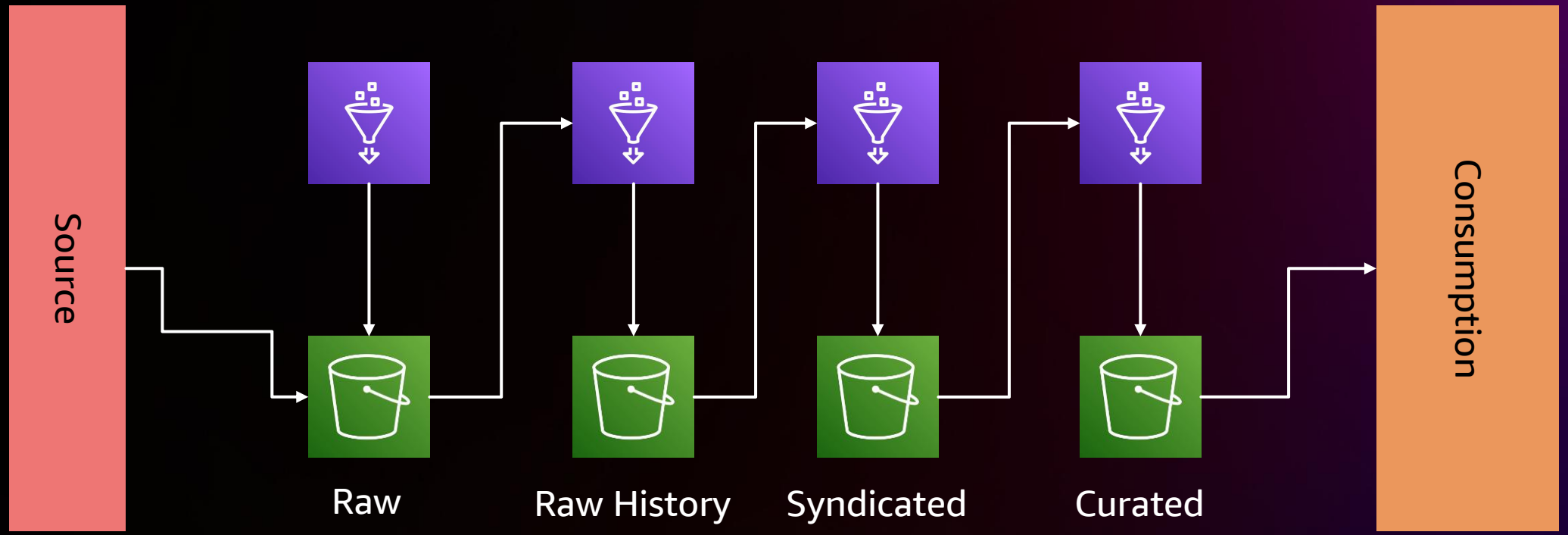
Data Analyst



aws

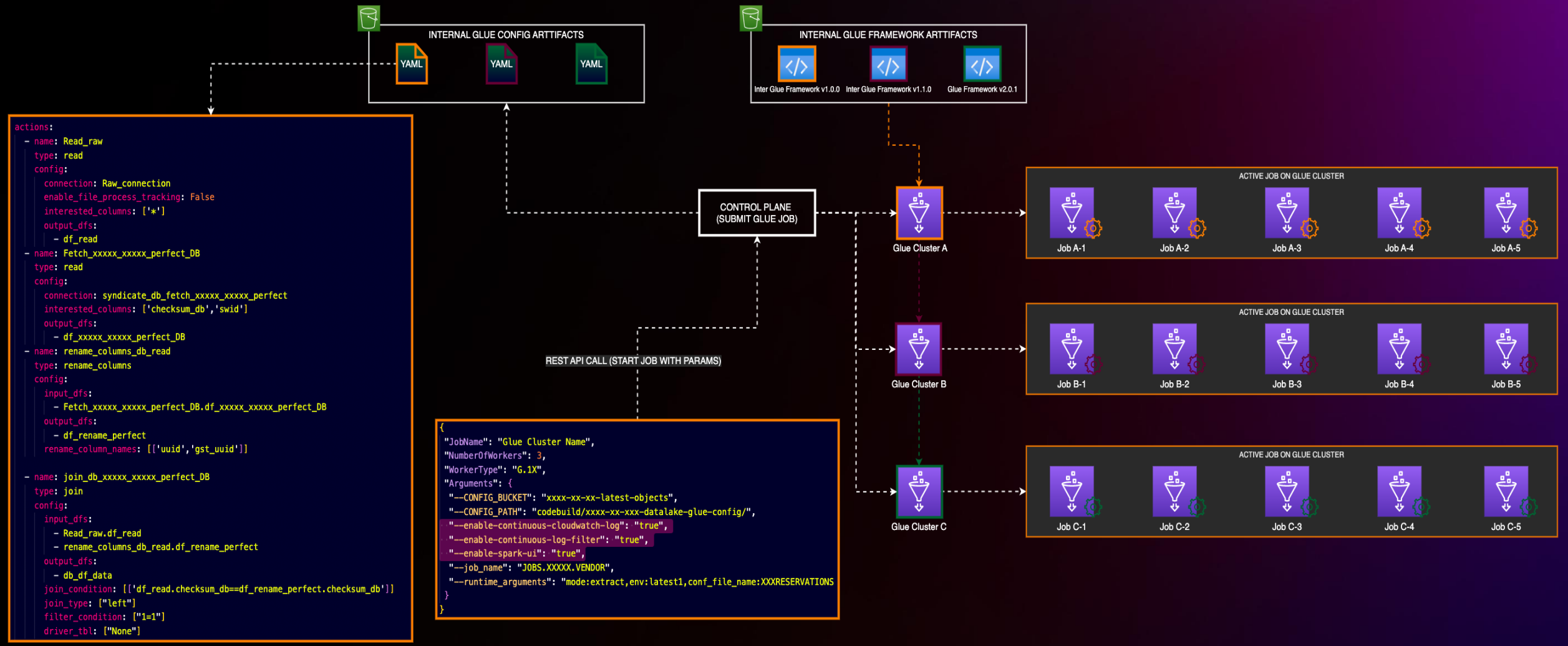
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Dataflow



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

High-Level Glue Job Runner Architecture



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

Configs

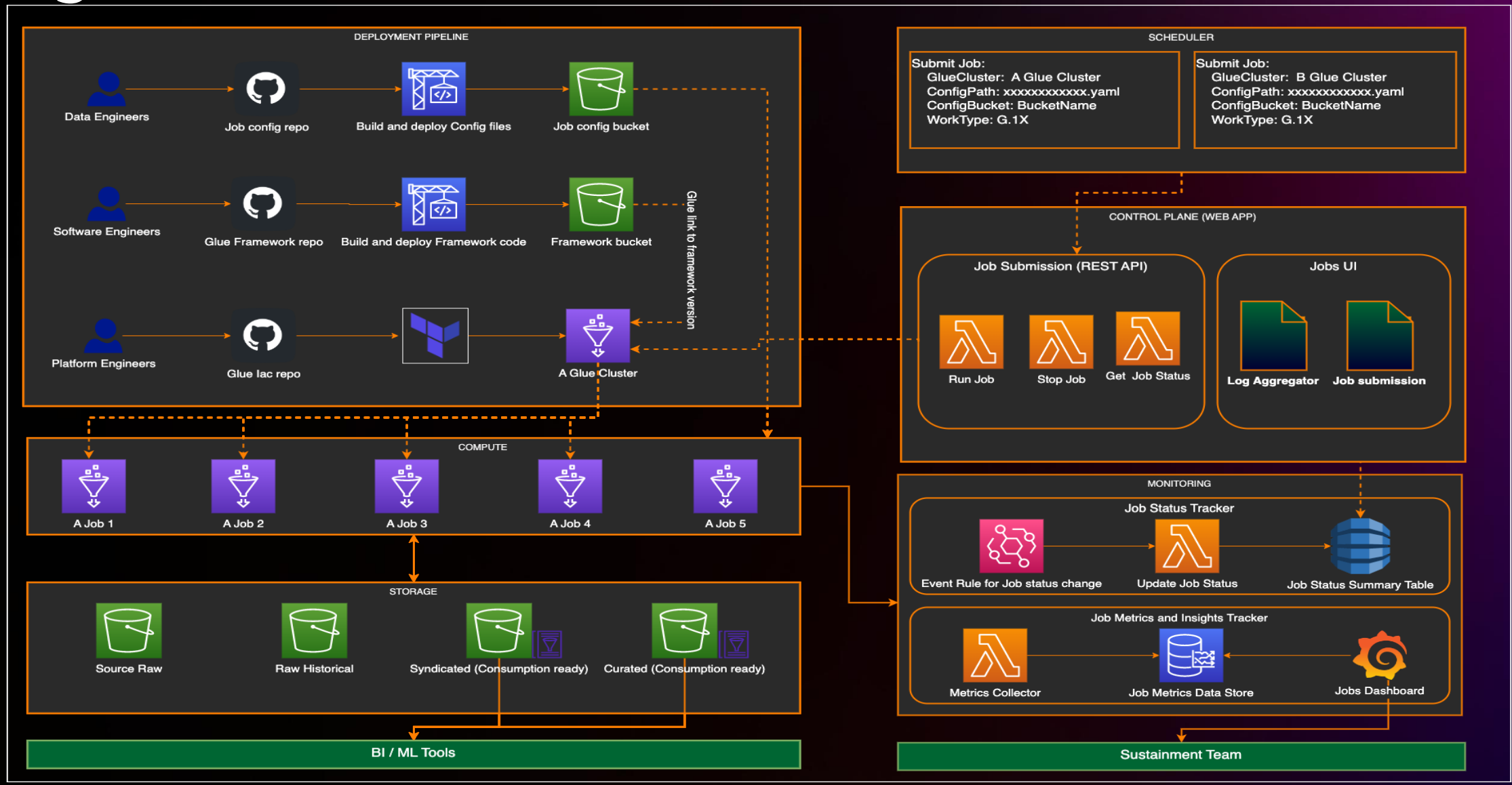
```yaml
connections:
  - name: Raw_connection
    type: s3.parquet
    config:
      bucket: xxxx-xx-xxxx-raw
      prefix: '{env}/syndication/xxxxx/ETL_DATE=*/data/ETL_DATE=*/'
      header: 'true'
  - name: Raw_history_write
    type: s3.parquet
    config:
      bucket: xxxx-xx-xxxx-datalake
      prefix: '{env}/syndication/xxxxx/db_perfect'
      mode: overwrite
  - name: syndicate_write
    type: s3.parquet
    config:
      bucket: xxxx-xx-xxxx-datalake
      prefix: '{env}/curated/xxxxx/db_perfect'
      mode: overwrite
  - name: rawhist_db_write
    type: rdbms.aurora
    config:
      table: raw_history.xxxxx_xxxxx_db_perfect
  - name: syndicate_db_write
    type: rdbms.aurora
    config:
      table: syndicate.xxxxx_xxxxx_db_perfect
  - name: syndicate_db_fetch_xxxxx_xxxxx_perfect
    type: rdbms.aurora
    config:
      aurora_query: select * from syndicate.xxxxx_xxxxx_perfect
```

```yaml
actions:
  - name: Read_raw
    type: read
    config:
      connection: Raw_connection
      enable_file_process_tracking: False
      interested_columns: ['*']
      output_dfs:
        - df_read
  - name: Fetch_xxxxx_xxxxx_perfect_DB
    type: read
    config:
      connection: syndicate_db_fetch_xxxxx_xxxxx_perfect
      interested_columns: ['checksum_db','uuid']
      output_dfs:
        - df_xxxxx_xxxxx_perfect_DB
  - name: join_db_xxxxx_xxxxx_perfect_DB
    type: join
    config:
      input_dfs:
        - Read_raw.df_read
        - rename_columns_db_read.df_rename_perfect
      output_dfs:
        - db_df_data
      join_condition: [['df_read.checksum_db==df_rename_perfect.checksum_db']]
      join_type: ["left"]
      filter_condition: ["1=1"]
      driver_tbl: ["None"]
  - name: User_Sql
    type: user_sql
    config:
      input_dfs:
        - De_Duplication.df_de_duplication
      output_dfs:
        - df_user_sql
      table_name: ['df','perfect','dta']
      query: ['SELECT * XXXXXXXXXX']
```

aws

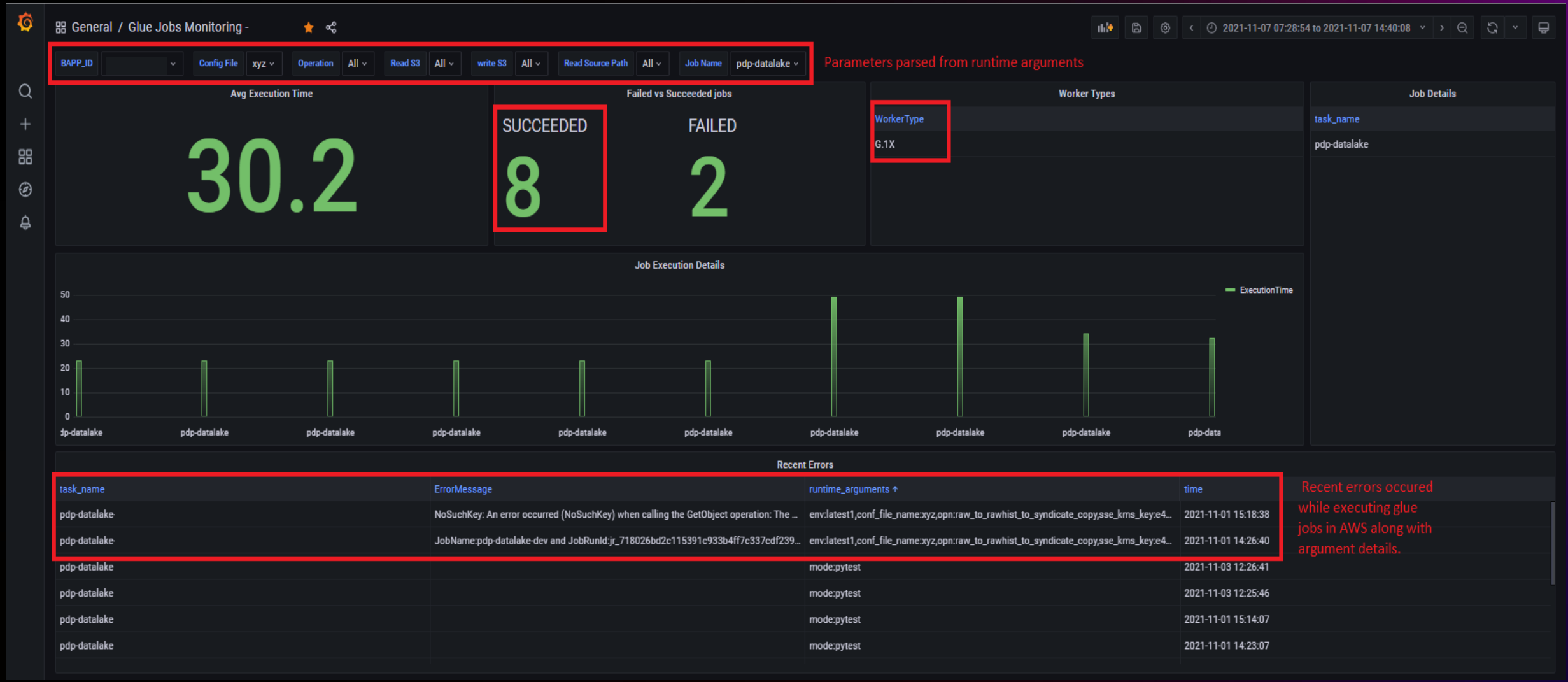
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

High-Level Architecture



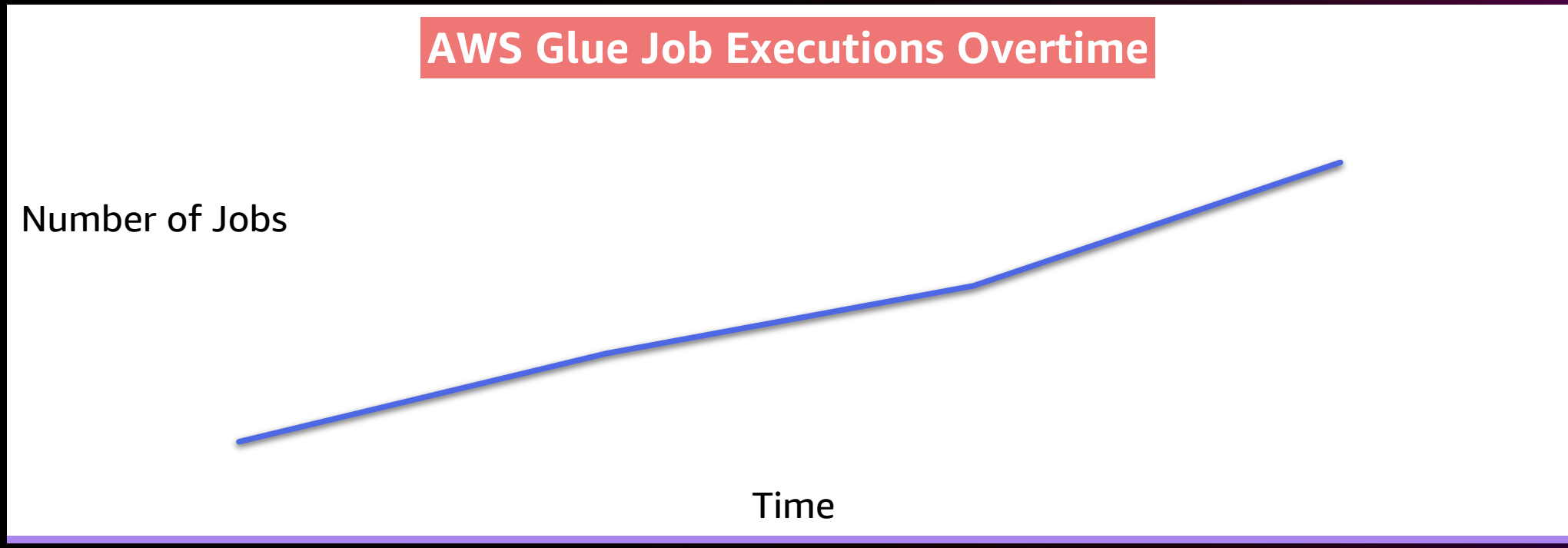
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Monitoring Glue Clusters



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

Where we are now



How we started

Covid workloads

A year in

Started Hadoop job migration

Present

Migrated Hadoop jobs while adding new jobs

aws

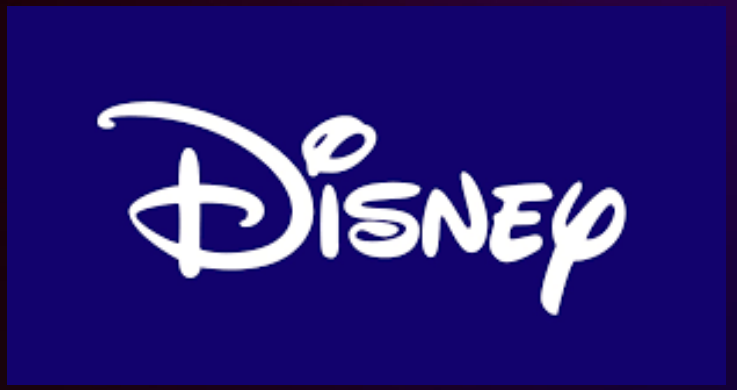
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Conclusion

AWS Glue has given us the ability to scale our workload.

We can now focus our engineering resources on building new data products instead of maintaining Hadoop clusters.

AWS Glue has transformed how we spin up analytical compute environments - new analytical platform in Disney Land Paris within a couple months.



aws

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

Thank you!

Alona Nadler
alonanad@amazon.com

Ralph Peterkin

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