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

AIM322

# Accelerate data preparation with Amazon SageMaker Data Wrangler

Arunprasath Shankar (he/him)
Senior AI/ML Solution Architect
AWS

Huong Nguyen (she/her)
Senior Product Manager
AWS

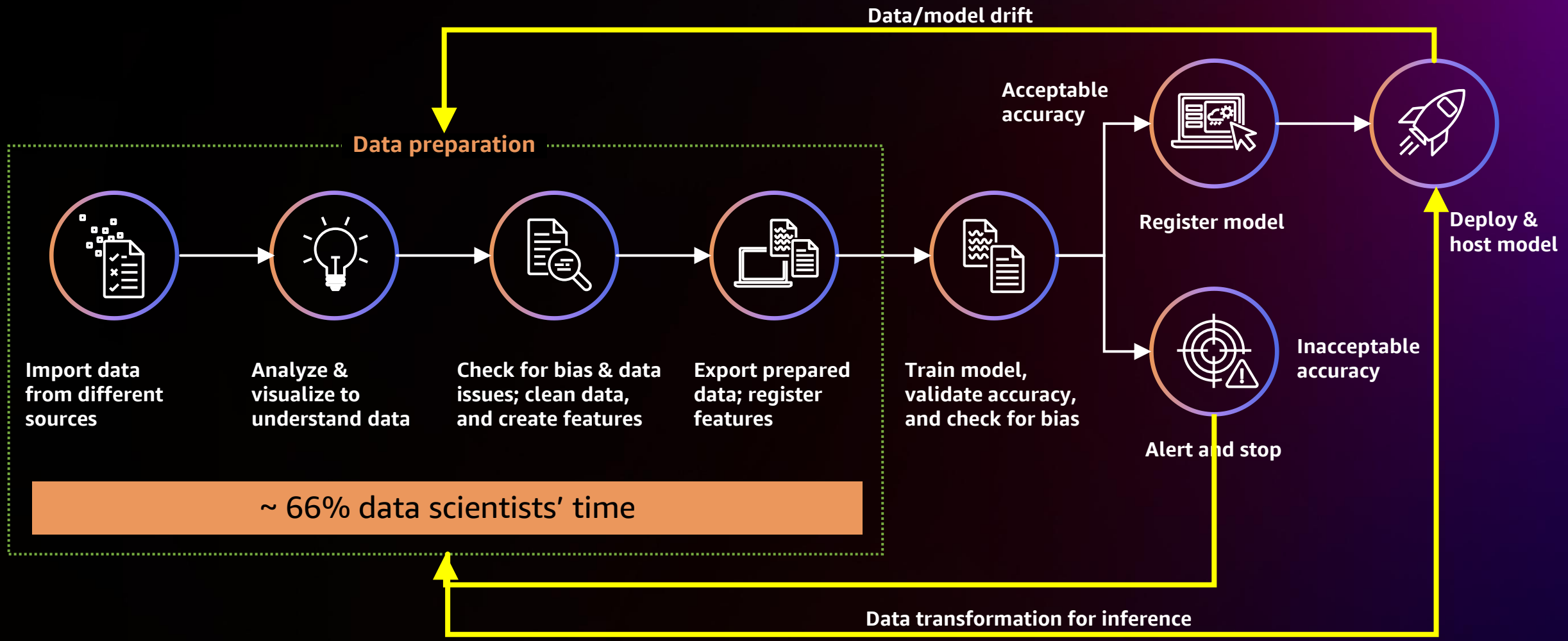
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agenda

- Common challenges with data preparation for machine learning (ML)
- Amazon SageMaker Data Wrangler capabilities
- Demo
- Q&A

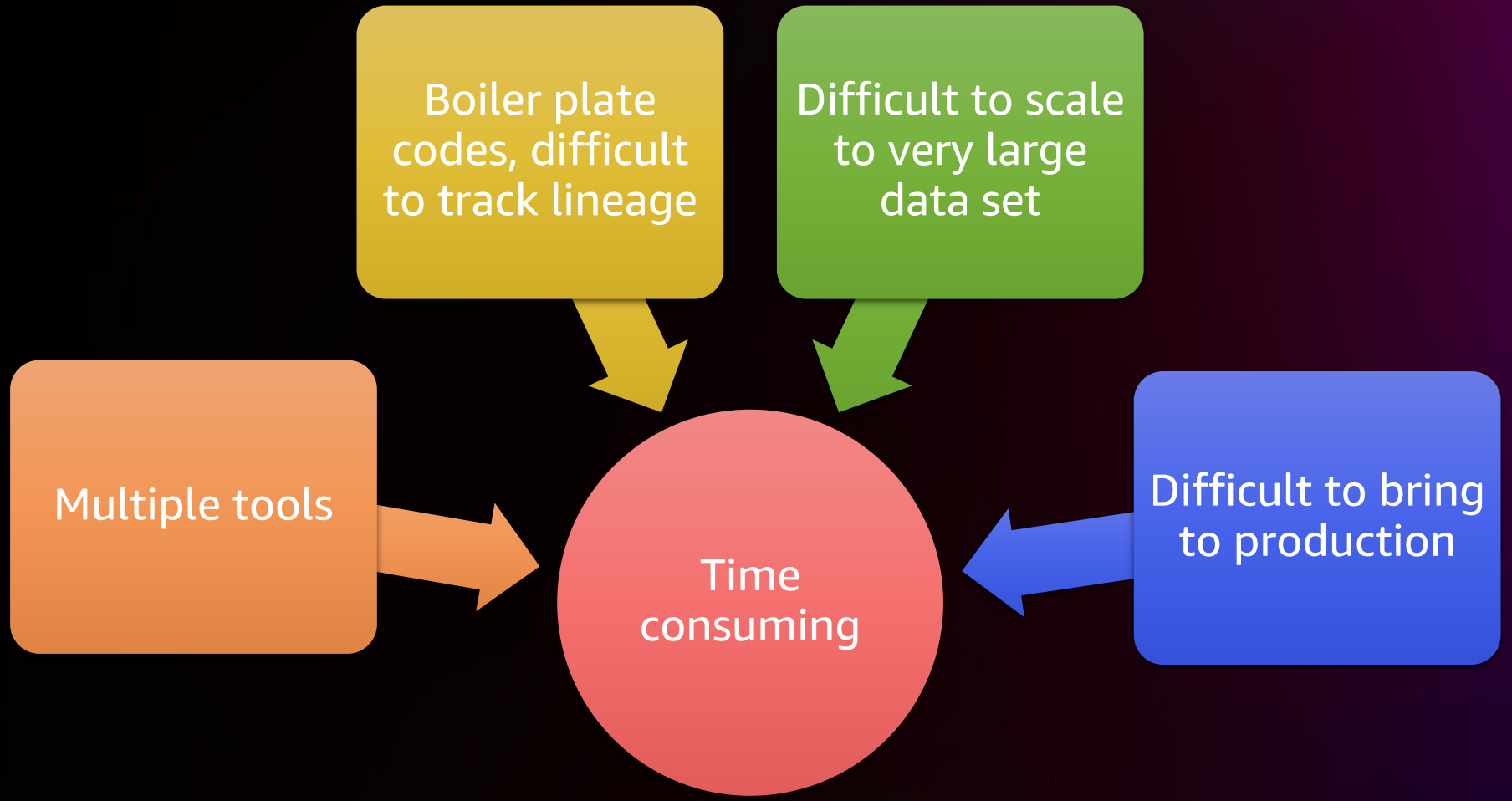
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Data prep: A critical part of the complete ML workflow




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

# Data preparation: Common challenges




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

# Amazon SageMaker Data Wrangler

The fastest and easiest way to prepare data for ML



## Single application for end to end data prep

Data import, analysis, cleansing, feature engineering & production



## No code / low code

Visual data analyses, 300+ built-in transforms, custom code library



## Scalable to very large data set

Backed by serverless Spark processing jobs



## Easy to operationalize

Schedule jobs, register features with SageMaker Feature Store, automate data prep workflows with SageMaker Pipelines

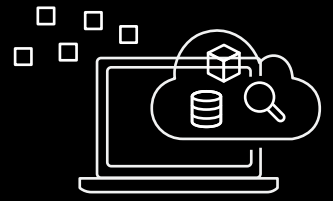
aws

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

# How SageMaker Data Wrangler works

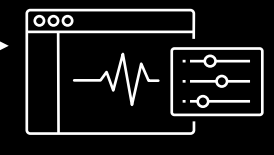
**Amazon SageMaker Data Wrangler**

A faster, visual way to aggregate and prepare data for machine learning



**Select & query**

from a wide variety of data sources using SQL



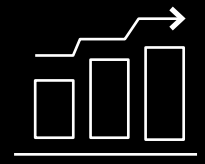
**Understand**

Detect quality issues and bias with built-in visualizations, analyses, & Amazon SageMaker Clarify



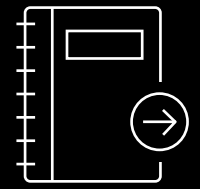
**Cleanse & enrich**

Perform feature engineering with 300+ built-in data transforms



**Improve accuracy**

Estimate model performance, accuracy & diagnose potential issues



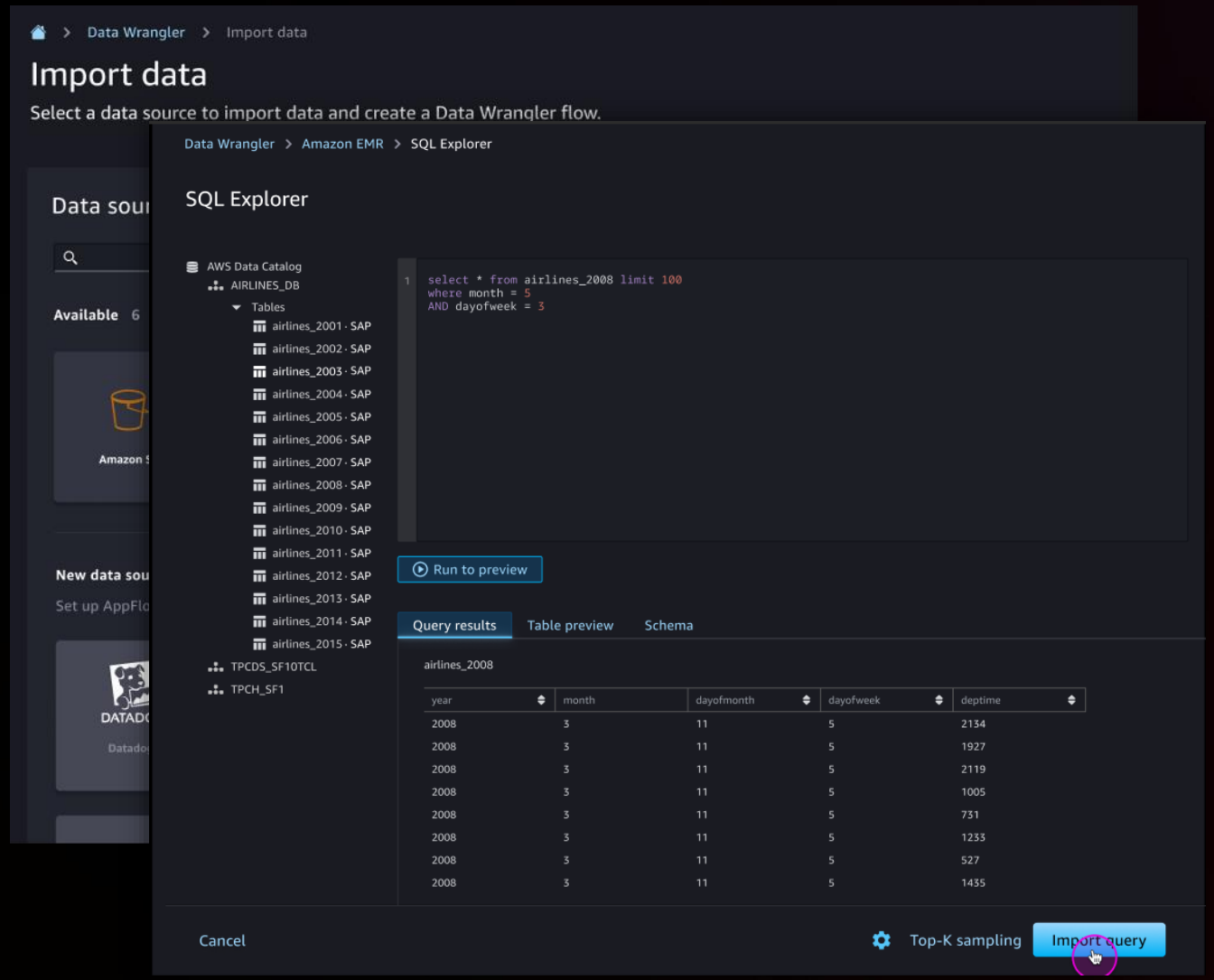
**Operationalize**

Export data, run or schedule job, deploy data flow for inference, automate for ML production

**Interactive** → **Production**

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

# Quickly connect to data sources

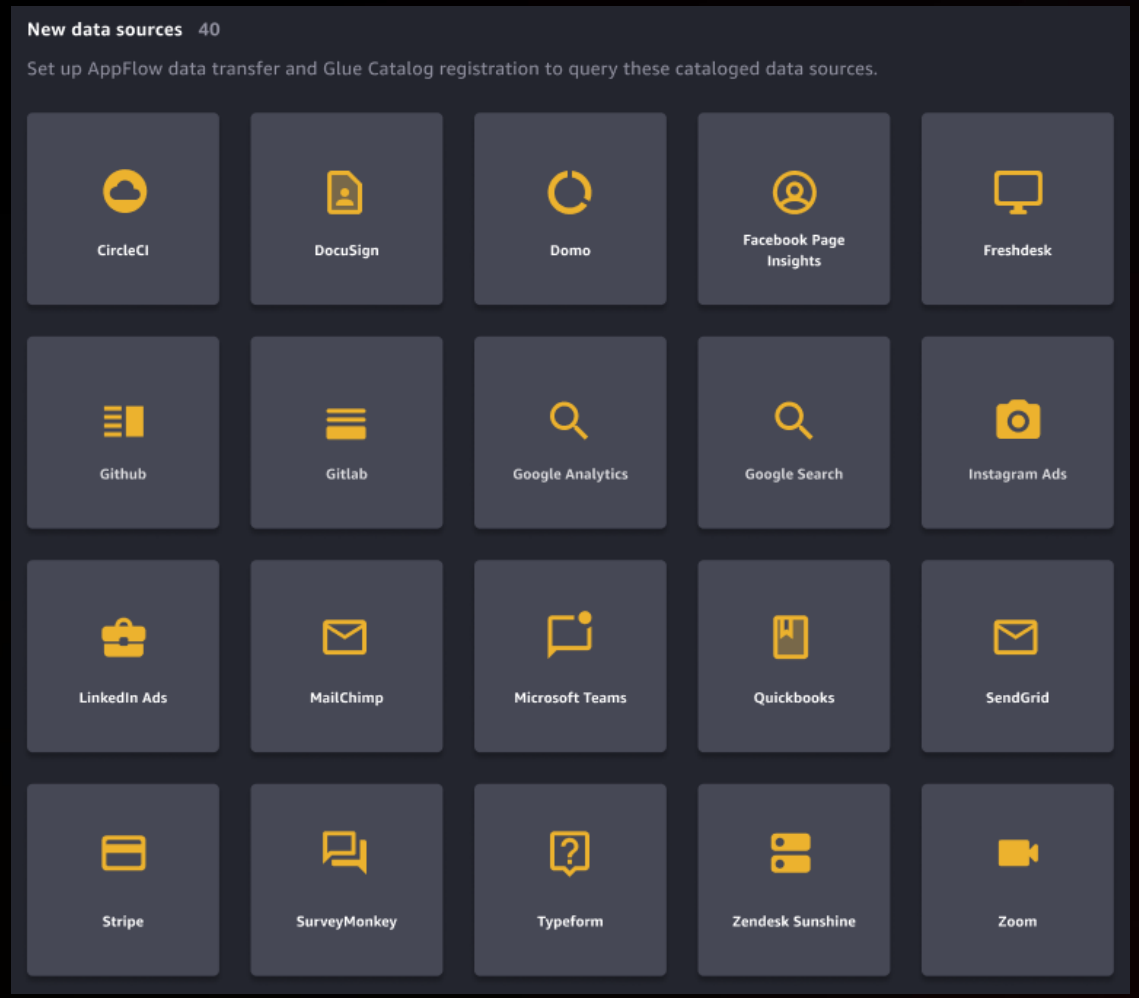



- **Connect to multiple data sources**
  - Amazon S3, Amazon Athena, Amazon Redshift, Snowflake, Databricks
  - Salesforce Genie (Preview)
  - **NEW** Amazon EMR
- **Support for common file formats:**
  - CSV, Parquet, JSON and JSONL, ORC, Database tables
- **Explore data with S3 browser or SQL explorer, and sample data**
  - **Sampling strategies: top-k, random, stratified**

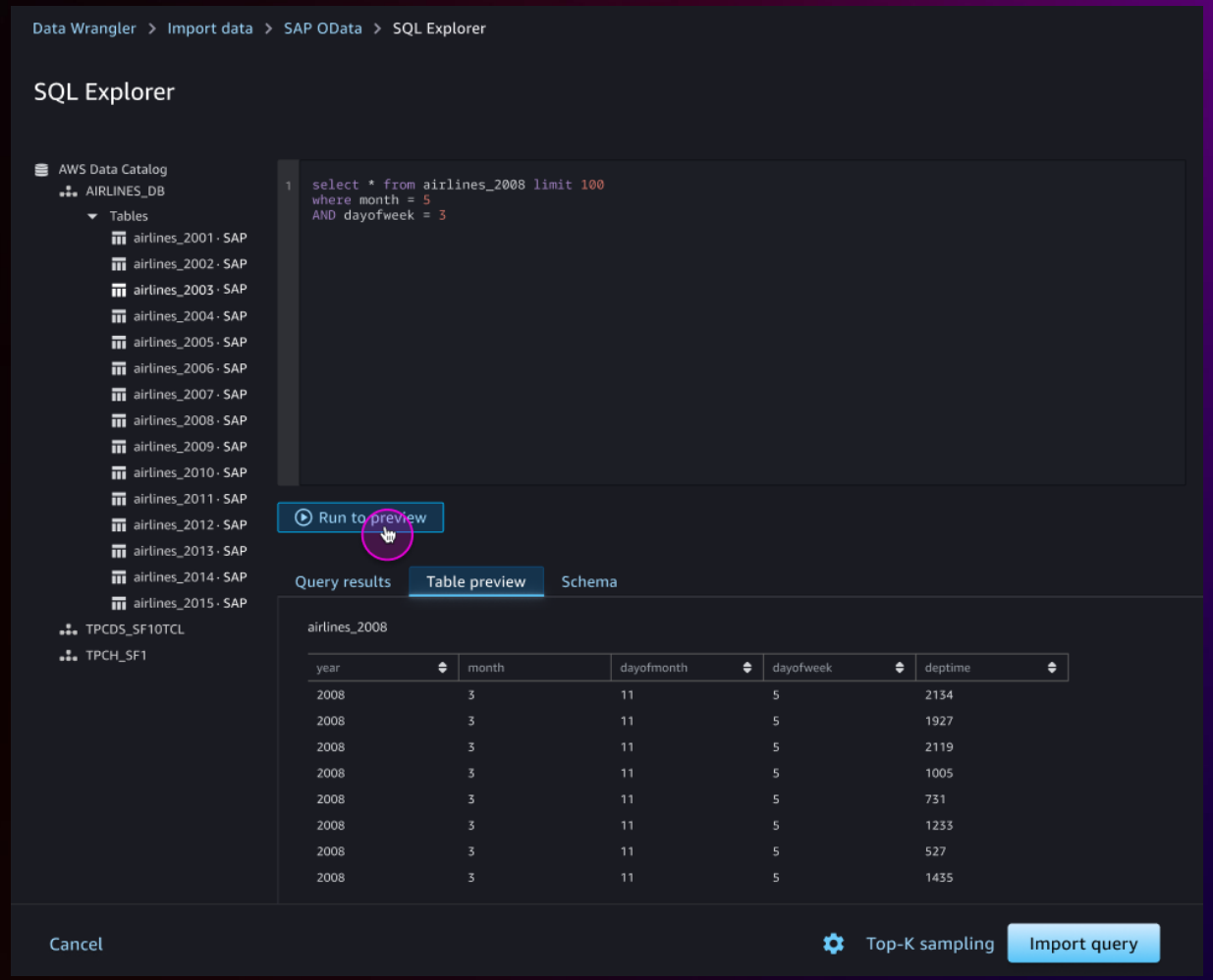
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# New: Support 40+ SaaS data sources

## Set up data sources via Amazon AppFlow

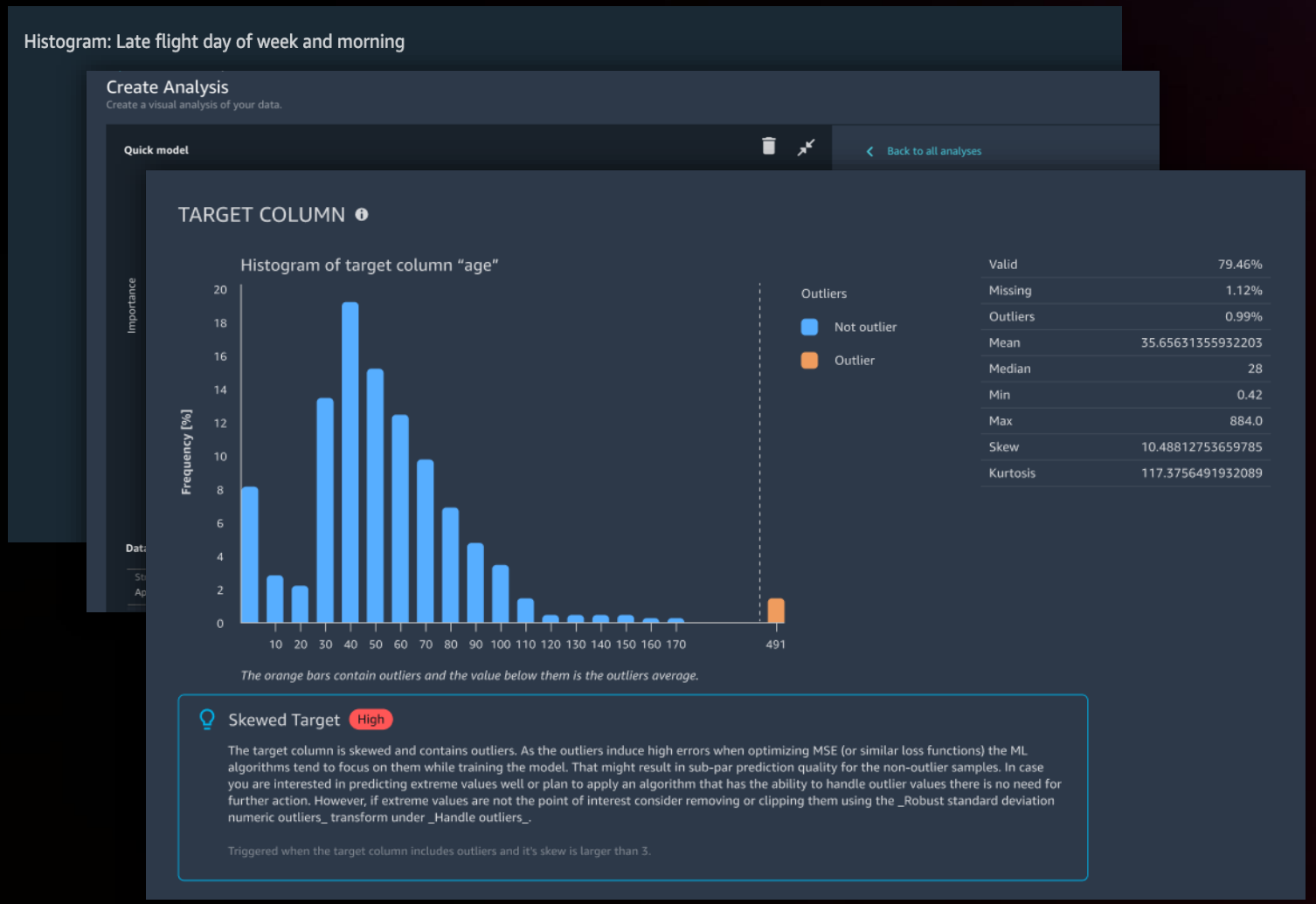



## Explore & import data using SQL




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

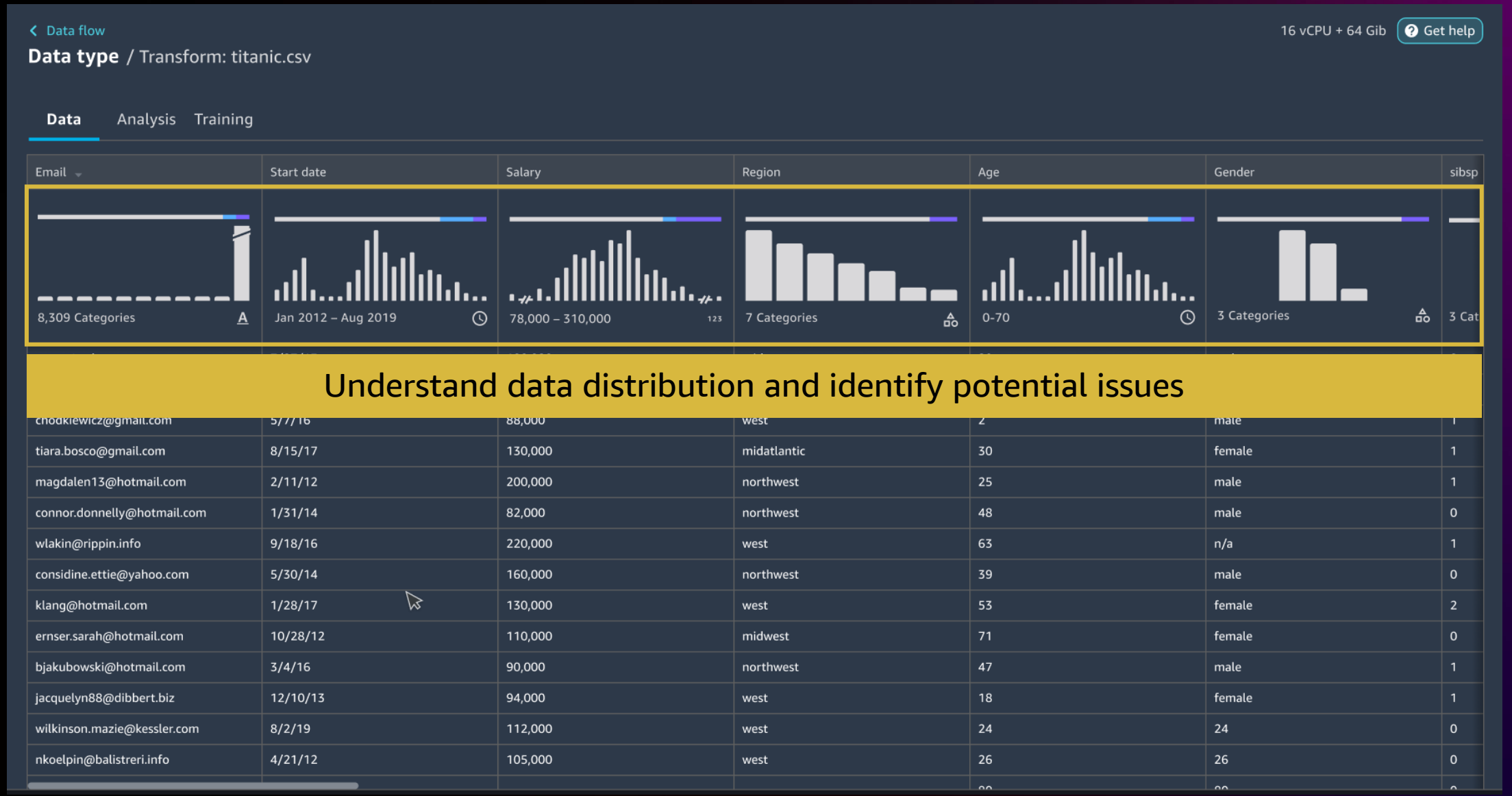
# Understand data & identify quality issues




- Popular visualizations & ML-powered analyses
- **Quick model** to predict model performance with prepared data
- **Data insight report**
  - Identifies quality issues & bias in features and target
  - Recommend transform to fix issue
  - Snapshot before and after

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

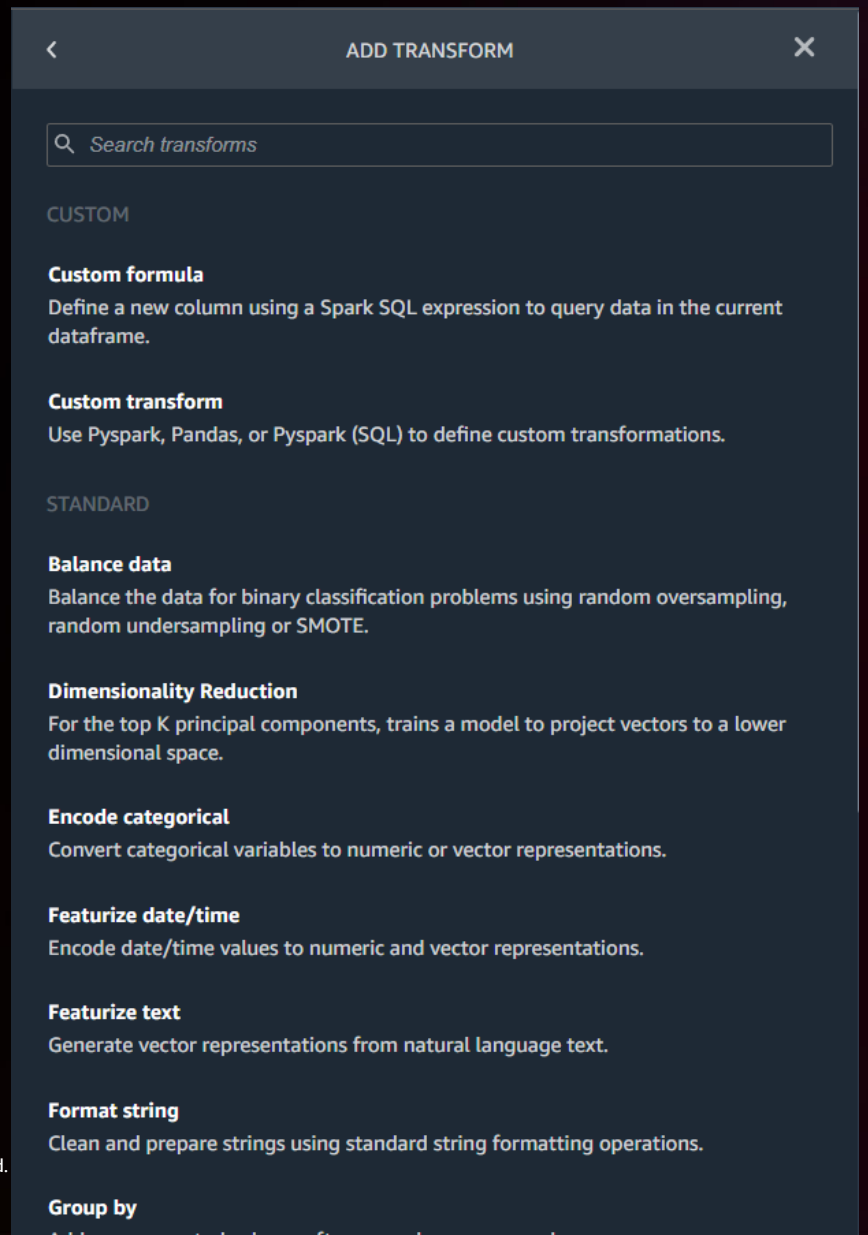
# New: Automatic visualization

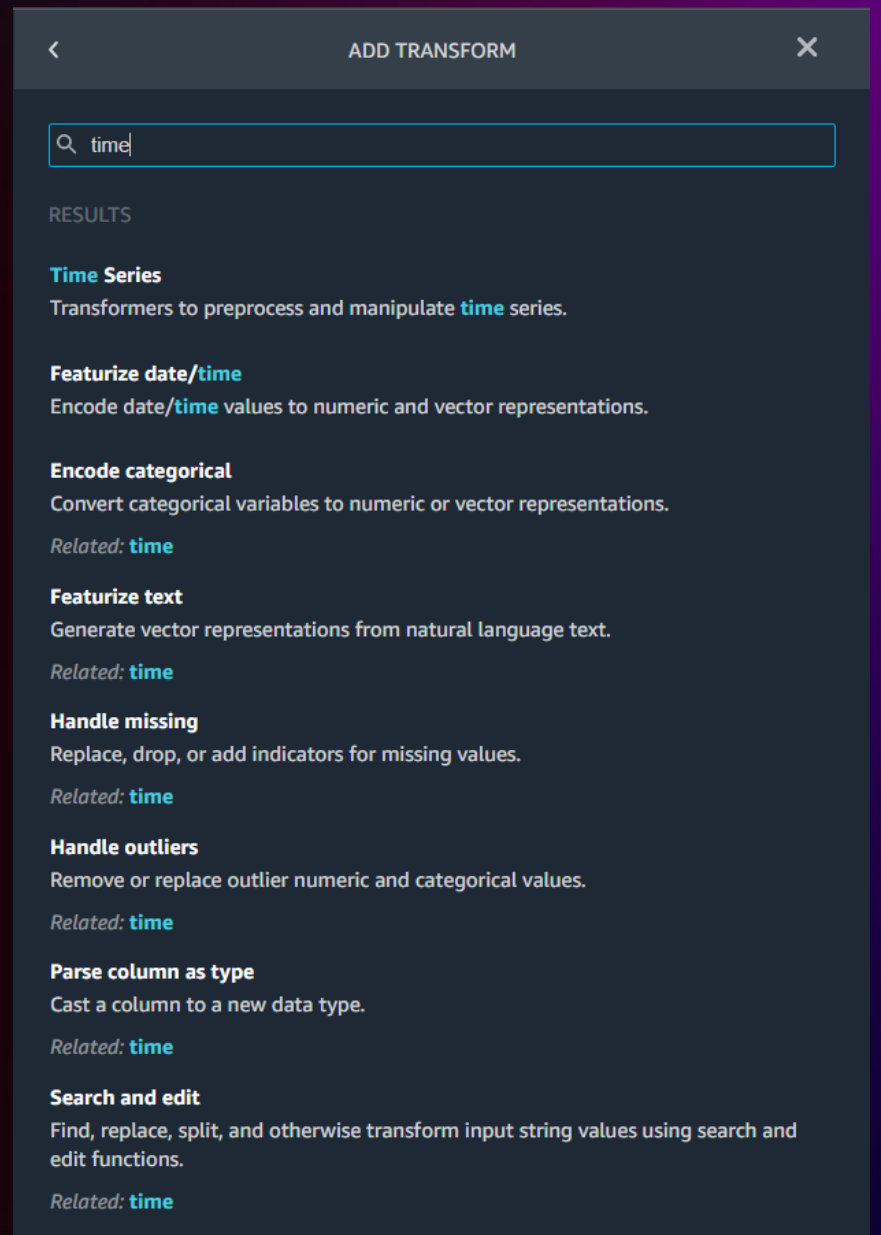



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Easily clean data and engineer features for ML

- With 300+ built-in transforms
- Added: Time Series, Train/Test split, Dimensionality Reduction, Multi-collinearity, PCA…
- Multi-column support
- Search to quickly find transform

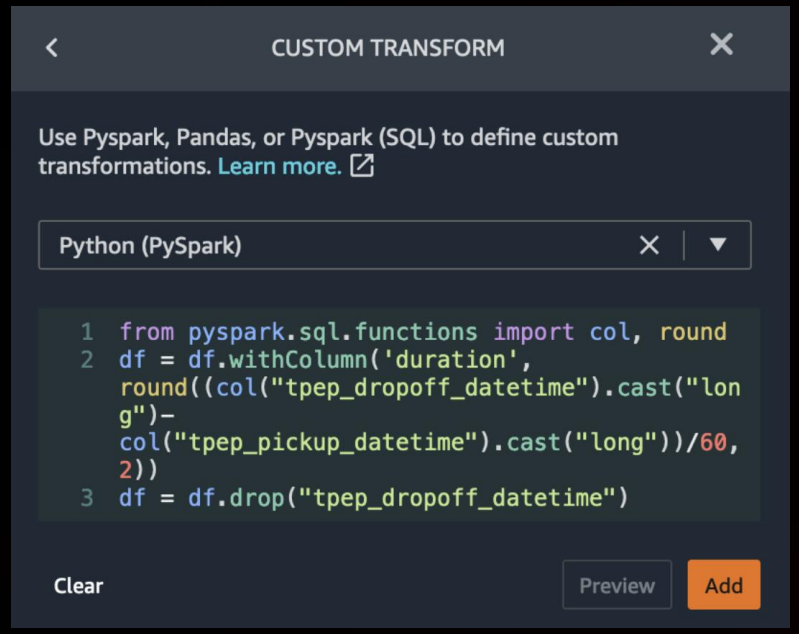


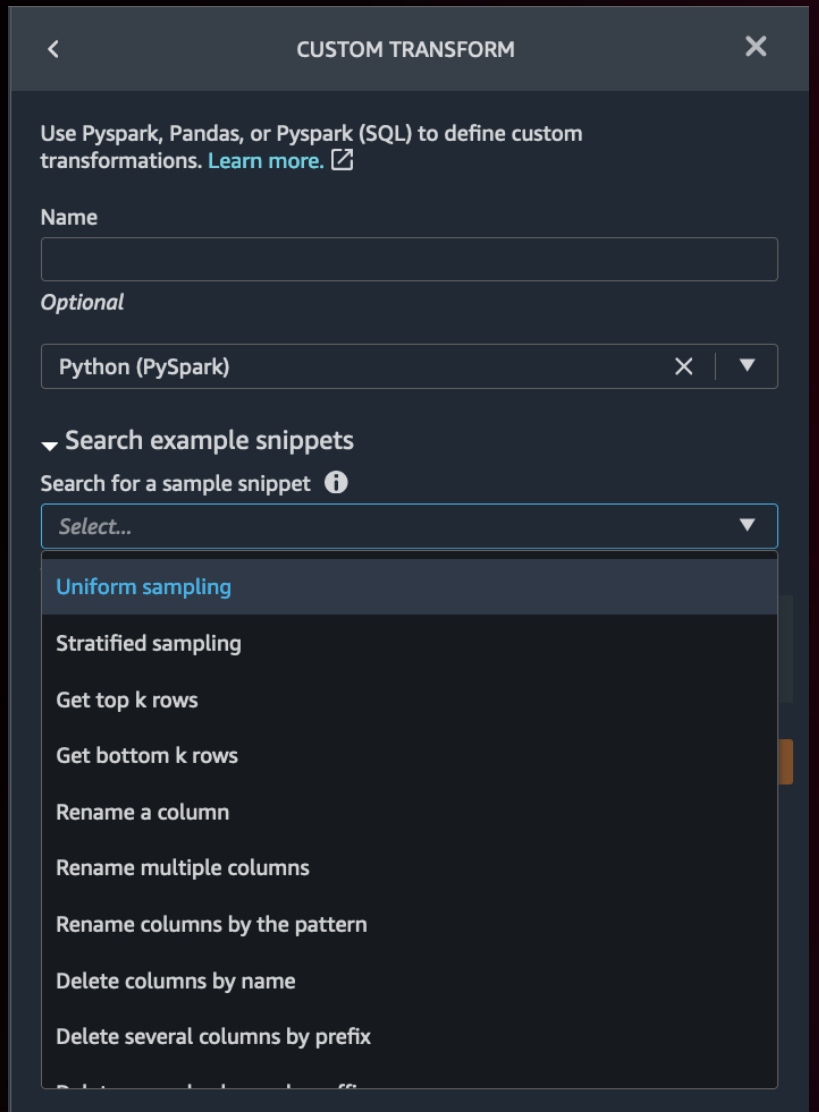

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

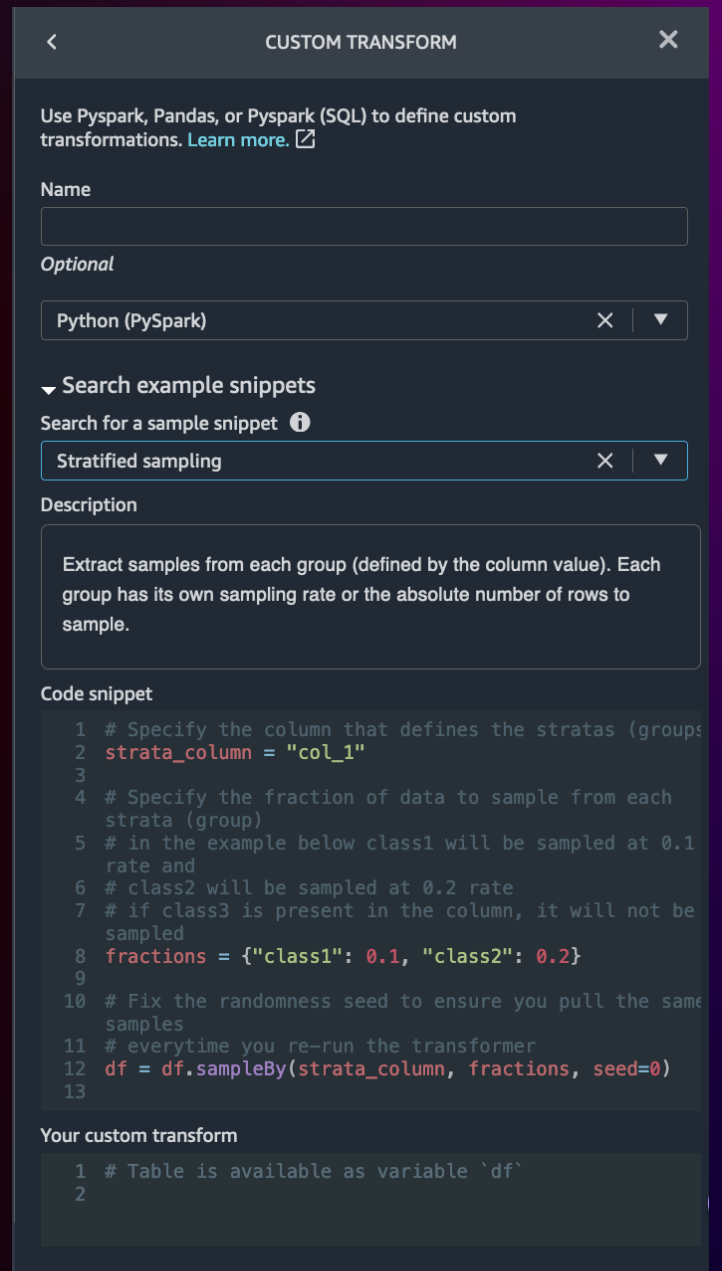
# Custom transforms & code snippets

## Panda, SQL, PySpark




## Common code snippet library

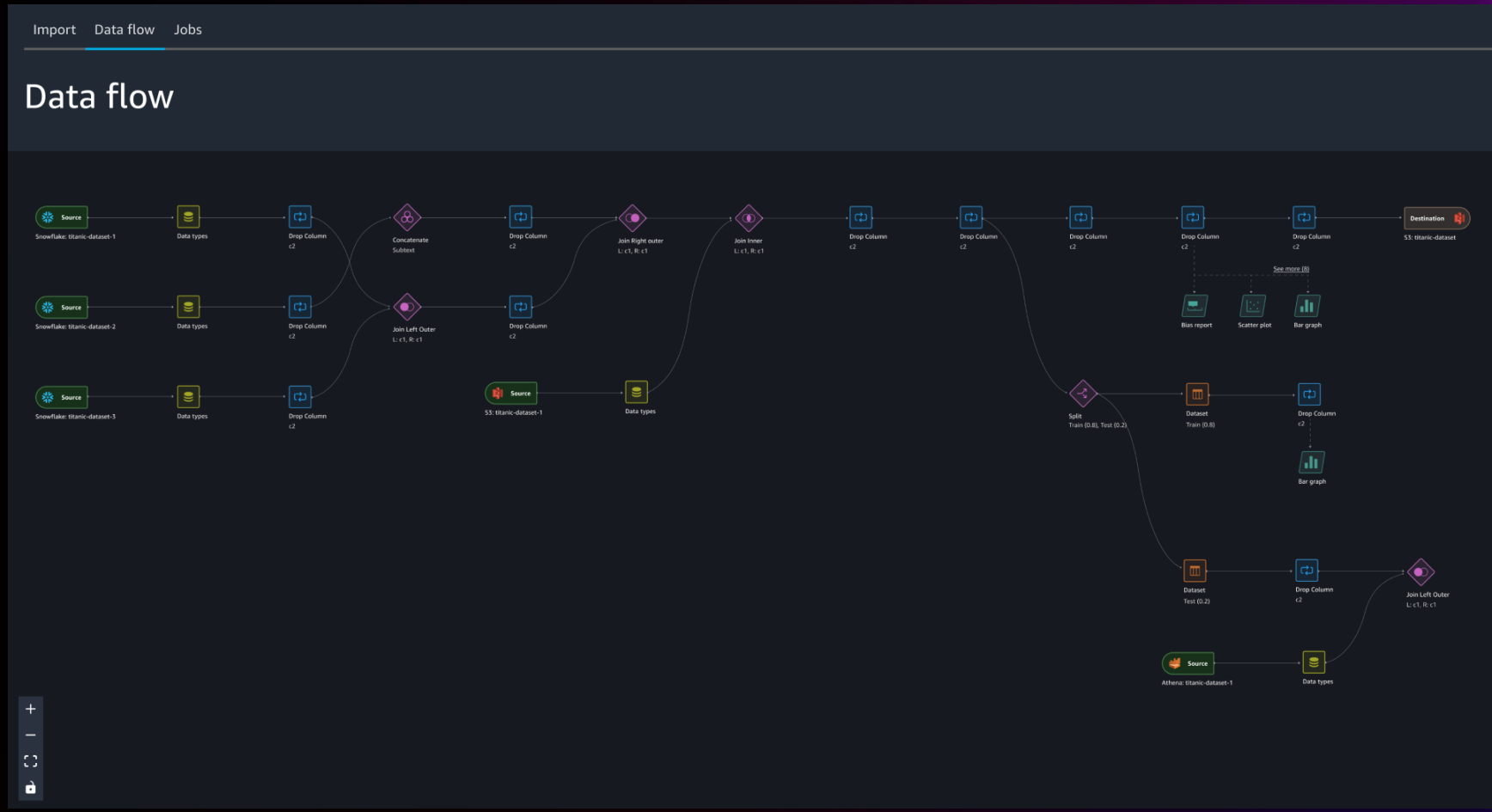



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

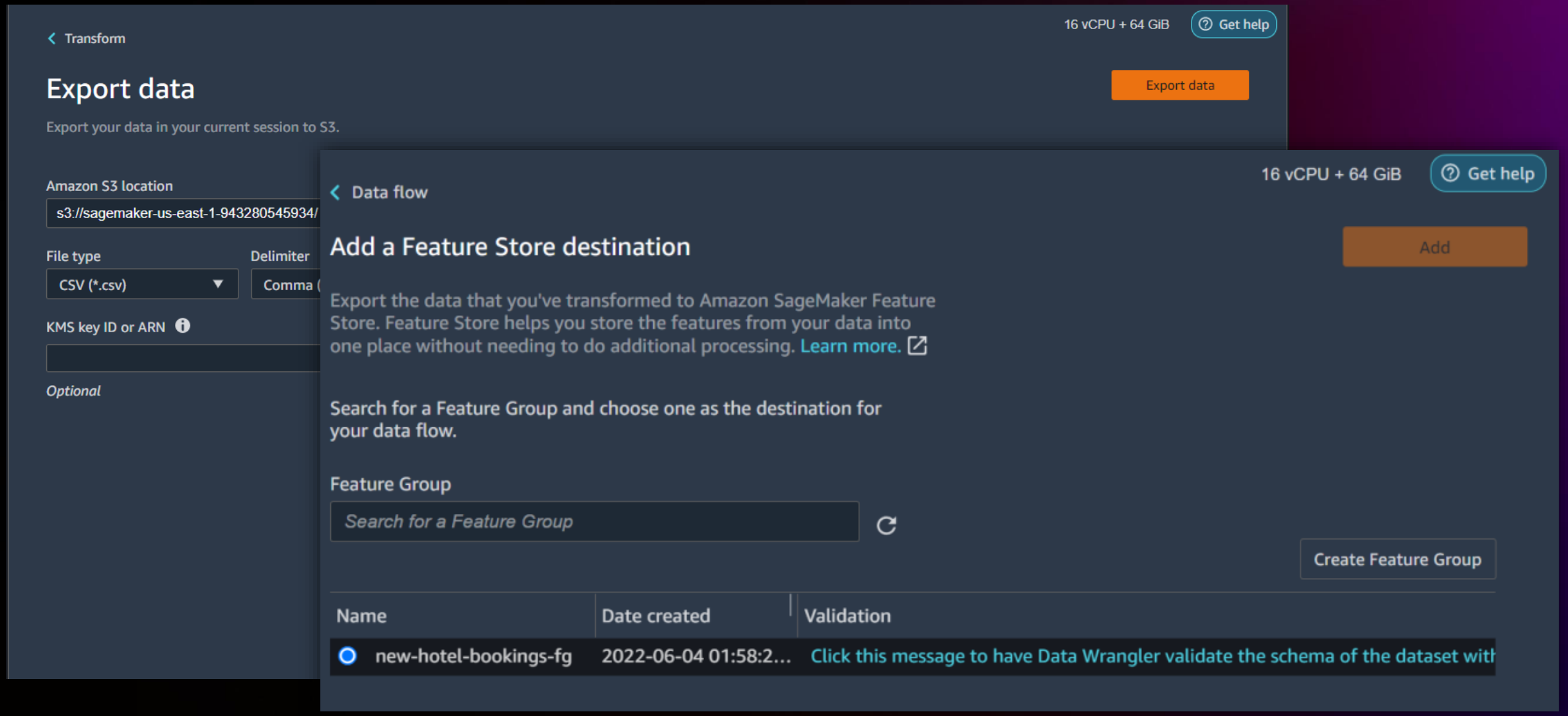
# Visualize data processing workflow steps

- Record all the steps in a data flow graph
- Easily navigate, view the order of transforms, modify, and delete steps iteratively
- Join and concatenate operators in the DAG



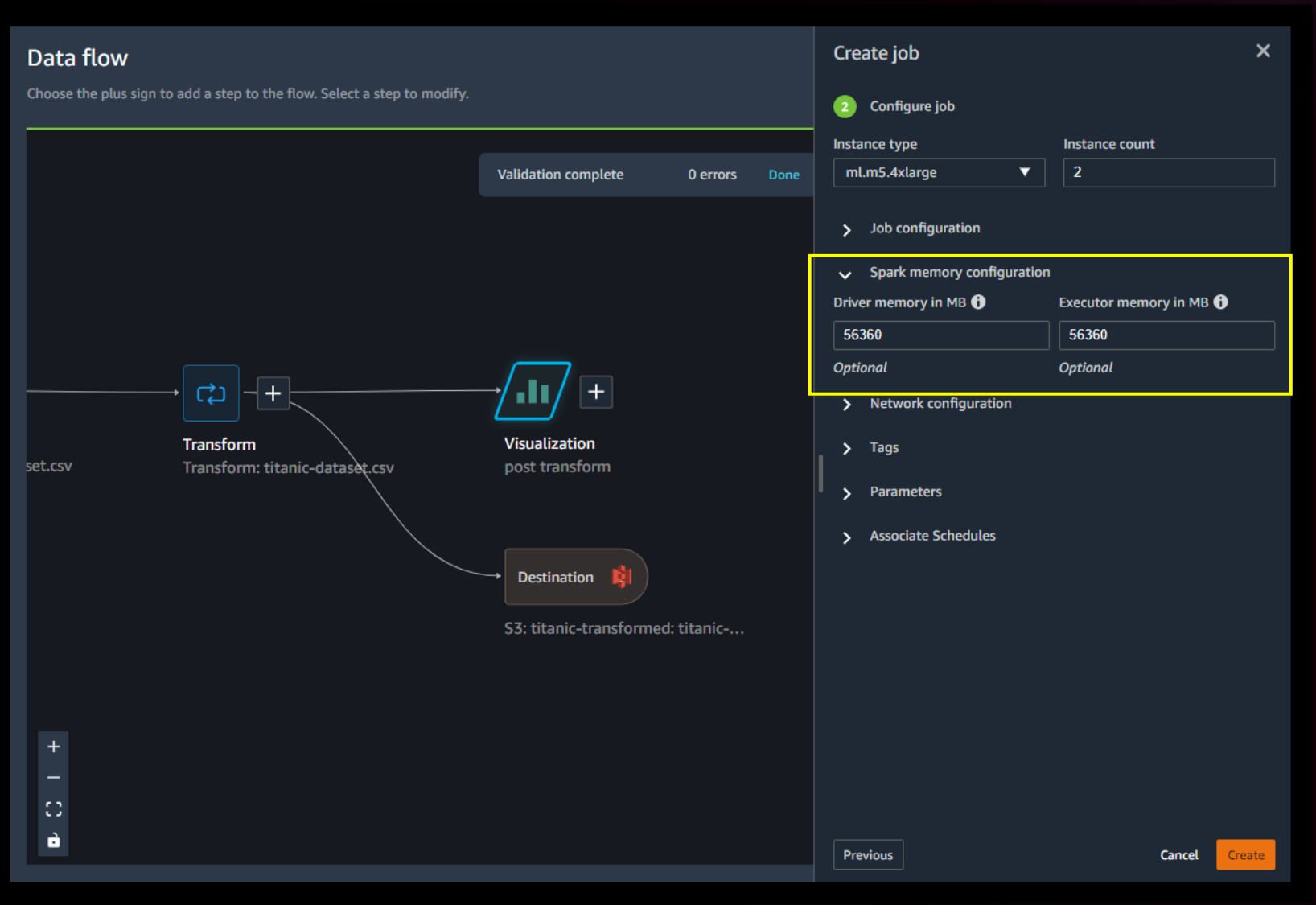

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Export to S3 or SageMaker Feature Store




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

# Scale to very large data set with processing job

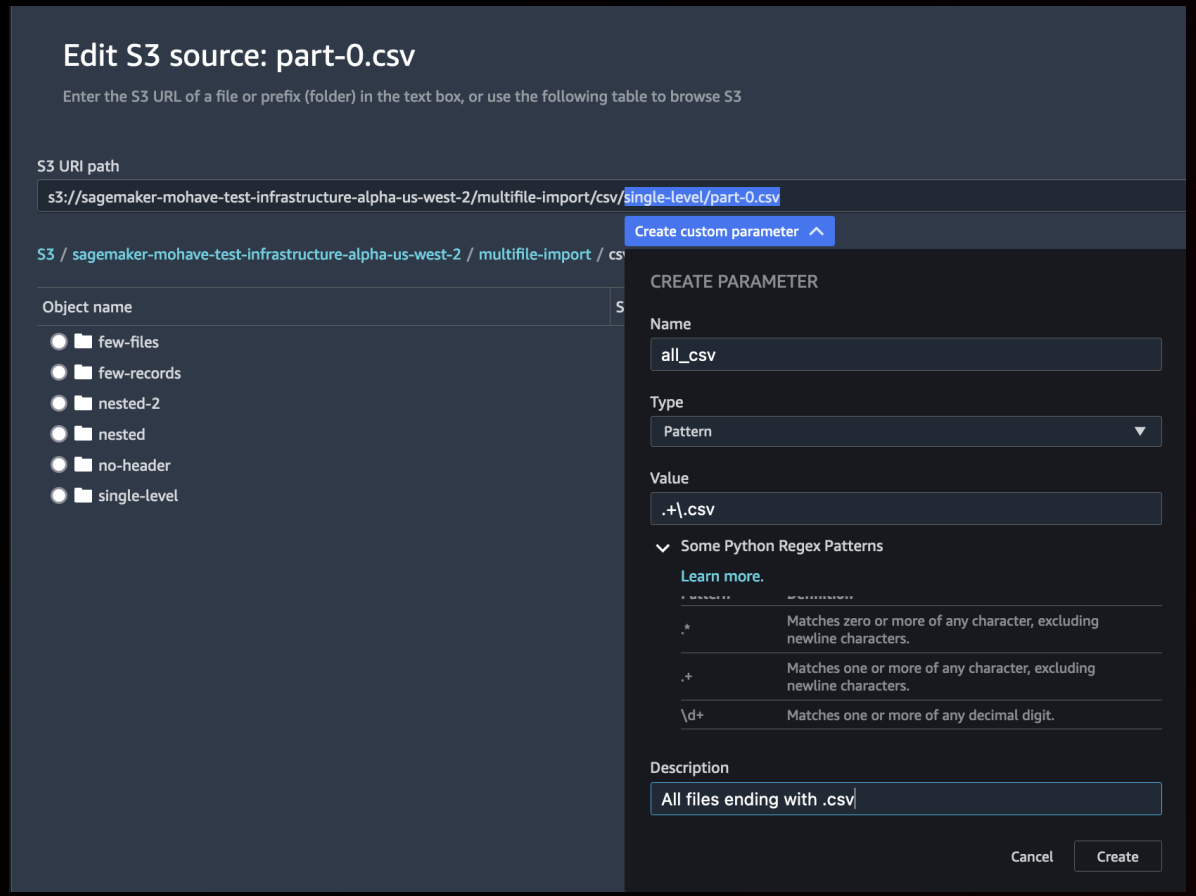



- Create a **serverless processing job** to scale to billions of rows of data via **UI & APIs**
- **Refit transform** for full dataset
- New Configure job partition settings and Spark memory

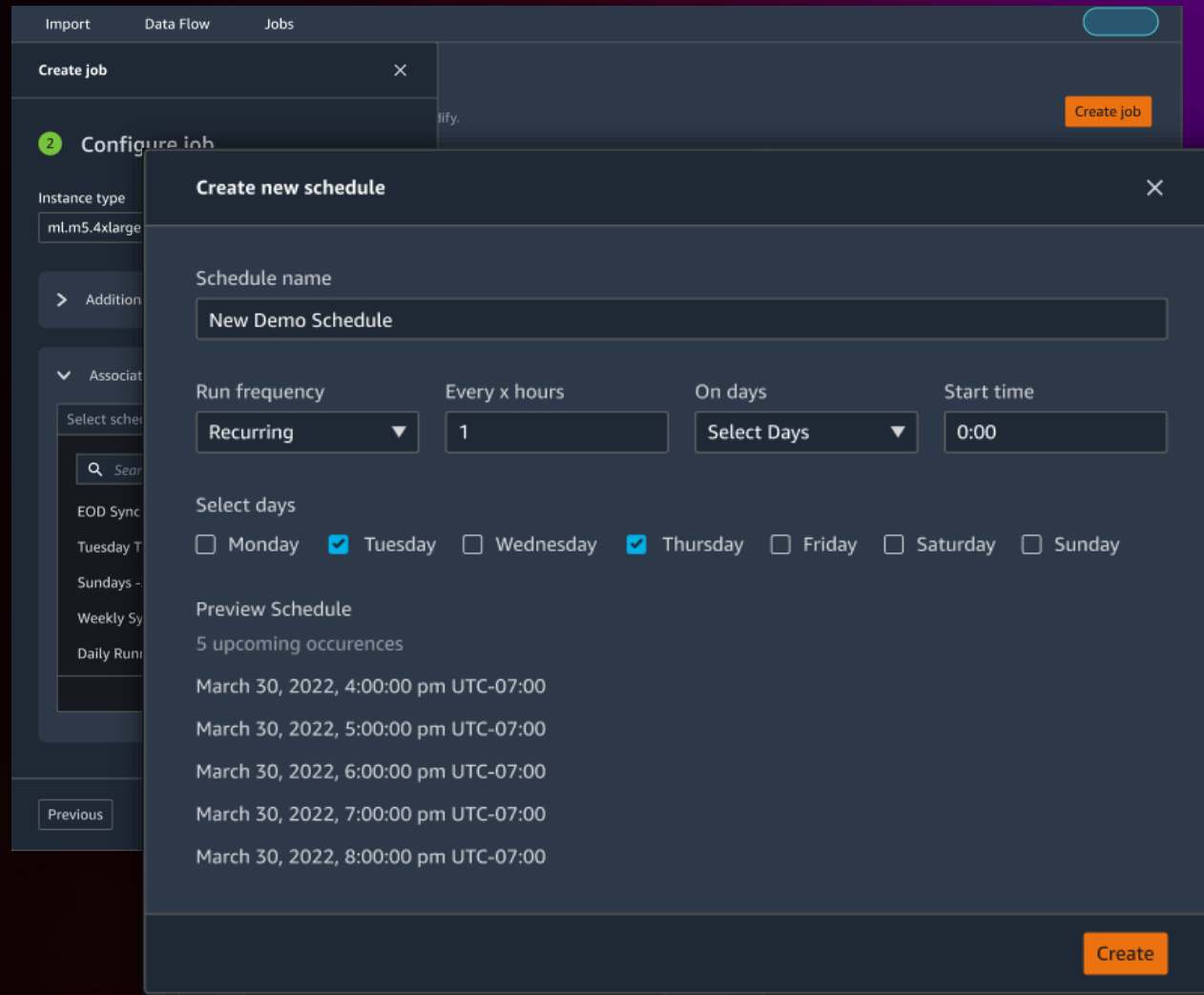
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Schedule job

## 1. Create custom parameter for S3 source path



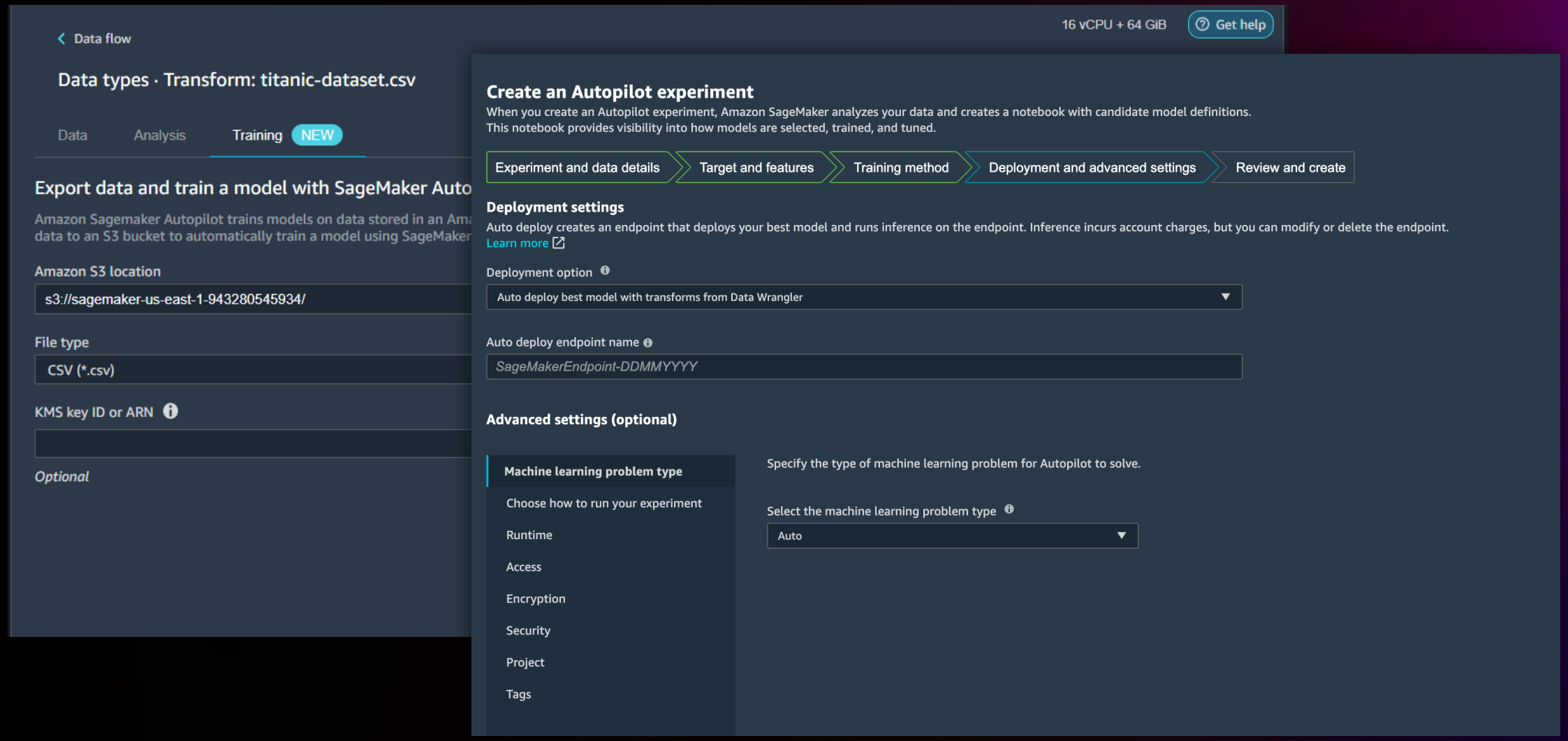

## 2. Schedule job to run on an hourly, daily, weekly, or custom schedule




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

# Train & deploy model using autoML

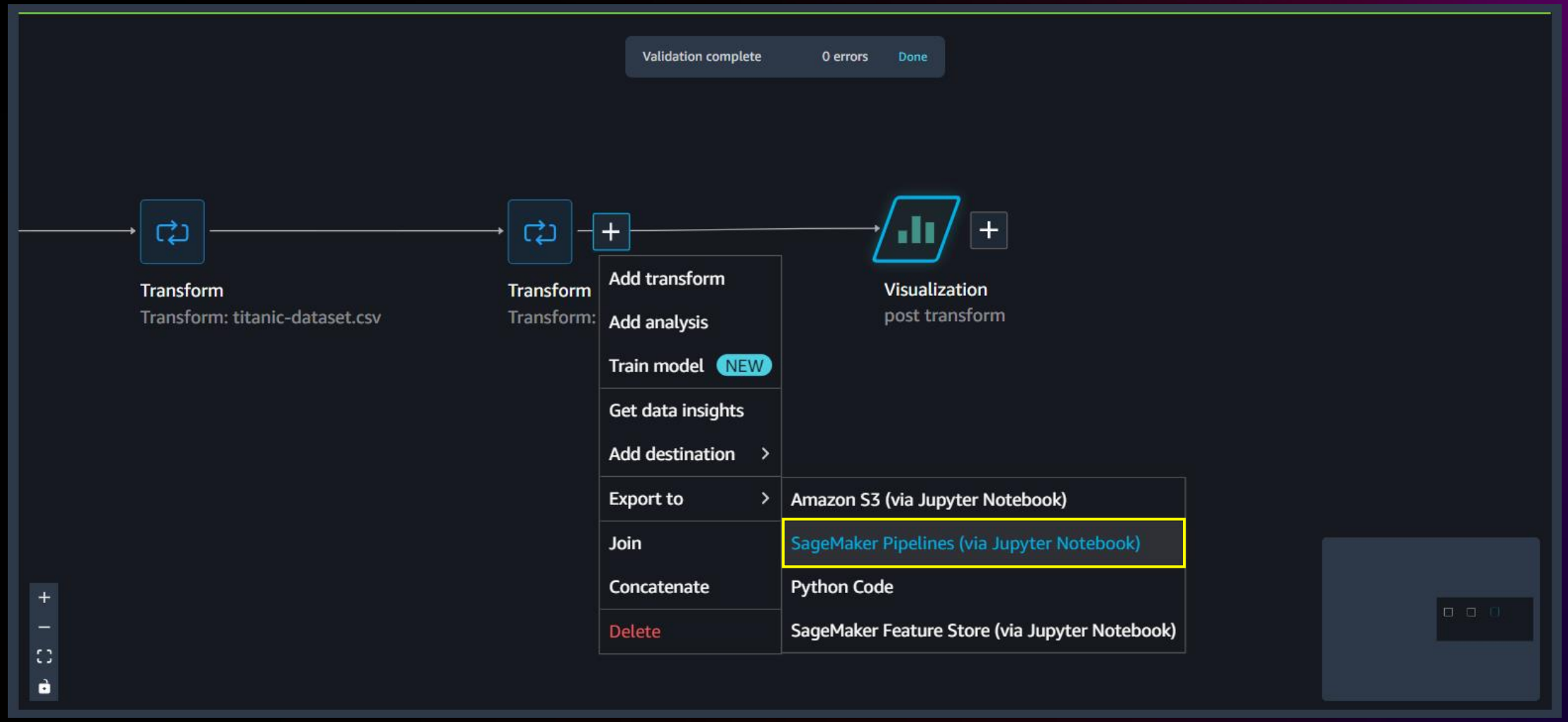
## WITH SAGEMAKER AUTOPILOT




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

# Automate data preparation in ML pipeline

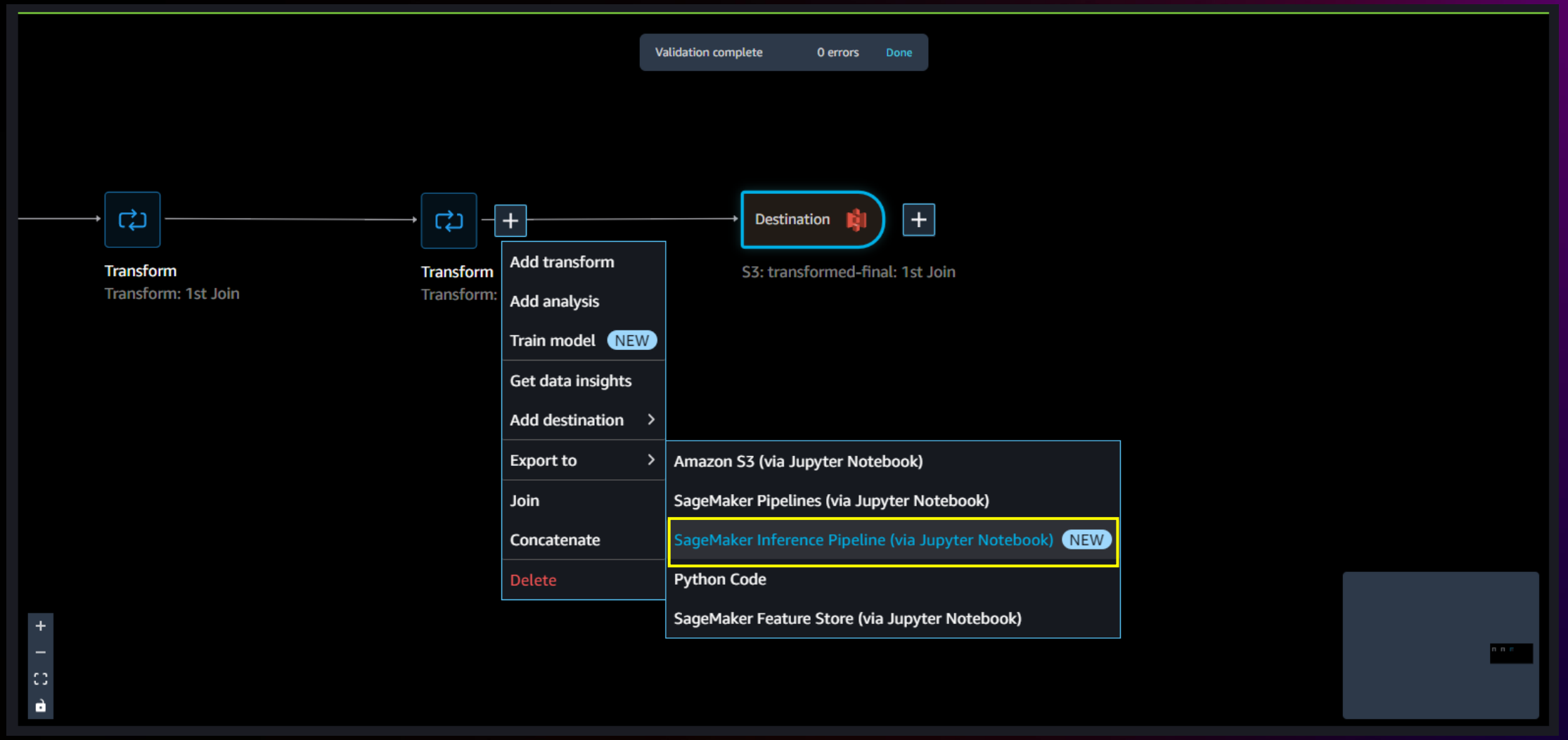
WITH SAGEMAKER PIPELINE




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

# New: Deploy data flow in Inference endpoint

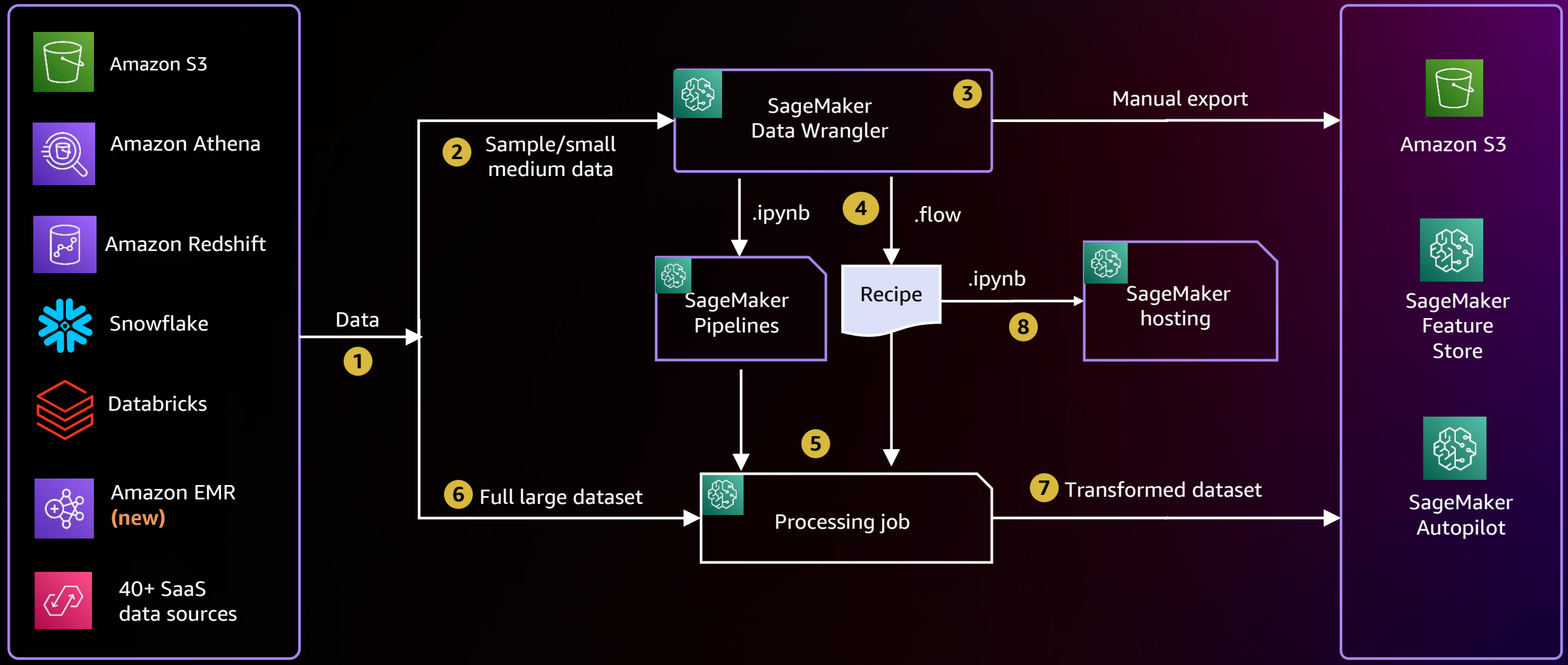
WITH SAGEMAKER INFERENCE PIPELINE




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

# Data preparation with SageMaker Data Wrangler

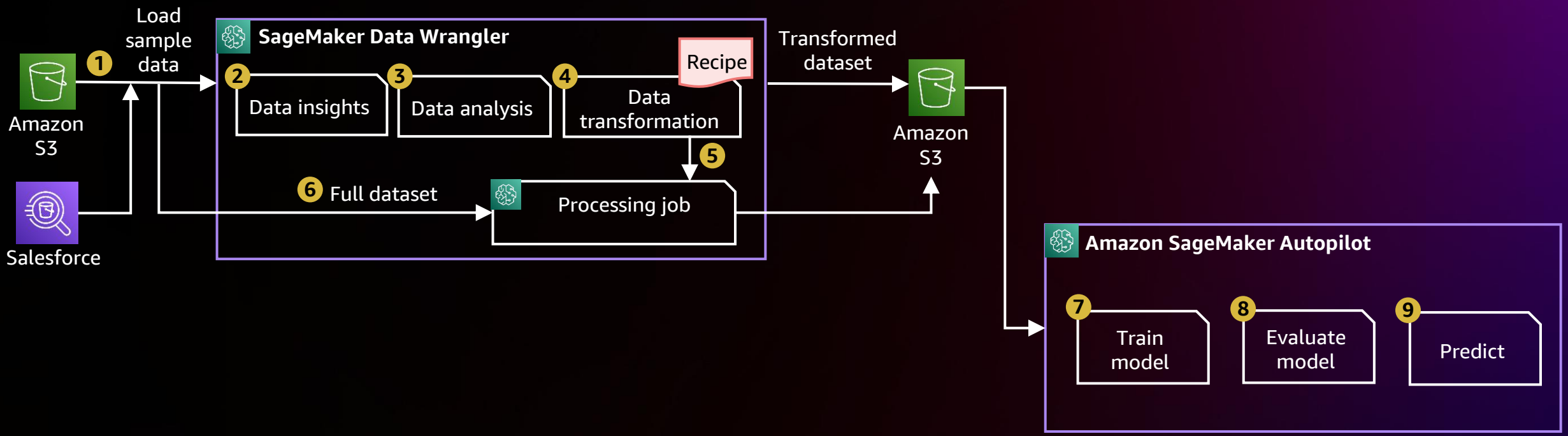
END-TO-END SOLUTION TO IMPORT, PREPARE, AND SCALE TO PRODUCTION




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

# Demo: Use case

## Loan default analysis and classification




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

# SageMaker Data Wrangler

USEFUL RESOURCES

- SageMaker Data Wrangler deep dive workshop – bit.ly/3tQweCI
- SageMaker low-code workshop – bit.ly/3SM8k5z
- Official developer guide – docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html
- Blogs – https://go.aws/3i92b6f
- Code examples – github.com/aws/amazon-sagemaker-examples/tree/main/sagemaker-datawrangler

aws

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

# SageMaker Data Wrangler workshop

bit.ly/3tQweCI



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

# Thank you!

Arunprasath Shankar
arunprsh@amazon.com

Huong Nguyen
huonguh@amazon.com



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

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