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

AIM326

Prepare data and model features for ML with ease, speed, and accuracy

Romi Datta, Ph.D.

SageMaker Product Management
AWS

Antonio Rodriguez

AI/ML Specialist Solutions Architect
AWS

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

Agenda

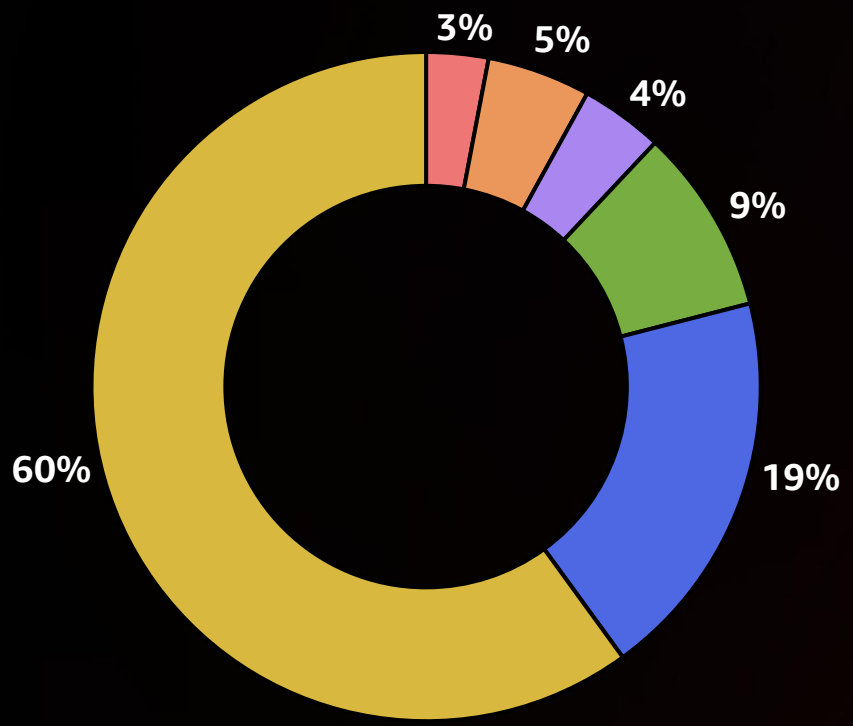
- Challenges in data preparation
- Amazon SageMaker ML flow and data prep tools
- SageMaker Data Wrangler
- SageMaker Feature Store
- Demo
- Discussion

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

Data preparation accounts for ~66% of data scientists time

According to Anaconda' State of Data Science survey, 2020

Common challenges with data preparation in machine learning



What data scientists spend the most time doing

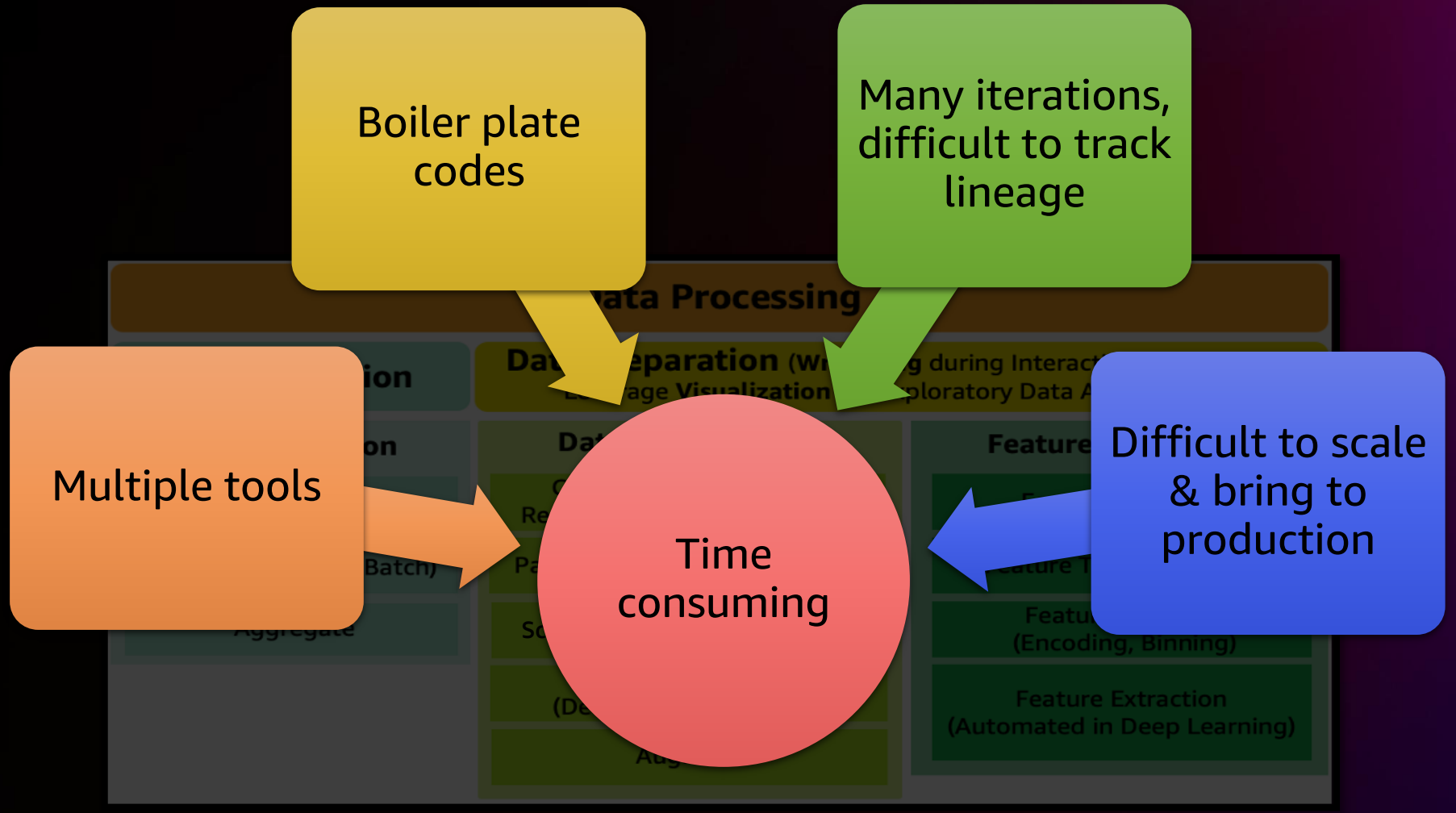
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting datasets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Source: Forbes survey of 80 data scientists, March 2016

aws

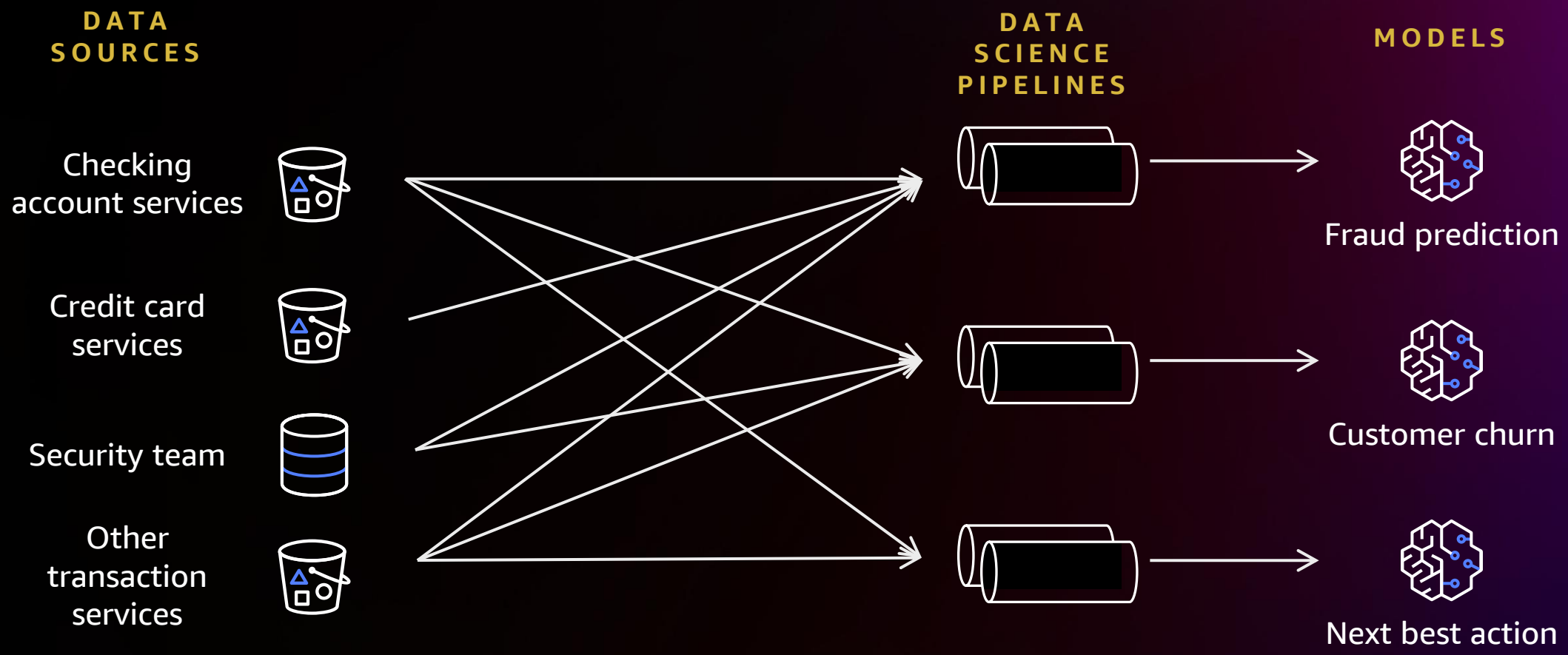
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data preparation: Common challenges



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

Gets worse when you have to repeat processes



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

Amazon SageMaker overview

Amazon SageMaker

PREPARE

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store
Store, update, retrieve, and share features

BUILD

SageMaker Studio notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and bring-your-own algorithms
Dozens of optimized algorithms or bring your own

Local mode
Test and prototype on your local machine

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart
Pre-built solutions for common use cases

TRAIN & TUNE

Managed training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic model tuning
Hyperparameter optimization

Distributed training libraries
Training for large datasets and models

SageMaker Debugger
Debug and profile training runs

Managed spot training
Reduce training cost by 90%

DEPLOY & MANAGE

Managed deployment
Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow integration
Simplify Kubernetes-based machine learning

Multi-model endpoints
Reduce cost by hosting multiple models per instance

SageMaker Model Monitor
Maintain accuracy of deployed models

SageMaker Edge Manager
Manage and monitor models on edge devices

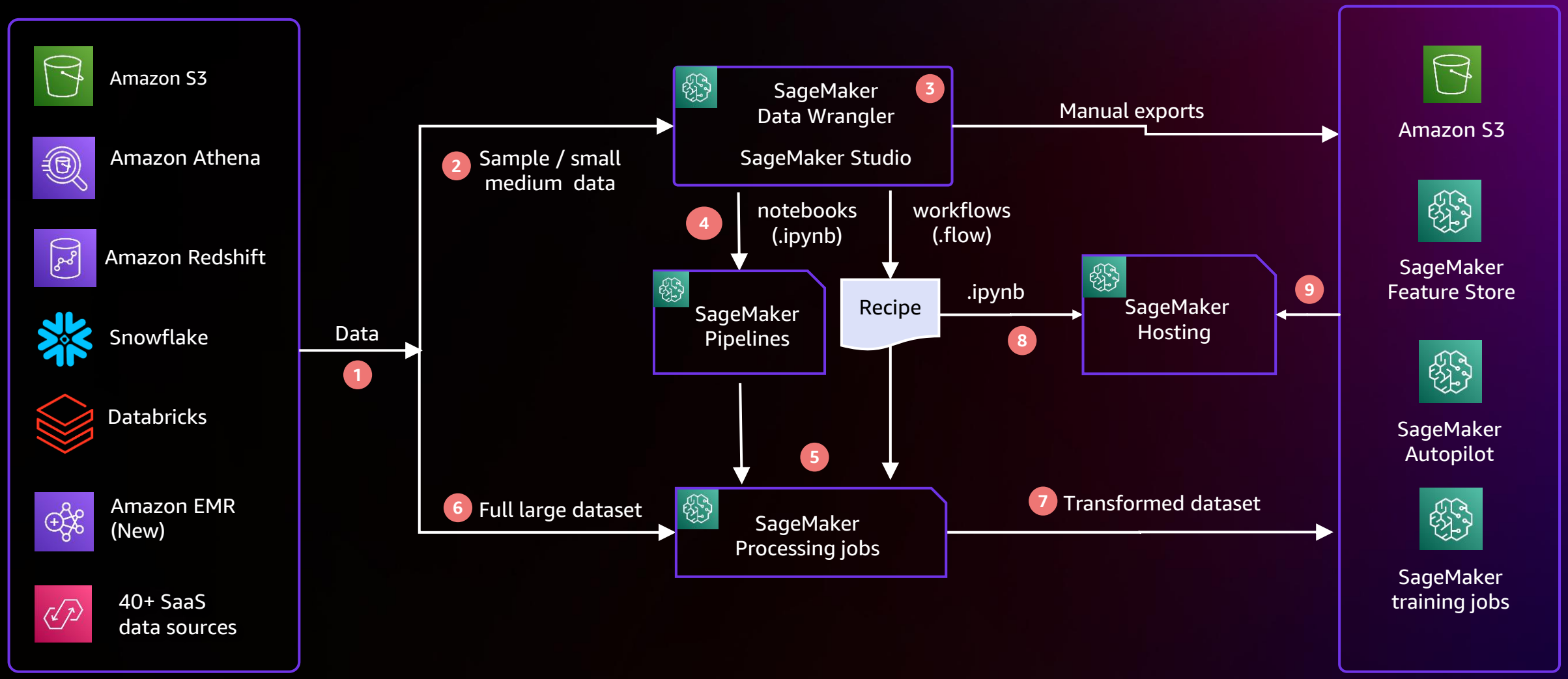
SageMaker Pipelines
Workflow orchestration and automation

SageMaker Studio
Integrated development environment (IDE) for ML

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

Data preparation with Amazon SageMaker

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



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

Amazon SageMaker Data Wrangler

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

Amazon SageMaker Data Wrangler

LOW-CODE/NO-CODE TOOL FOR INTERACTIVE AND PRODUCTION DATA PREP FOR ML



Connect to a data source & import data for interactive data prep

(4 native & 26+ via Athena)

Sample data or work on full dataset



Visualize, understand, cleanse, and enrich data

(300+ built-in Spark operators)

Build & visualize data transformation flows



Run data transformation flows in production on large data sets

(Spark processing job)

Export to S3, SageMaker Feature Store, or Pipelines for E2E ML

Interactive **Production**

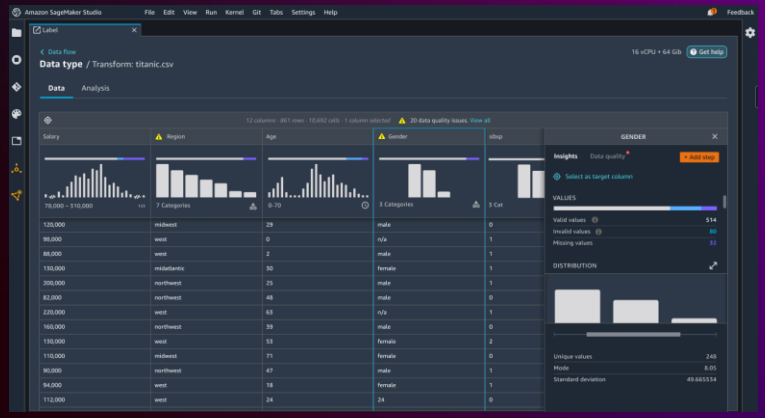
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

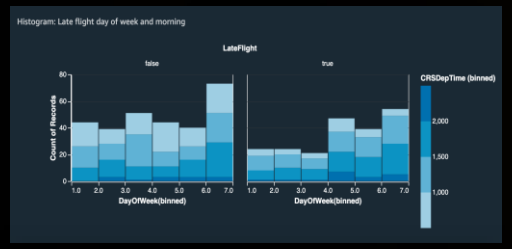
SageMaker Data Wrangler

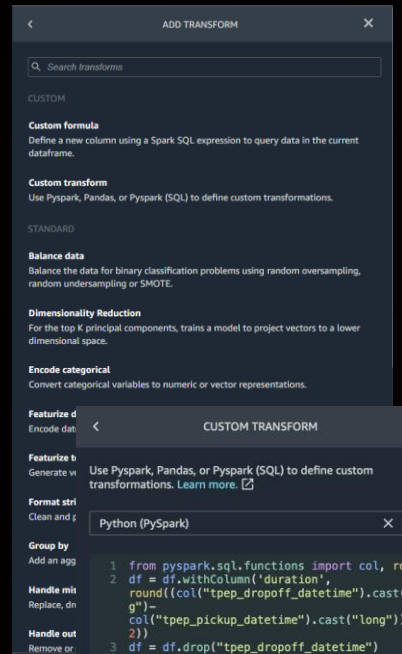
THE FASTEST AND EASIEST WAY TO PREPARE DATA FOR ML

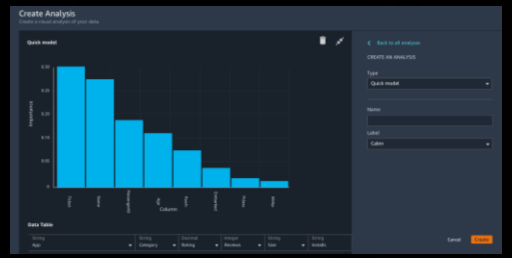
- ✓ Quickly connect and import data
- ✓ Understand data visually
- ✓ Easily prepare data for ML, 300+ built-in or custom transforms

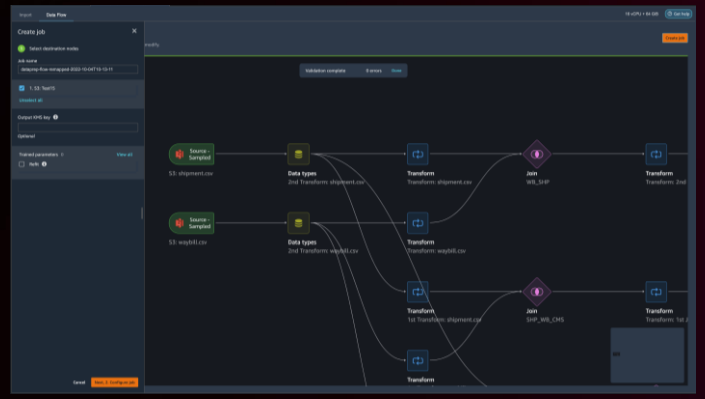


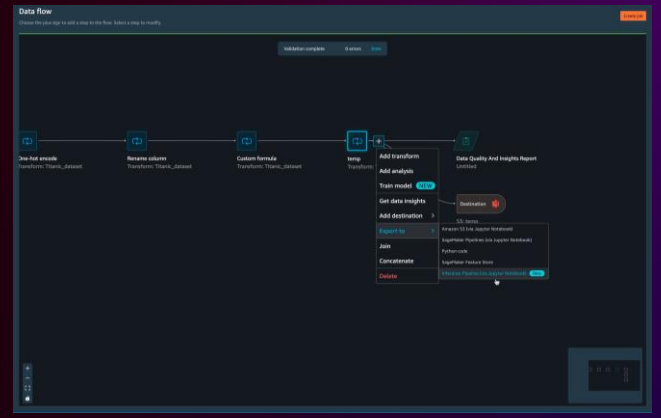












- ✓ Data insight and quality reports
- ✓ Data processing workflows
- ✓ Scale to very large datasets
- ✓ Deploy in production or schedule executions

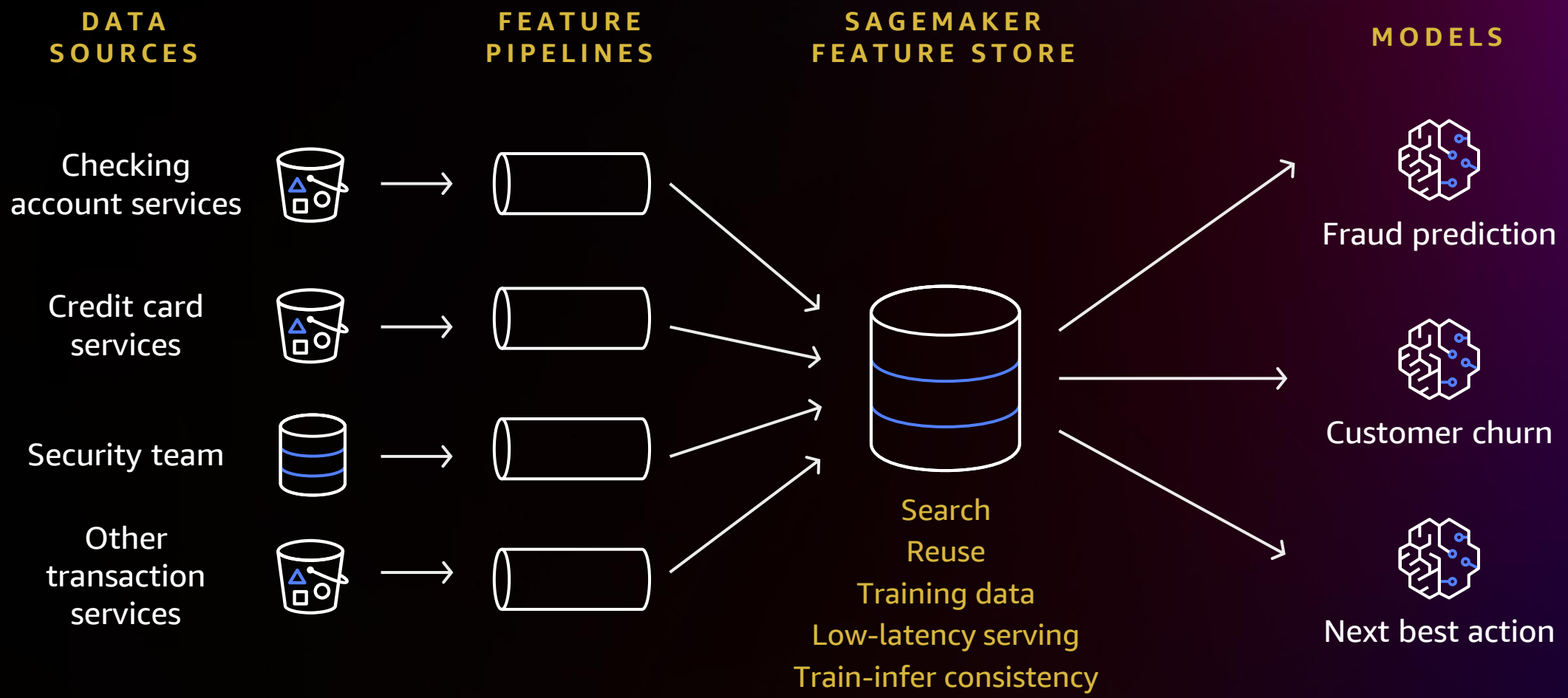
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker Feature Store

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

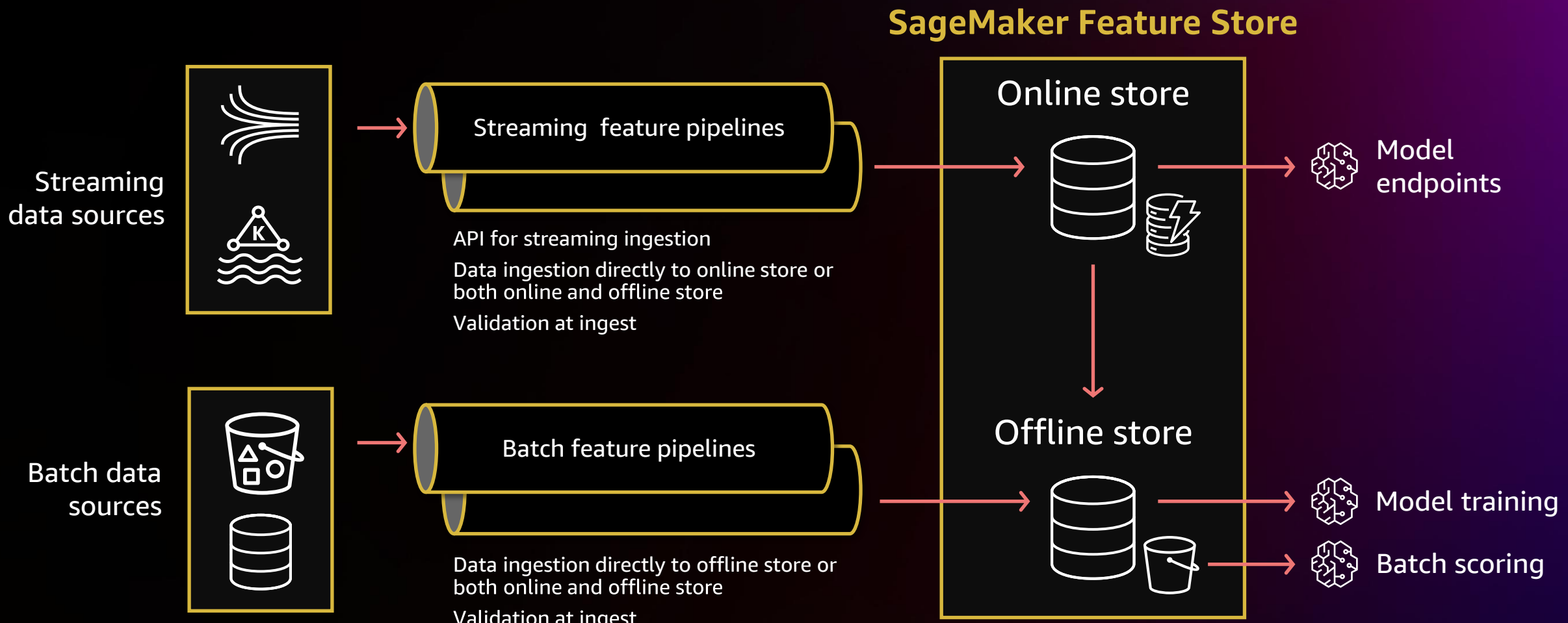
With SageMaker Feature Store ...

BUILD FEATURES ONCE, AND REUSE THEM ACROSS TEAMS AND MODELS



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

SageMaker Feature Store



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

Feature consistency with online and offline store

```
# 3 uniquely identified records

put_record('id-1', t1, 0.1)

put_record('id-2', t2, 0.2)

put_record('id-3', t3, 0.3)


# new records for existing id

put_record('id-1', t4, 0.4)

put_record('id-1', t5, 0.5)
```

Feature group after all PutRecord calls

Online store

id	event time	feature
id-1	t5	0.5
id-2	t2	0.2
id-3	t3	0.3



Offline store

id	event time	feature
id-1	t1	0.1
id-2	t2	0.2
id-3	t3	0.3
id-1	t4	0.4
id-1	t5	0.5

aws

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

Feature ingestion APIs

PutRecord API

```
record = [{'FeatureName': 'feature_1', 'ValueAsString': 'val_1'},
                    …
          {'FeatureName': 'feature_N', 'ValueAsString': 'val_N'}]

sm_fs.put_record(FeatureGroupName='my-fg-name', Record=record)
```

fg.ingest()

```
fg = FeatureGroup(name='my-fg-name',
                  sagemaker_session=fs_session)

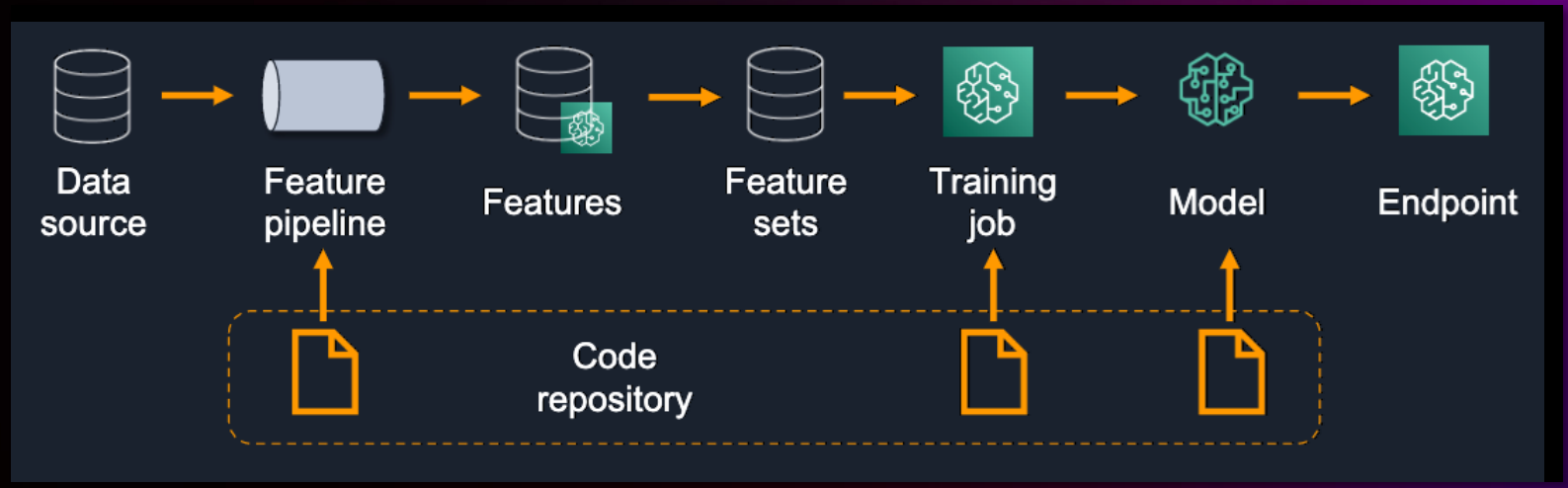
fg.ingest(df, max_processes=20)
```

aws

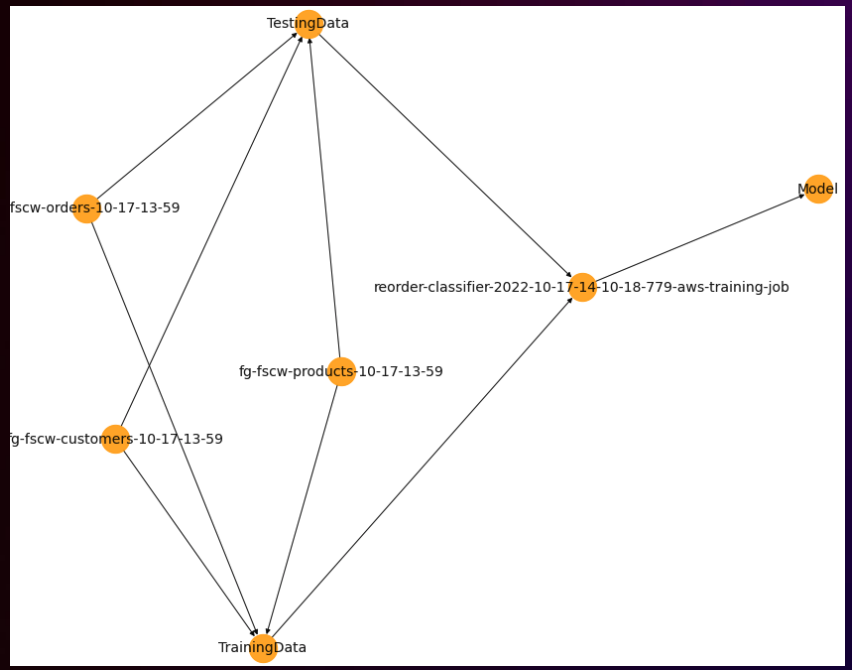
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Features lineage tracking

- Extending ML lineage to SageMaker Feature Store
- Keep track of feature's lifecycle in the E2E of the ML projects
- Query lineage for reusing existing features, increasing efficiency

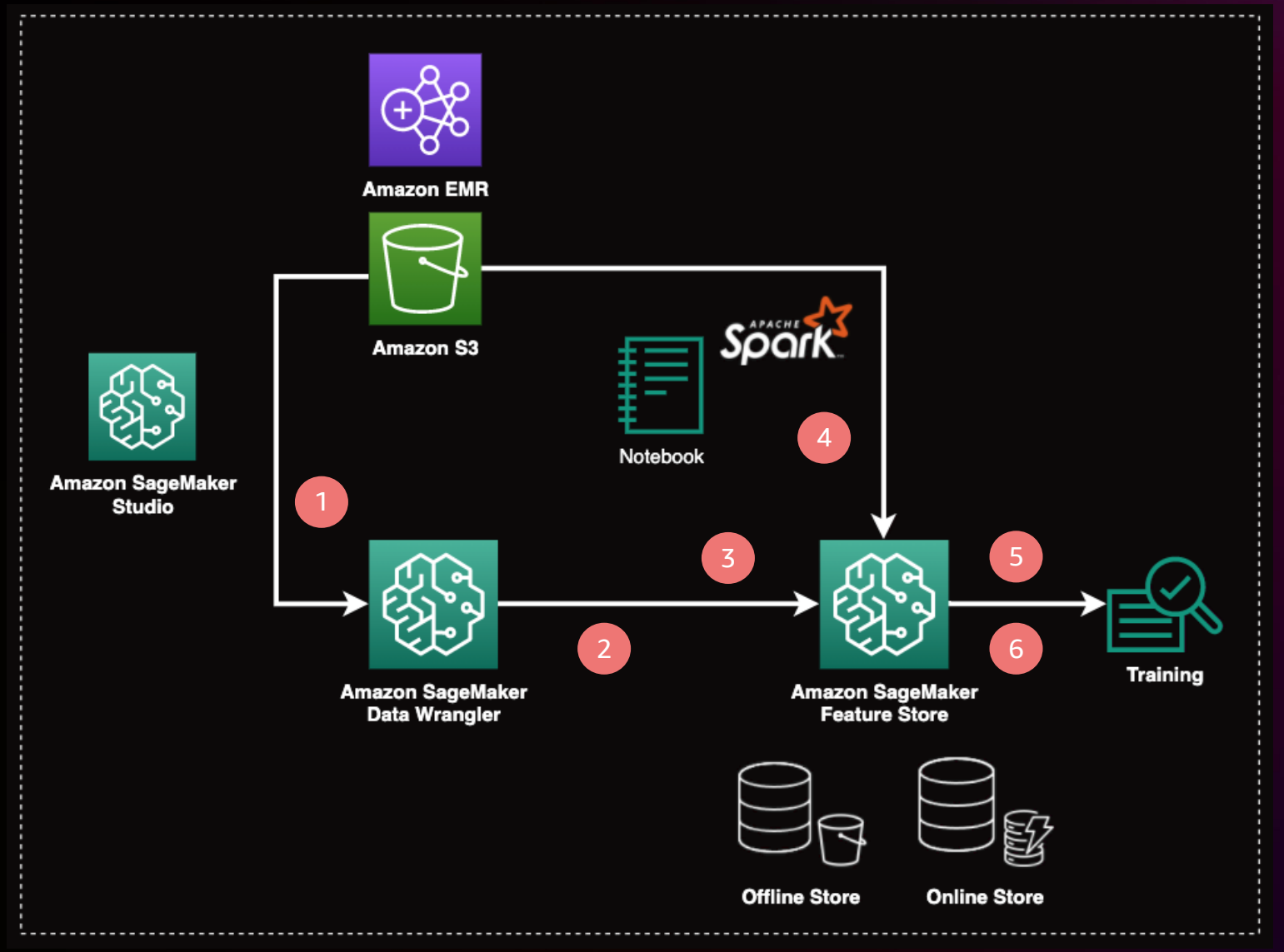


	Name/Source	Association	Name/Destination	Artifact Source ARN	Artifact Destination ARN	Source URI
0	reorder-classifier-2022-10-17-14-10-18-779-aws...	Produced	Model	arn:aws:sagemaker:us-west-2:889960878219:exper...	arn:aws:sagemaker:us-west-2:889960878219:artif...	None
1	TestingData	ContributedTo	reorder-classifier-2022-10-17-14-10-18-779-aws...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:exper...	s3://sagemaker-us-west-2-889960878219/sagemake...
2	TrainingData	ContributedTo	reorder-classifier-2022-10-17-14-10-18-779-aws...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:exper...	s3://sagemaker-us-west-2-889960878219/sagemake...
3	fg-fscw-orders-10-17-13-59	ContributedTo	TestingData	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:featu...
4	fg-fscw-products-10-17-13-59	ContributedTo	TestingData	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:featu...
5	fg-fscw-customers-10-17-13-59	ContributedTo	TestingData	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:featu...
6	fg-fscw-orders-10-17-13-59	ContributedTo	TrainingData	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:featu...
7	fg-fscw-products-10-17-13-59	ContributedTo	TrainingData	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:featu...
8	fg-fscw-customers-10-17-13-59	ContributedTo	TrainingData	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:artif...	arn:aws:sagemaker:us-west-2:889960878219:featu...

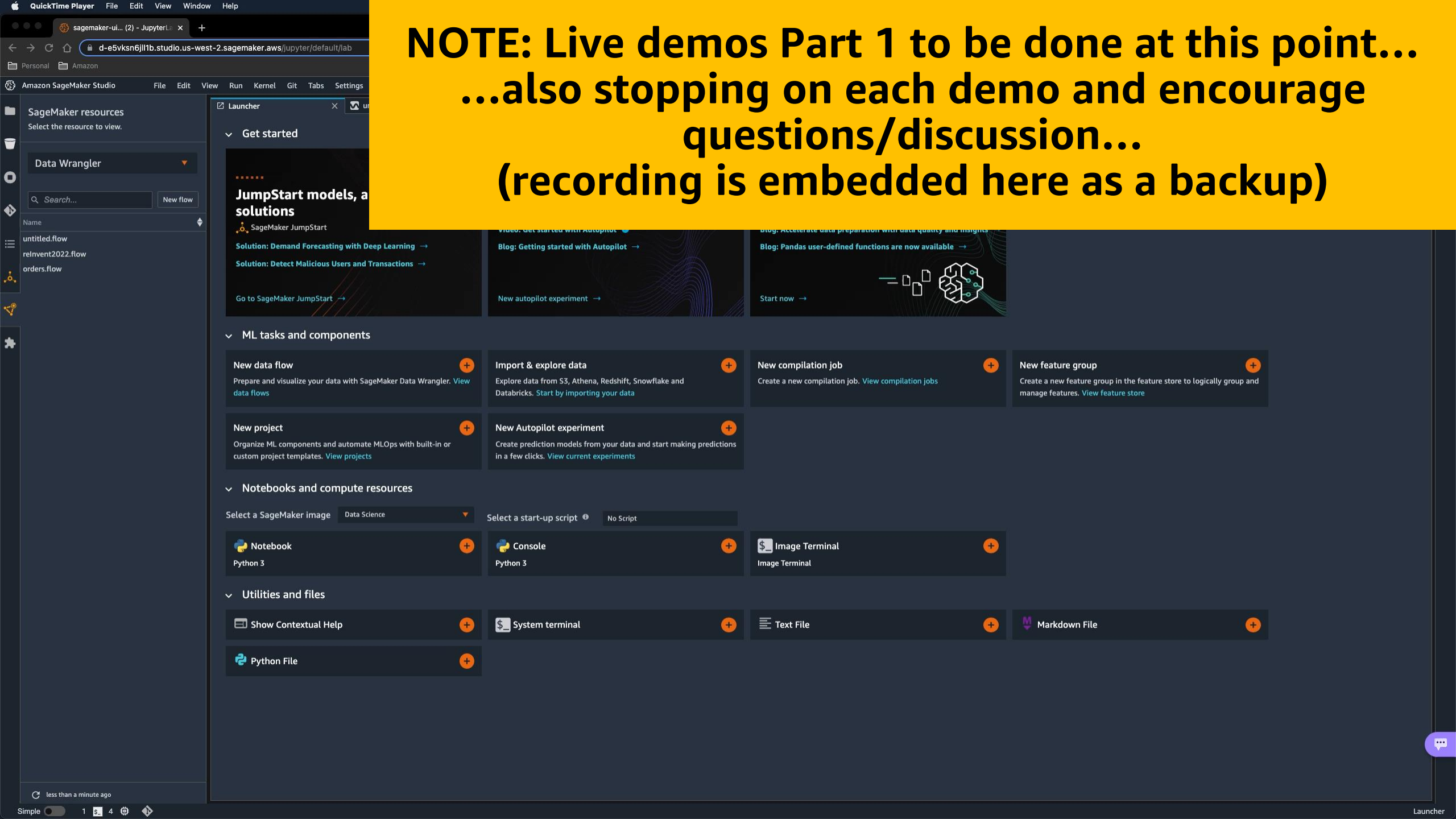


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Preparing data for ML – demos



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


NOTE: Live demos Part 1 to be done at this point...
...also stopping on each demo and encourage questions/discussion...
(recording is embedded here as a backup)
QuickTime Player File Edit View Window Help
sagemaker-ui... (2) - JupyterLa
d-e5vksn6jll1b.studio.us-west-2.sagemaker.aws/jupyter/default/lab
Personal Amazon
Amazon SageMaker Studio File Edit View Run Kernel Git Tabs Settings
SageMaker resources
Select the resource to view.
Data Wrangler
Search...
New flow
Name
untitled.flow
reInvent2022.flow
orders.flow
less than a minute ago
Launcher
Get started
JumpStart models, a
solutions
SageMaker JumpStart
Solution: Demand Forecasting with Deep Learning
Solution: Detect Malicious Users and Transactions
Go to SageMaker JumpStart
Blog: Getting started with Autopilot
New autopilot experiment
Blog: Pandas user-defined functions are now available
Start now
ML tasks and components
New data flow
Prepare and visualize your data with SageMaker Data Wrangler. View data flows
Import & explore data
Explore data from S3, Athena, Redshift, Snowflake and Databricks. Start by importing your data
New compilation job
Create a new compilation job. View compilation jobs
New feature group
Create a new feature group in the feature store to logically group and manage features. View feature store
New project
Organize ML components and automate MLOps with built-in or custom project templates. View projects
New Autopilot experiment
Create prediction models from your data and start making predictions in a few clicks. View current experiments
Notebooks and compute resources
Select a SageMaker image
Data Science
Select a start-up script
No Script
Notebook
Python 3
Console
Python 3
Image Terminal
Image Terminal
Utilities and files
Show Contextual Help
System terminal
Text File
Markdown File
Python File
Simple
Launcher

NOTE: Live demos part 2 to be done at this point...
...also stopping on each demo and encourage questions/discussion...
(recording is embedded here as a backup)

Contents

1. Overview
2. Feature Group Monitoring Preparation
3. Run Feature Group Monitoring using DataBrew Profile Job
4. Visualisation of Feature Group statistics
5. Clean up

Overview

In previous module (`Module-1 Introduction to SageMaker Feature Store`), we demonstrated how to create multiple features groups inside a Amazon SageMaker Feature Store and ingest data into it.

In this notebook, we will illustrate how to implement feature monitoring using AWS Glue DataBrew to create feature statistics.

AWS Glue DataBrew is a visual data preparation tool that helps you clean and normalize data without writing code. DataBrew also allows customers to specify which data quality statistics to auto-generate for datasets when running a profile job. This allows users to customize data profile statistics such as determining duplicate values, correlations, and outliers based on the nature and size of their datasets, and create a custom data profile overview with only the statistics that meet their needs.

We will cover the following aspects:

- Preparation step to Feature Monitoring
- Executing Feature Monitoring using a DataBrew Profile job to generate statistics for a feature group
- Extract statistics from profile job execution and persist to S3
- Additional visualisation of the feature group statistics



Thank you!

Romi Datta
romidat@amazon.com

Antonio Rodriguez
rodzanto@amazon.com



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

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