AWS re:Invent

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

AIM319

Prepare data for ML with ease, speed, and accuracy

Romi Datta, Ph.D.
SageMaker Product Management
Amazon Web Services

Kiran Kandala (he/him)
VP Engineering, Machine Learning & Advanced Analytics
Fidelity Investments

© 2021, 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

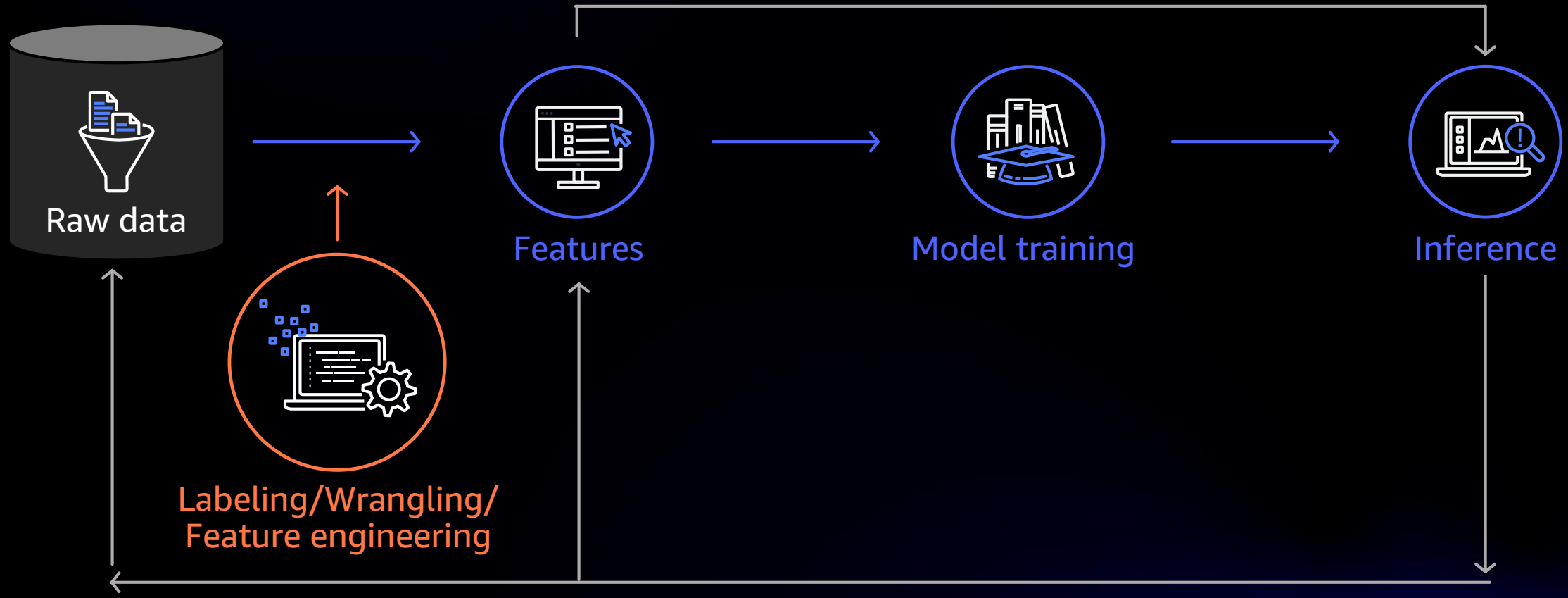
SageMaker Feature Store

Customer use case with Feature Store

SageMaker Ground Truth

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

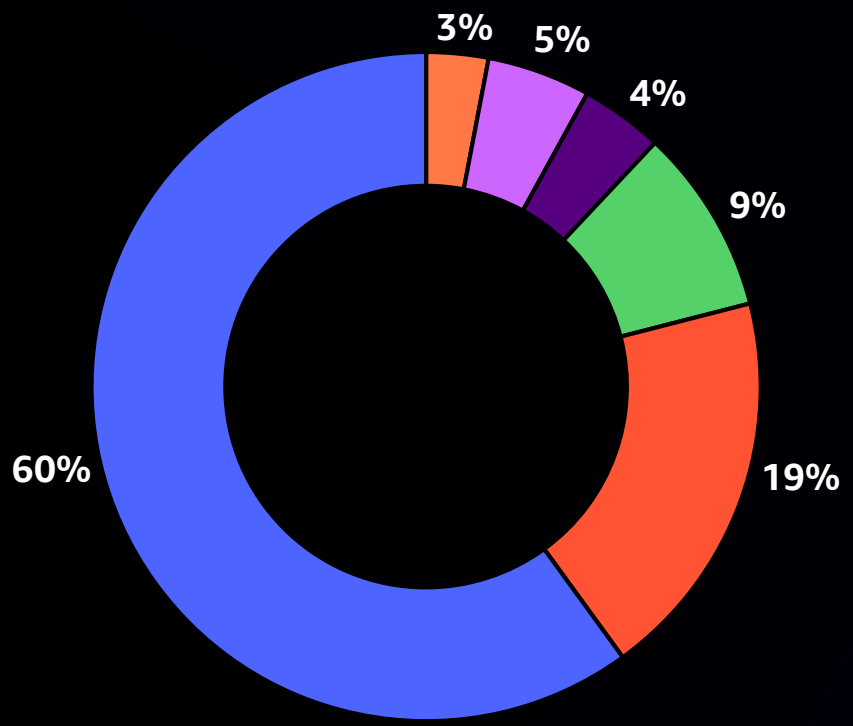
Machine learning workflow



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

Common challenges with data preparation in machine learning

Data preparation accounts for 80% of the work of data scientists



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

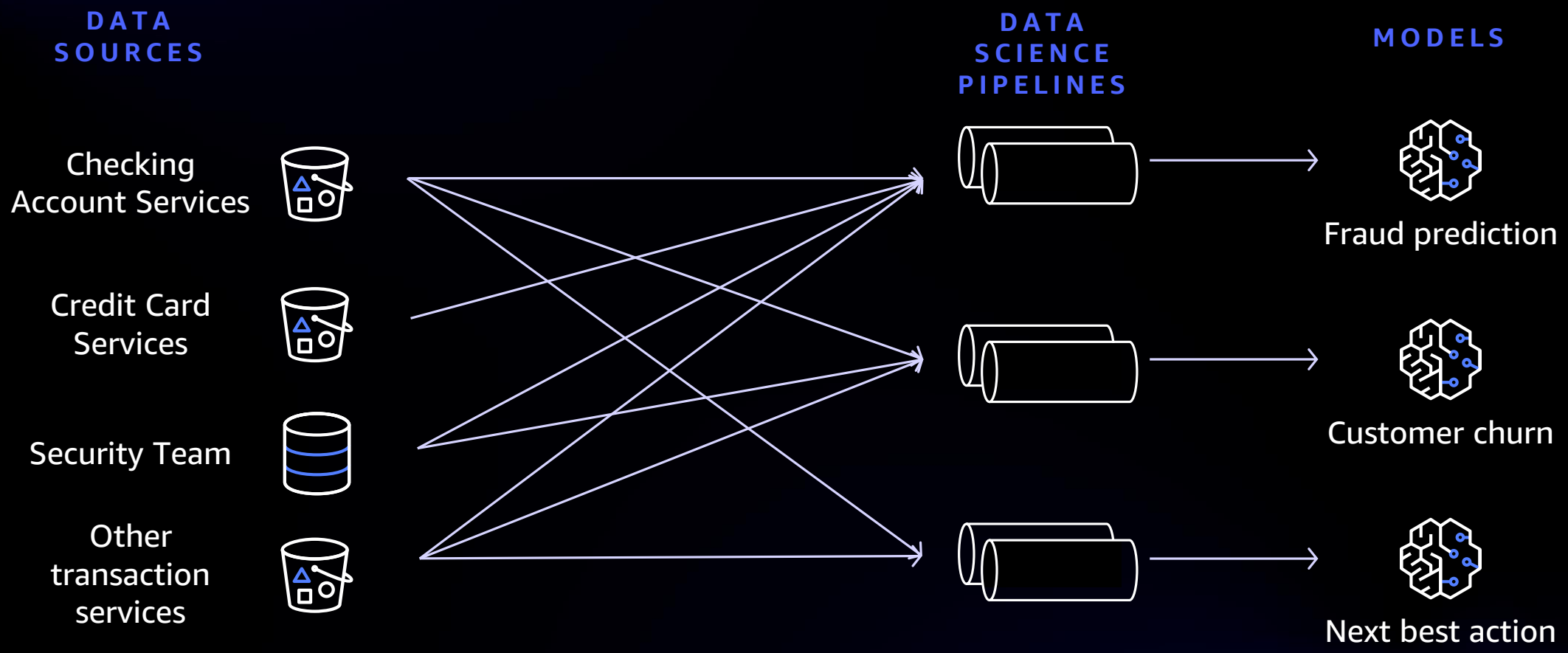
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Common challenges with data preparation in machine learning

1. Time-consuming & resource intensive
2. Requires multiple tools
3. Data preparation tools are typically not DevOps friendly
4. Common data preparation tasks require writing code

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

Gets worse when you have to repeat processes



© 2021, 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

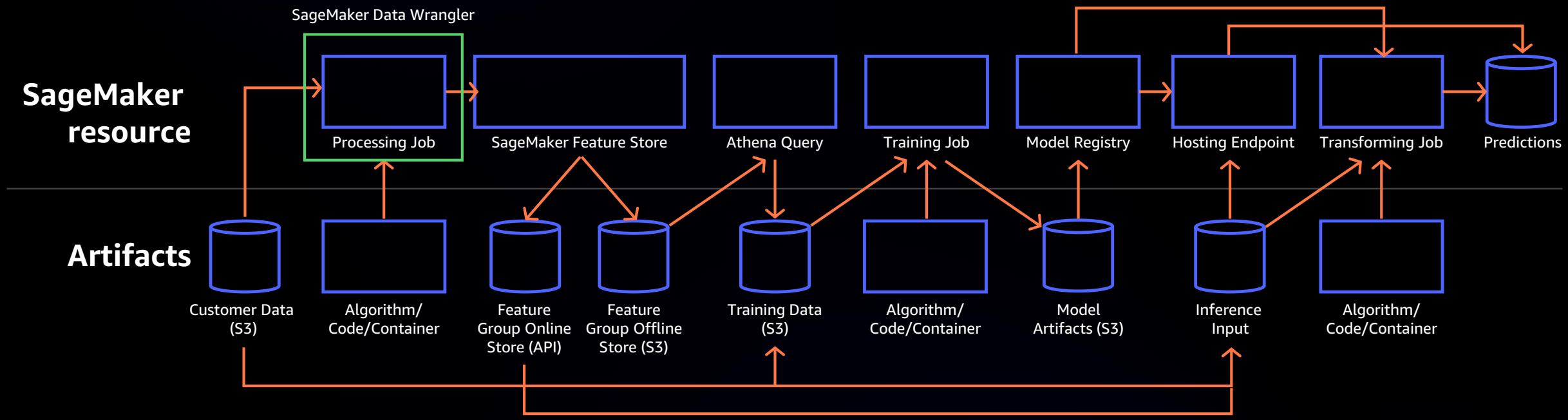
SageMaker Pipelines
Workflow orchestration and automation

SageMaker Studio
Integrated development environment (IDE) for ML

aws

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

SageMaker machine learning workflow



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

Amazon SageMaker Data Wrangler

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

SageMaker Data Wrangler

The fastest and easiest way to prepare data for machine learning



Quickly select and query data

Support for data from multiple sources



Easily transform data with built-in data transformations

Use built-in data transformations to convert raw data to features for machine learning



Customize data transformations

Complete flexibility to bring your own custom transformations in PySpark, SQL, or Pandas



Understand data visually

Quickly detect outliers or extreme values – all without writing code



Quickly estimate ML model accuracy

Diagnose potential issues in data preparation workflows that could hinder ML model accuracy



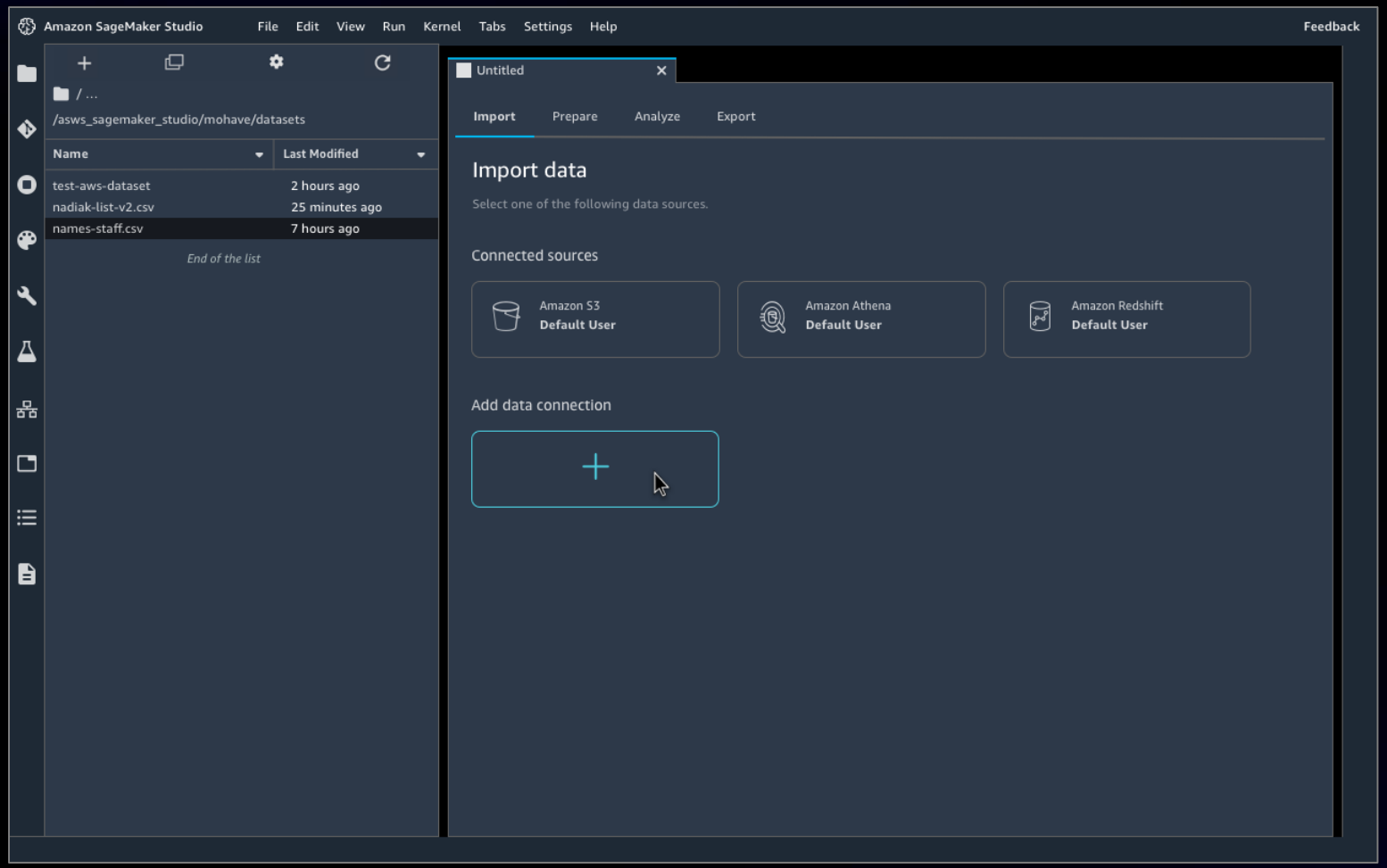
Deploy data preparation workflows into production with a single click

Manage all steps of the data preparation workflow through a single visual interface to quickly operationalize workflows into production settings

aws

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

Quickly select and query data



Select data from Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon S3, and features from SageMaker Feature Store

Write queries for data sources before importing data over to SageMaker Data Wrangler

Import data in various file formats, such as CSV files, Parquet files, and database tables directly into Amazon SageMaker

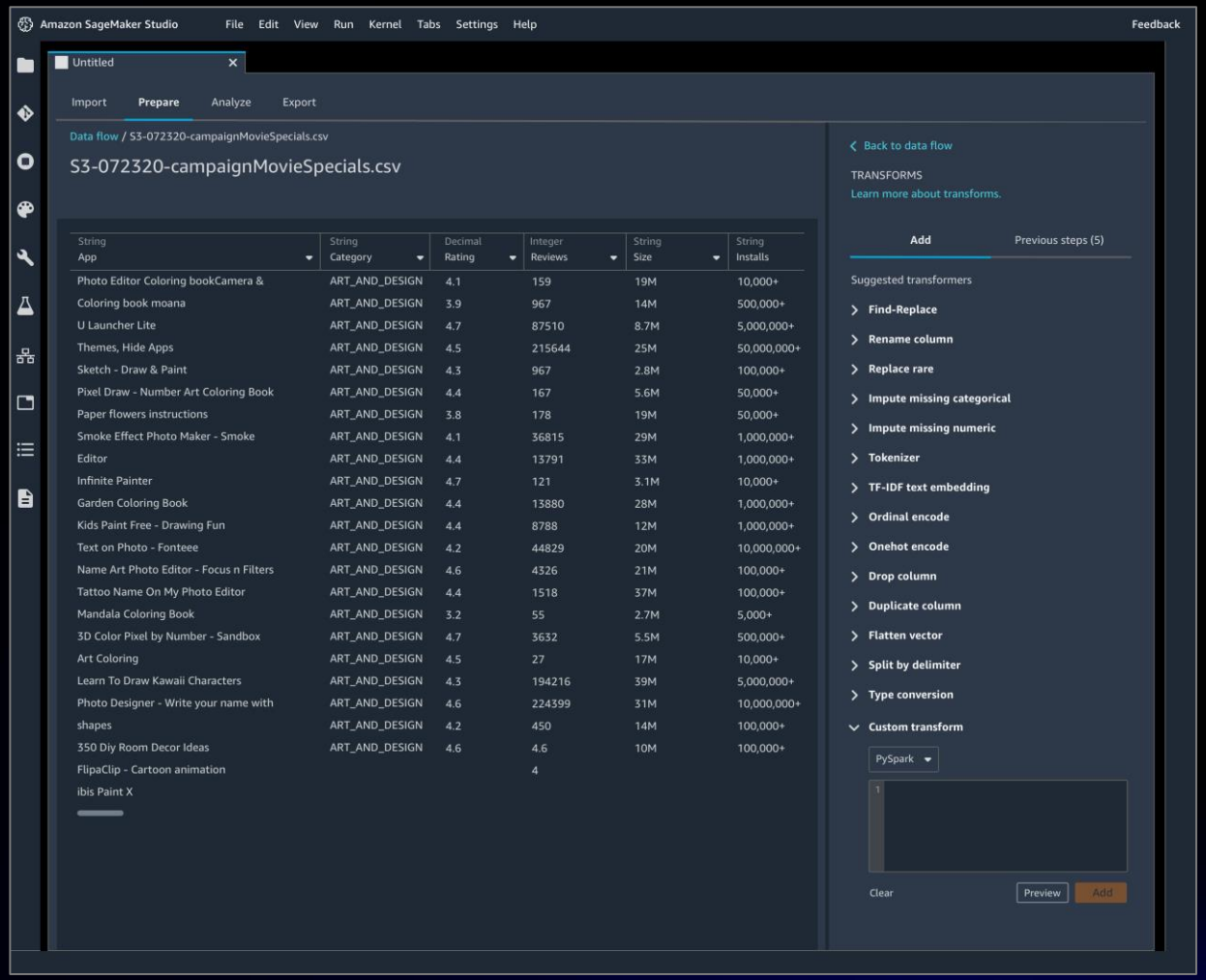
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Easily transform data

Transform your data without writing a single line of code using over 300 built-in data transformations

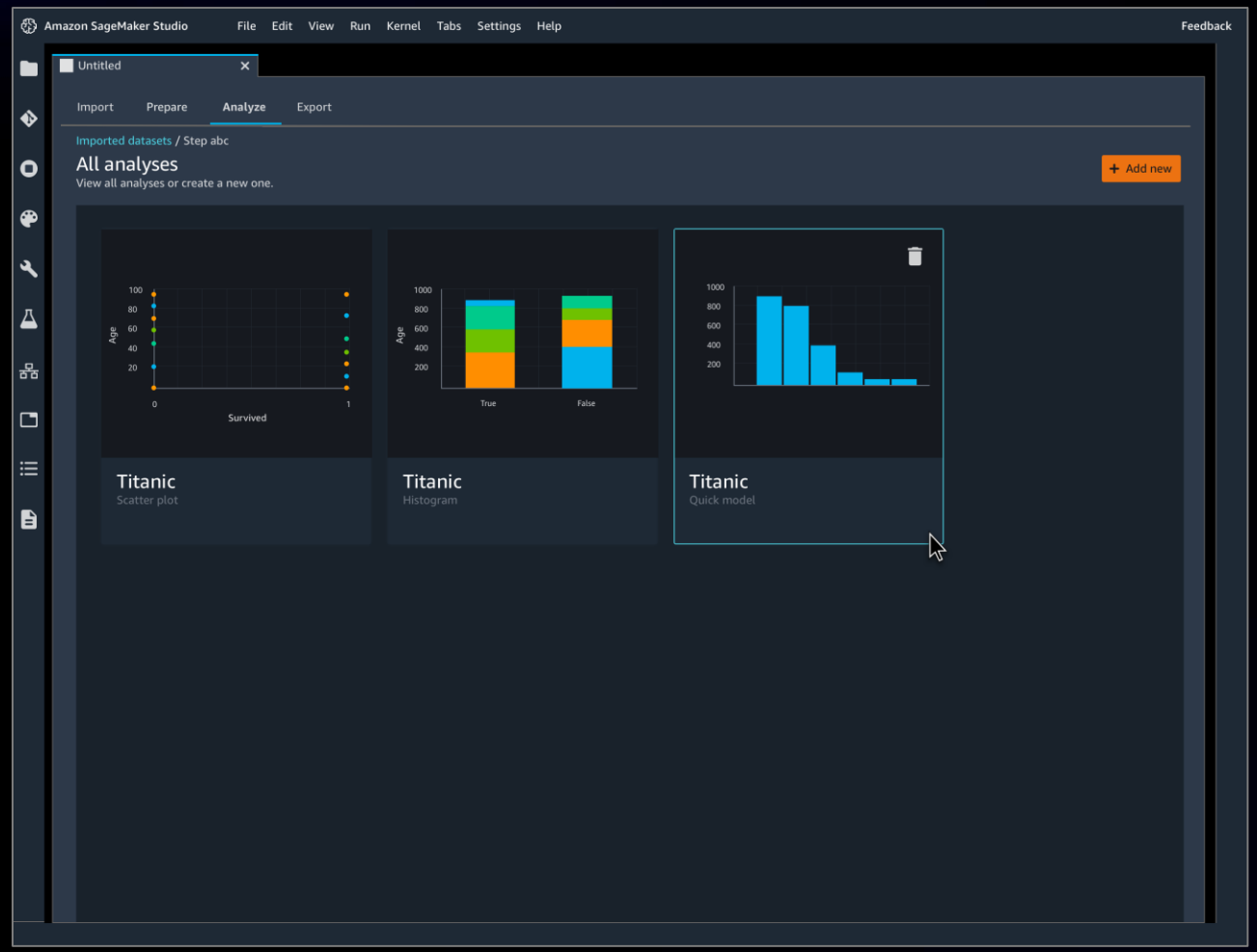
Built-in data transformations include convert column type, rename column, and delete column

Author custom transformations in PySpark, SQL, and Pandas



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

Understand your data visually



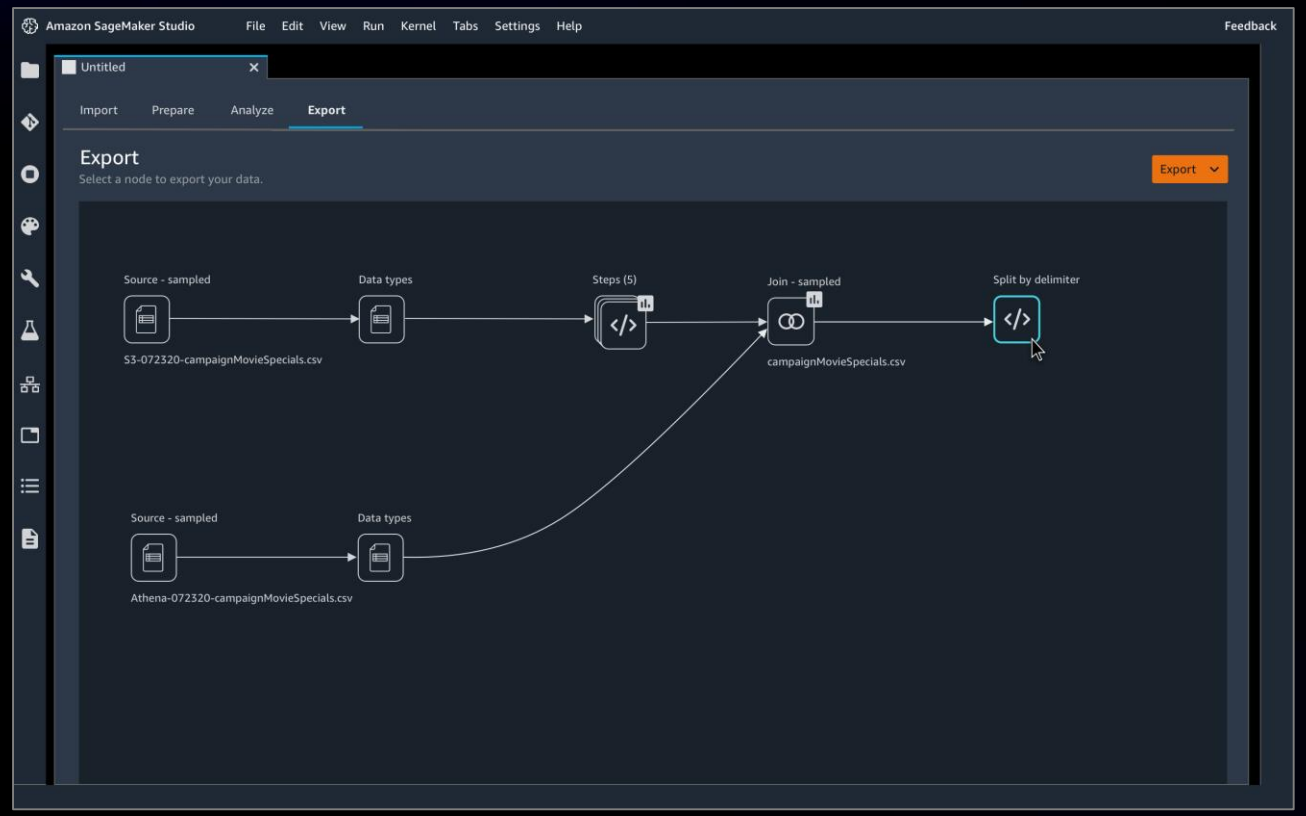
Intuitively understand your data with a set of preconfigured visualization templates

Preconfigured visualization templates include histograms, scatter plots, box and whisker plots, line plots, and bar charts

Interactively create and edit your own visualizations so you can quickly detect outliers or extreme values

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

Deploy data preparation workflows into production



Export data preparation workflows to a notebook or Python code

Integrate your workflow with SageMaker Pipelines to automate model deployment and management

Publish created features to SageMaker Feature Store for reuse and syndication across teams and projects

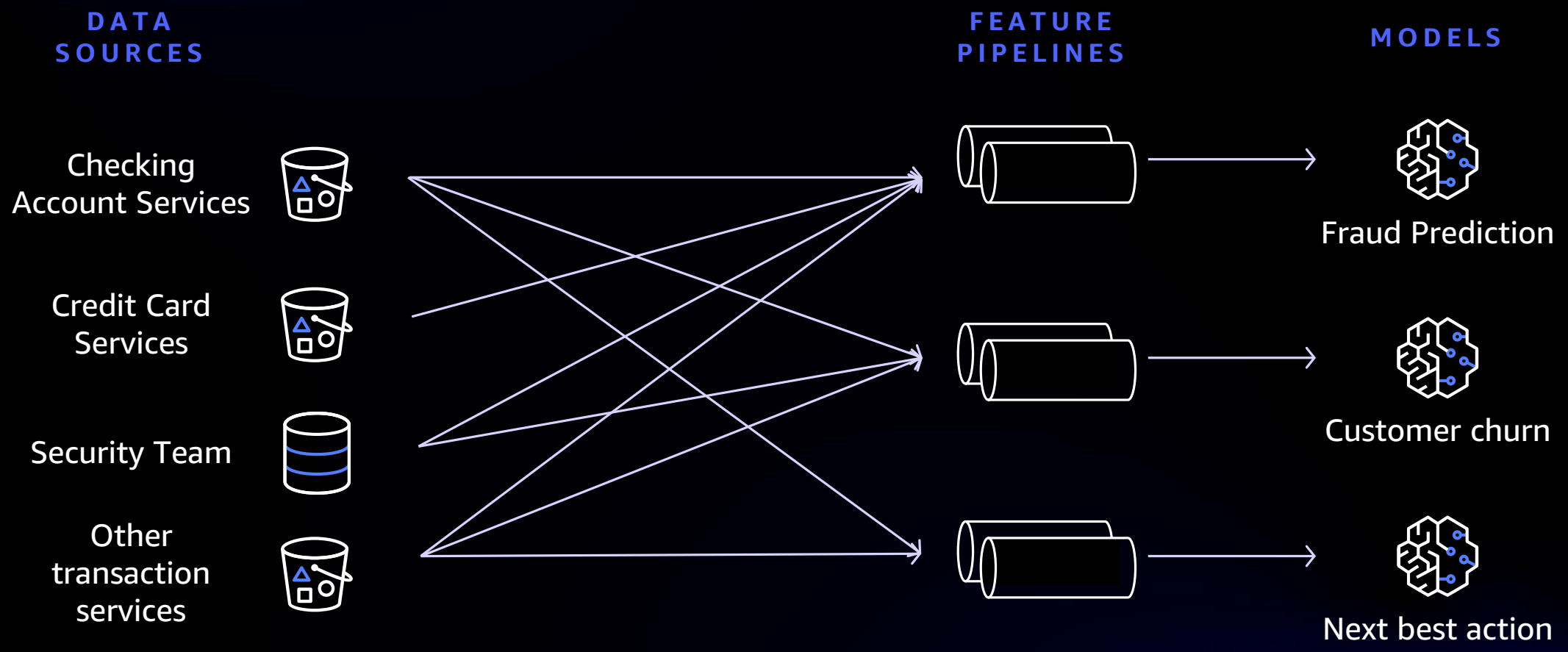
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker Feature Store

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

Without a feature store

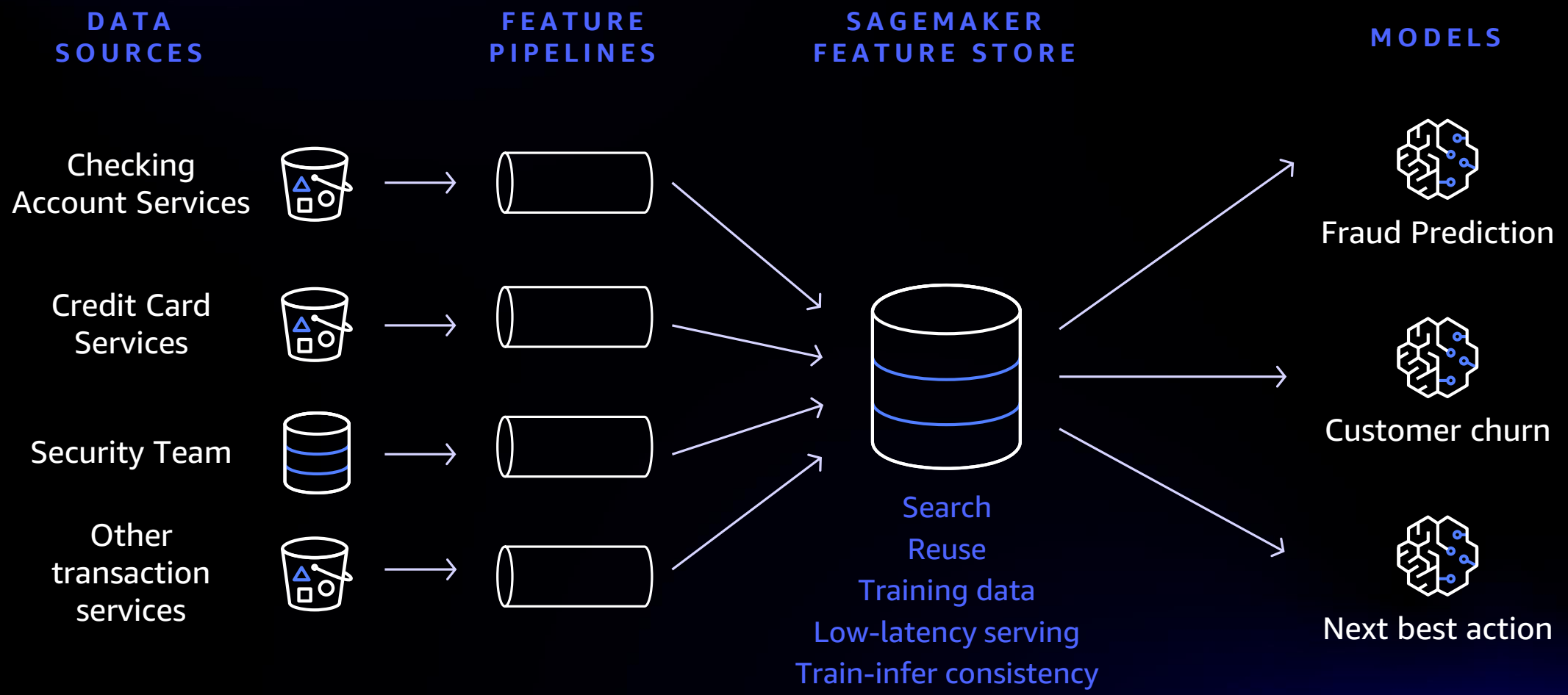
STANDALONE FEATURE ENGINEERING FOR EACH NEW MODEL



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

With SageMaker Feature Store ...

BUILD FEATURES ONCE, AND REUSE THEM ACROSS TEAMS AND MODELS



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

SageMaker Feature Store

Accelerate machine learning with a feature store



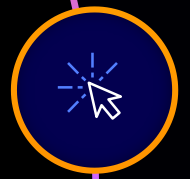
Batch and streaming ingestion

High-throughput writes and reads for features



Feature metadata and data cataloging

Store metadata for features; leverage automatic data cataloging



Online and offline features

Online features for real-time prediction; offline features for model training and batch prediction



Feature consistency

Service maintains consistency between Online and Offline Store to avoid training-inference skew



Security and access control

Access control for feature data and feature metadata



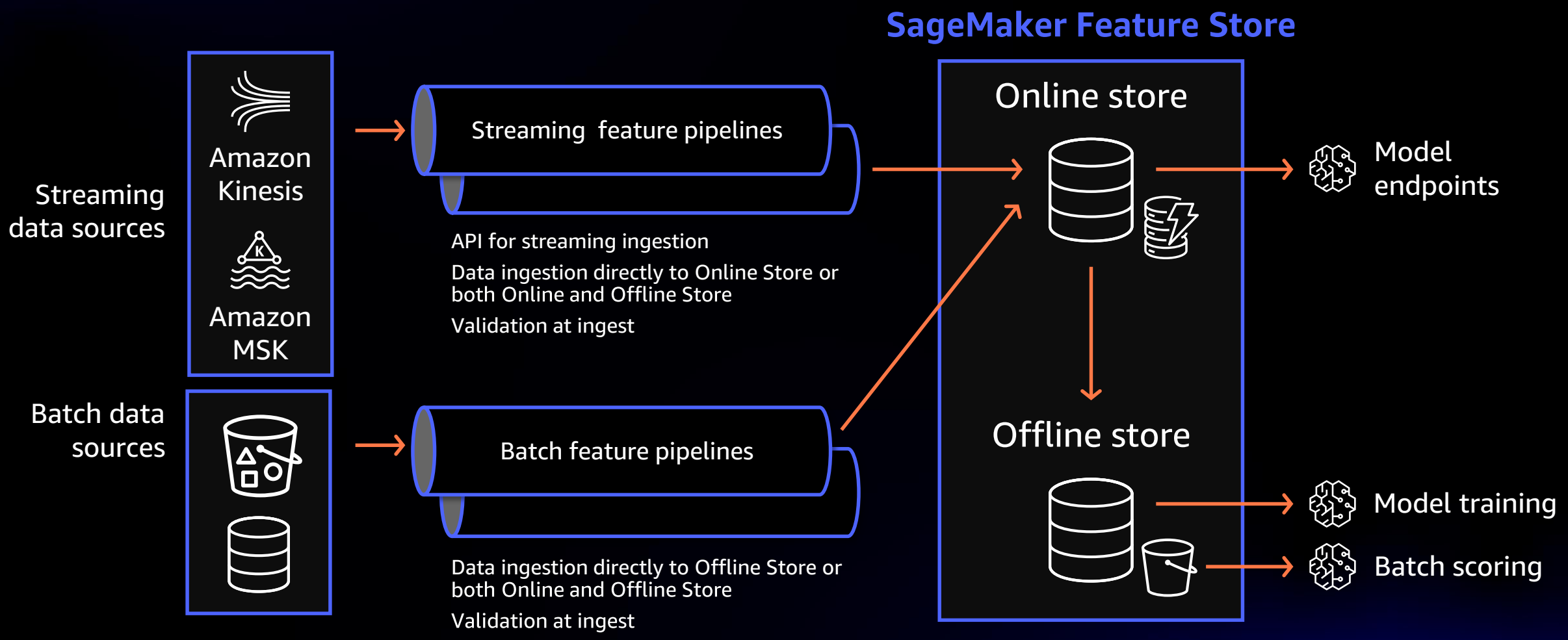
Serve features with millisecond latency

Millisecond-latency serving of features for real-time inference

aws

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

SageMaker Feature Store



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

Feature consistency with Online and Offline Store

- Online store keeps **latest feature values** for each record identifier
- Online feature values are available **immediately**, for use by any model

- Offline store appends each new record, keeping a **history of all feature values**
- Offline store provides <15 minute read after write consistency

```
# 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

© 2021, 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)
```

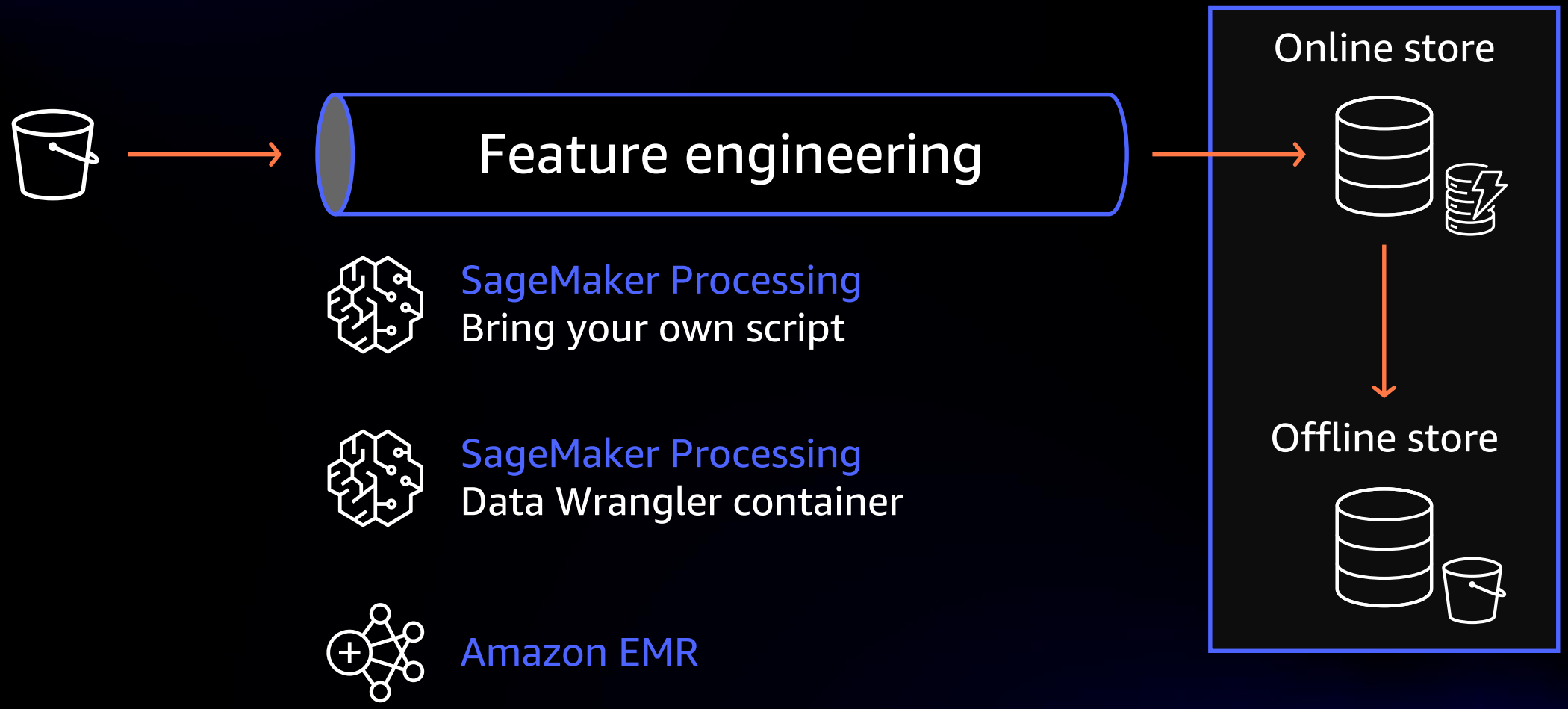
fg.ingest()

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

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

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

Feature engineering options



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

SageMaker Feature Store time travel

SageMaker Feature Store appends all feature values to Offline Store

Feature producer records *event-time* for each entity

SageMaker Feature Store records *write-time* and *api-invocation time*

Query the Offline Store to reconstruct datasets as of a point in time

user-id	number-of-changes	number-of-logins	event-time	api-invocation-time	write-time
1089756453	15	5	2016-02-20T03:26:25	2016-02-20T03:26:27	2016-02-20T03:26:27
2039867858	85	12	2016-02-20T03:26:32	2016-02-20T03:27:37	2016-02-20T03:27:37
1129867567	20	7	2018-02-20T05:26:32	2018-02-20T05:30:32	2018-02-20T05:30:32
5567890234	35	10	2019-02-20T10:26:32	2019-02-20T10:30:35	2019-02-20T10:30:35
3456782341	10	13	2019-02-20T14:26:32	2019-02-20T14:26:32	2019-02-20T14:26:32

```
SELECT *
FROM
(SELECT *,
row_number()
OVER (PARTITION BY Record_ID

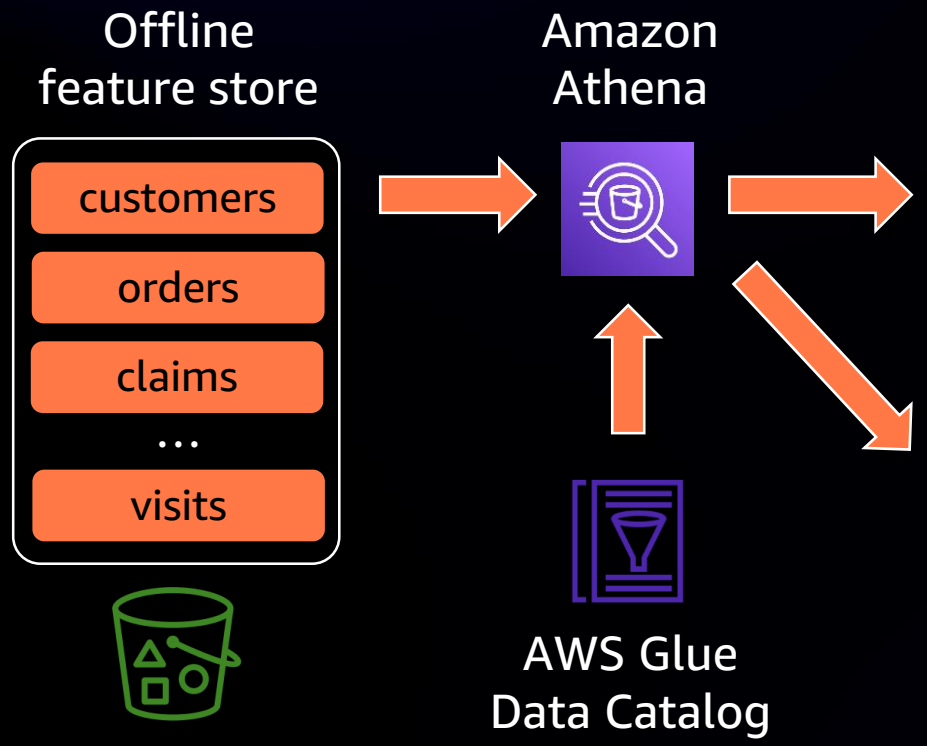
ORDER BY Event_Time desc, Api_Invocation_Time DESC, write_time
DESC) AS row_num
FROM US-Accidents-
FeatureGroup.FeatureStoreDataCatalog.FeatureStoreDatabase.US-
Accidents-FeatureGroup.FeatureStoreDataCatalog.Us-Accidents-
Table

WHERE Event_Time <= '2020-09-01'
WHERE row_num = 1 AND
NOT is_deleted
```

aws

eserved.

Query from Offline Store for training or batch transform datasets



Training

```
SELECT   loyalty_status,   last_purch_date,
         last_purch_amt,   churn_label
FROM     customers
```

customer_train.csv → SageMaker Training

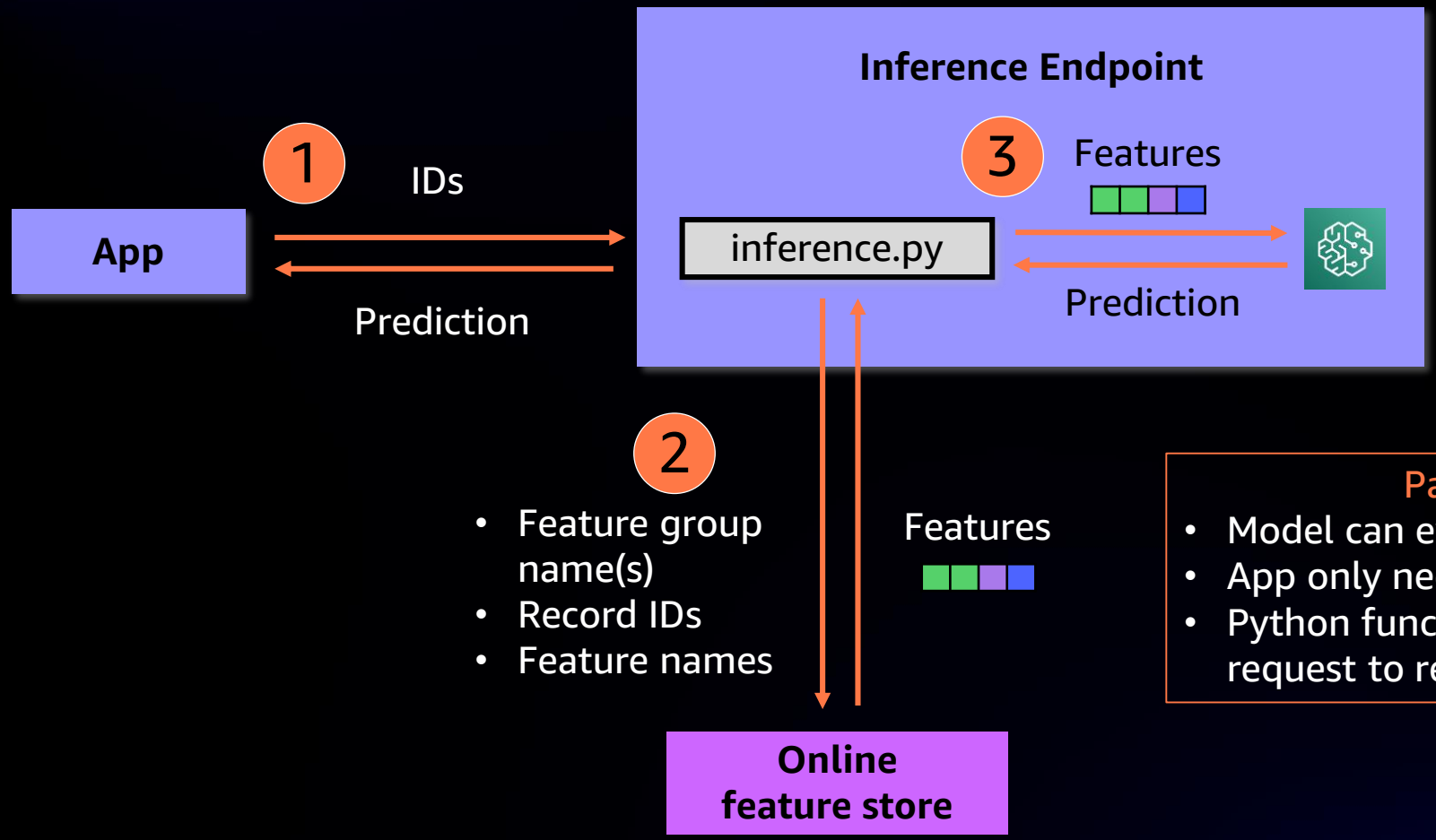
Batch scoring

```
SELECT   id,  loyalty_status,
         last_purch_date, last_purch_amt
FROM     customers
```

customer_score.csv → SageMaker Batch Transform

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

Inference pattern 1: Endpoint-based feature lookup

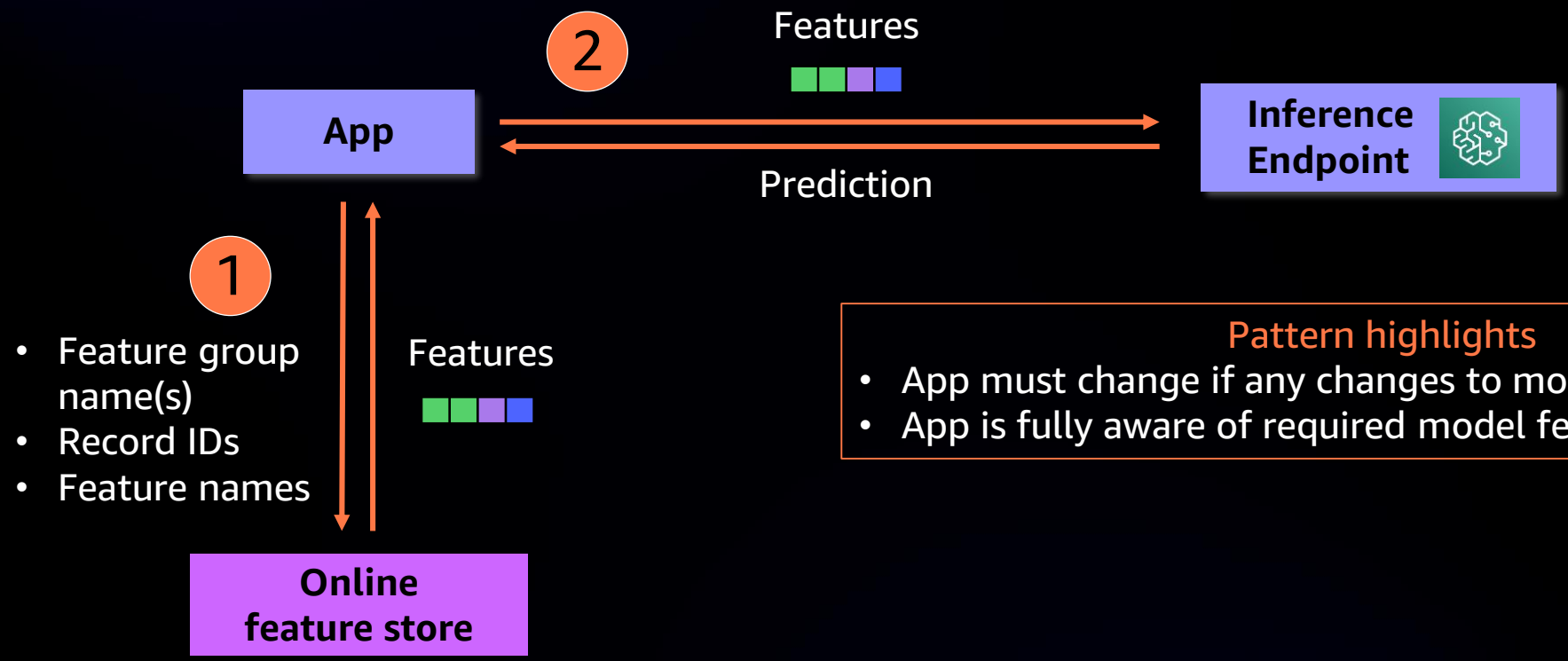


Pattern highlights

- Model can evolve without changing app
- App only needs IDs
- Python function intercepts prediction request to retrieve latest features

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

Inference pattern 2: App-based feature lookup



Pattern highlights

- App must change if any changes to model input features
- App is fully aware of required model features

aws

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

SageMaker Feature Store @Fidelity

This presentation is a case study of the
Fidelity Investments experience

This is not an endorsement or recommendation
of any vendor, product, or service

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

1946

Fidelity Management & Research Company Founded

1965

Fidelity Purchases First Computer





1995

Fidelity Becomes an Internet Pioneer with First Mutual Fund to Create a Home Page

2016

Fidelity's First Product Application Deployed to the Cloud

2019

Fidelity Launches Multi-Cloud Hybrid Strategy

TODAY

4,300+ Applications on Public Cloud

aws Fidelity INVESTMENTS

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

What are we sharing today?

Why Feature Store?

Architecture

Feature ingestion

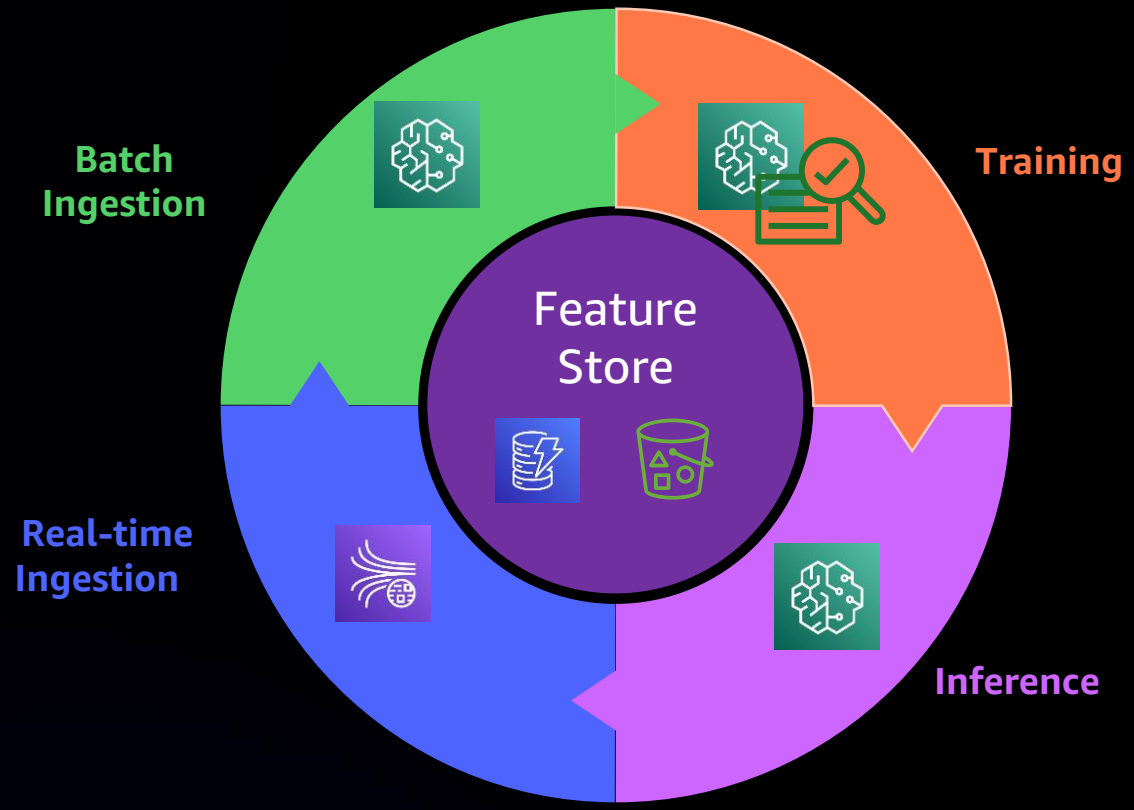
Real-time inference

Model training

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

Why Feature Store?

- Faster time-to-market
- Reusability, search and discovery
- Feature consistency
- Feature lineage, versioning
- Low latency access of features



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

Architecture

Event-driven

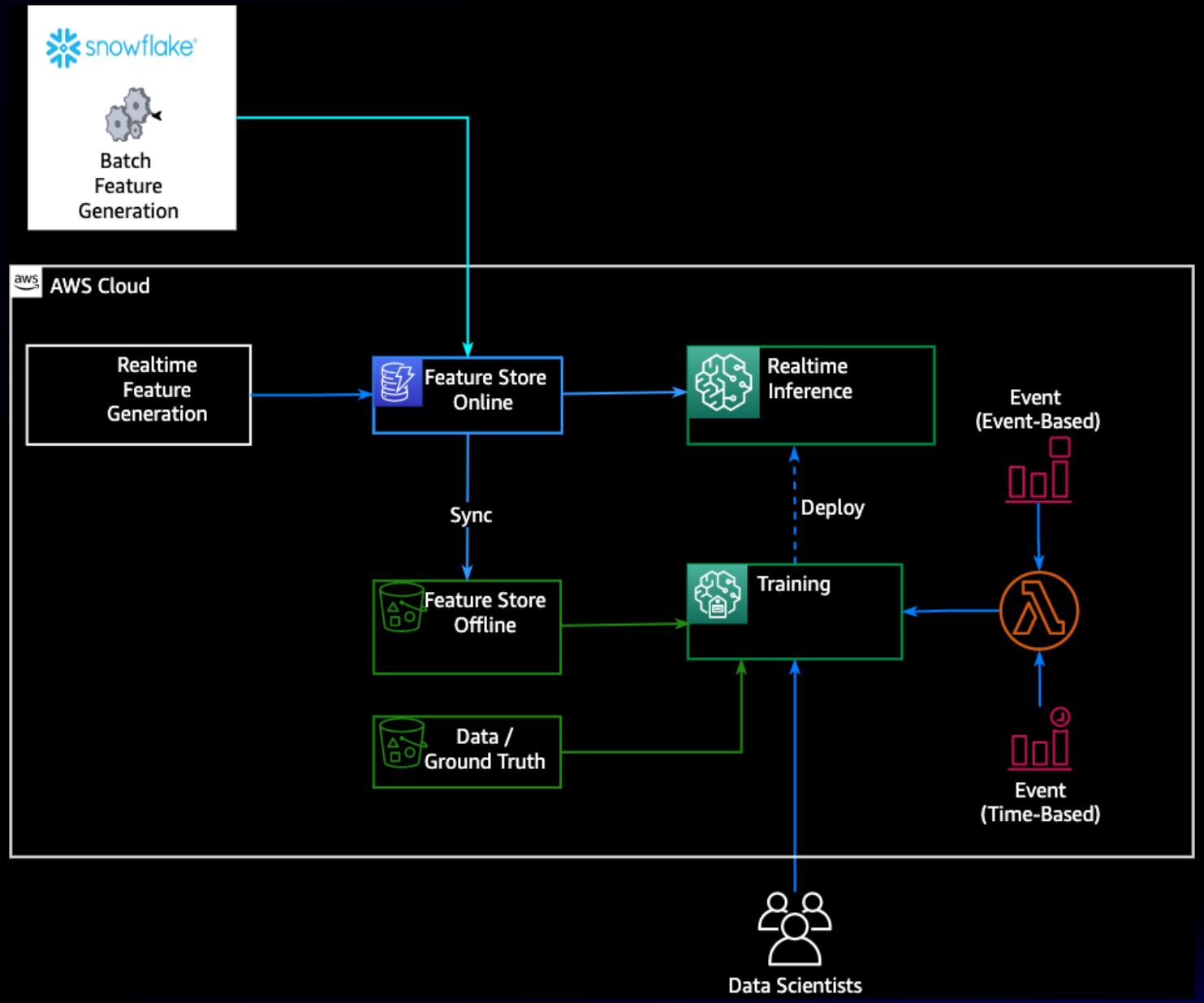
Configurable

Managed services

Serverless

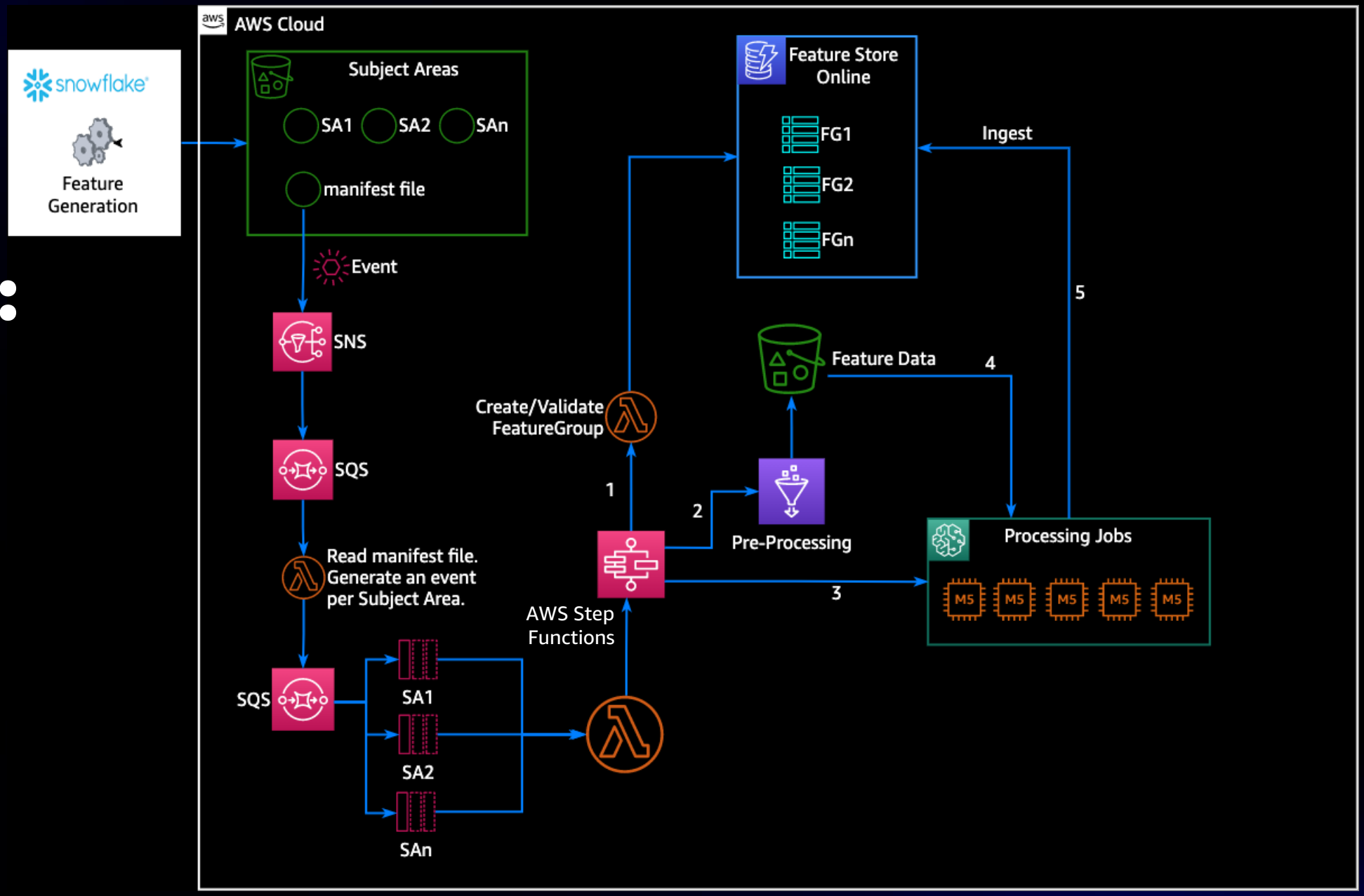
Auto scaling

Highly available



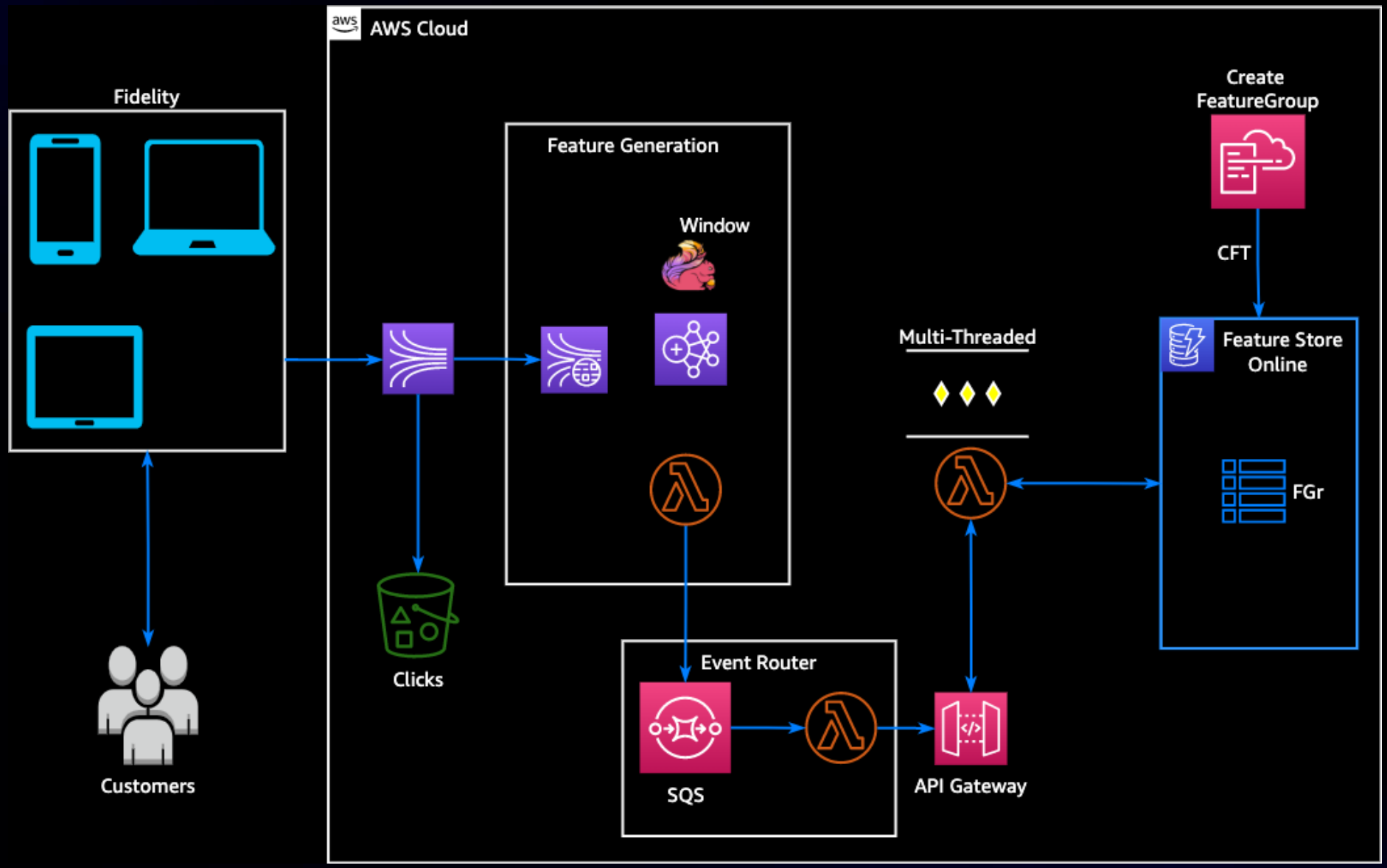
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Feature Ingestion: Batch



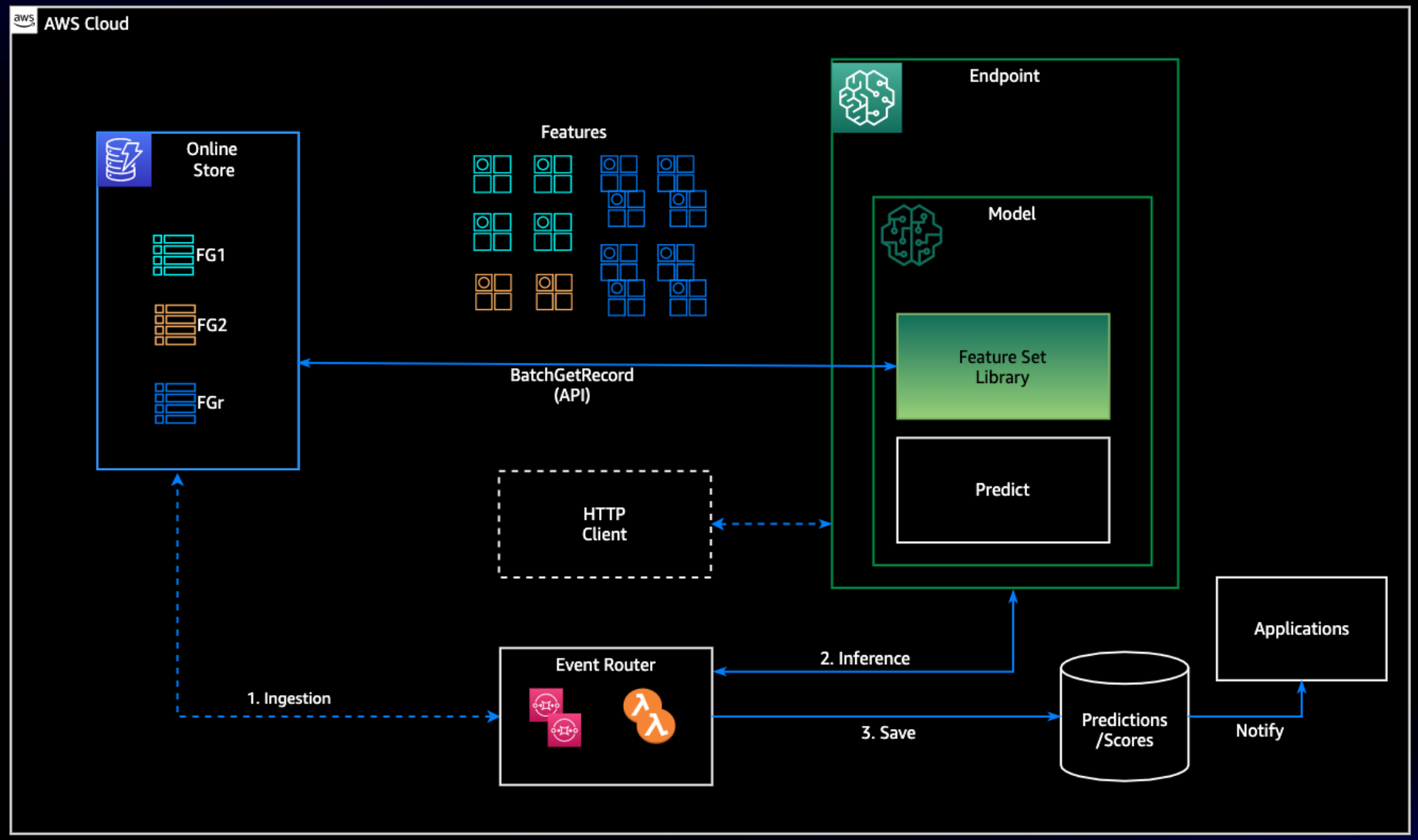
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Feature Ingestion: Real-time



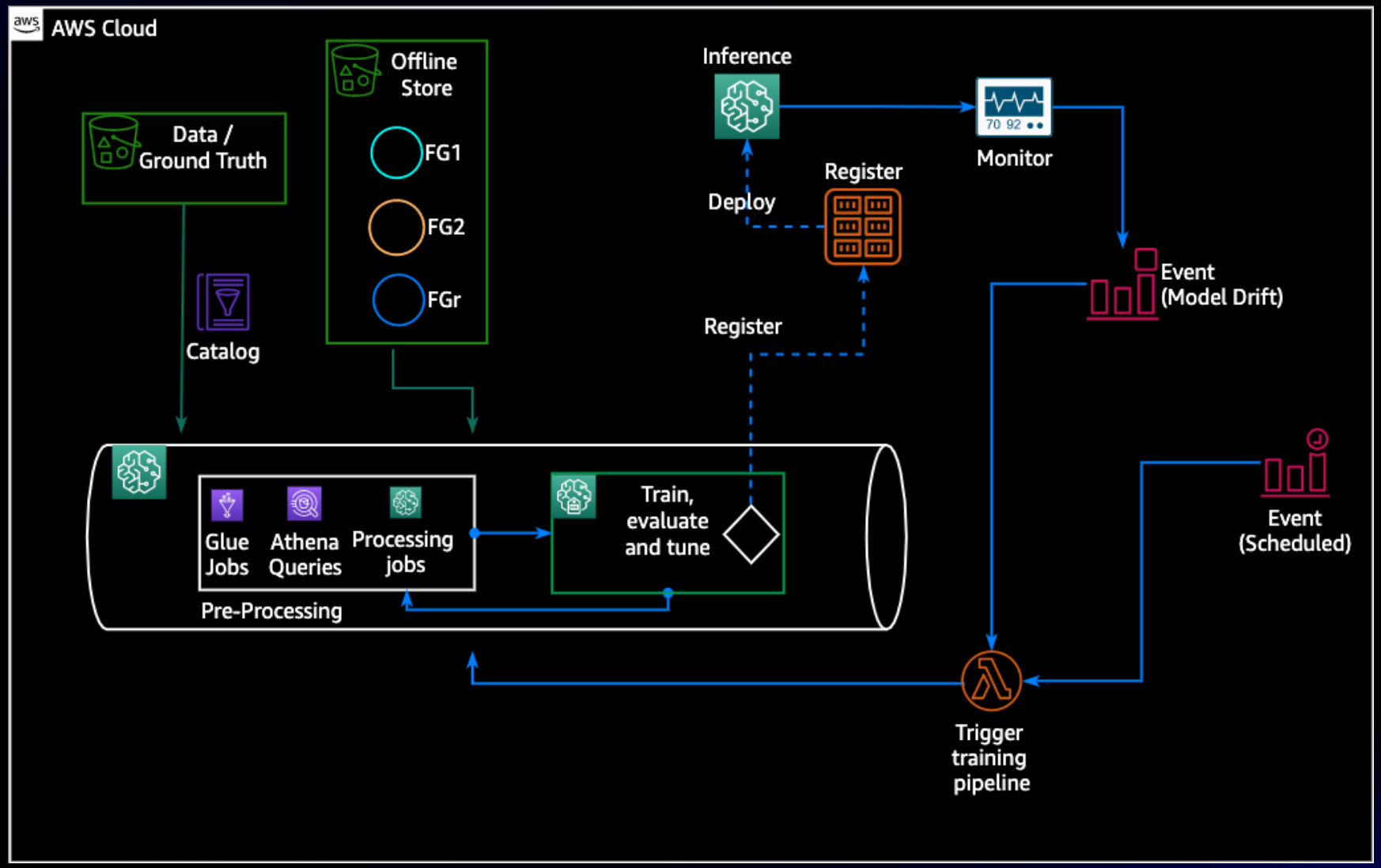
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Real-time inference



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

Model training



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

It’s only the beginning …

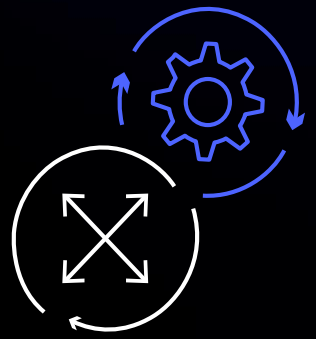
- Flexible feature generation
- Onboard additional subject areas to Feature Store
- Onboard additional real-time models

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

Amazon SageMaker Ground Truth

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

Data labeling challenges



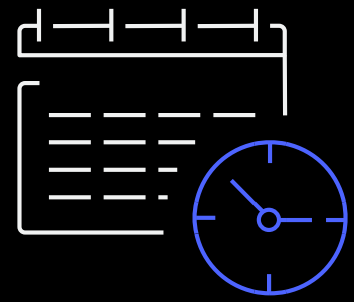
Massive scale

ML models need large, labeled datasets



High accuracy

ML models depend on accurately labeled data



Time consuming

Building training datasets takes up to 80% of a data scientist's time

aws

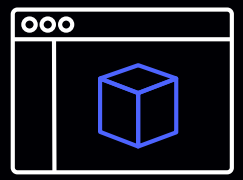
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker Ground Truth capabilities

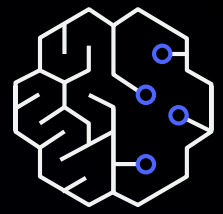
Easily label training data for machine learning at scale



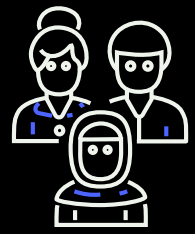
Fully managed and secure



30+ labeling workflows



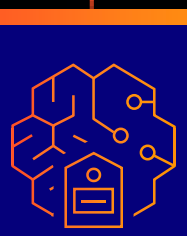
Assistive tooling, auto-labeling, consensus



Your choice of workforce

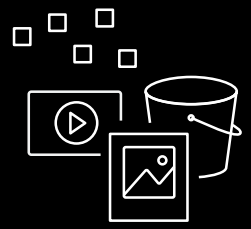
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How Amazon SageMaker Ground Truth works



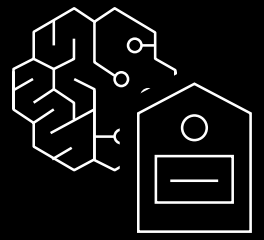
SageMaker Ground Truth

Easily label training data for machine learning at scale



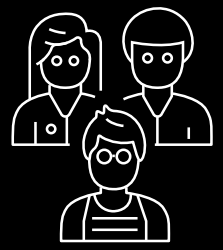
Raw data

Input raw data into Amazon S3



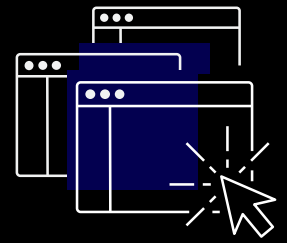
Labeling

Create a labeling job using custom or built-in workflows



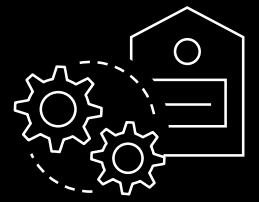
Labelers

Choose from different groups of labelers



Assistive labeling

Labelers complete tasks using the labeling UI with assistive labeling features



An accurate training dataset is ready for use

aws

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

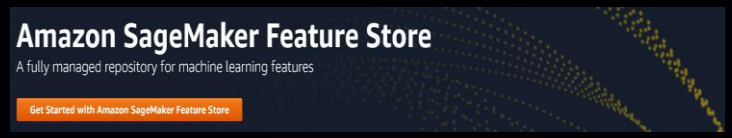
Resources to get started

Amazon SageMaker Data Wrangler



https://aws.amazon.com/sagemaker/data-wrangler/

Amazon SageMaker Feature Store



https://aws.amazon.com/sagemaker/feature-store/

Amazon SageMaker Ground Truth



https://aws.amazon.com/sagemaker/groundtruth/

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

Thank you!

Romi Datta
romidat@amazon.com

Kiran Kandala
kiran.kandala@fmr.com

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