
aws
INNOVATE
AI/ML EDITION

aws INNOVATE
AI/ML EDITION

Build and manage training datasets for machine learning

Praveen Jayakumar
Principal Solutions Architect, AI/ML
AISPL

Agenda

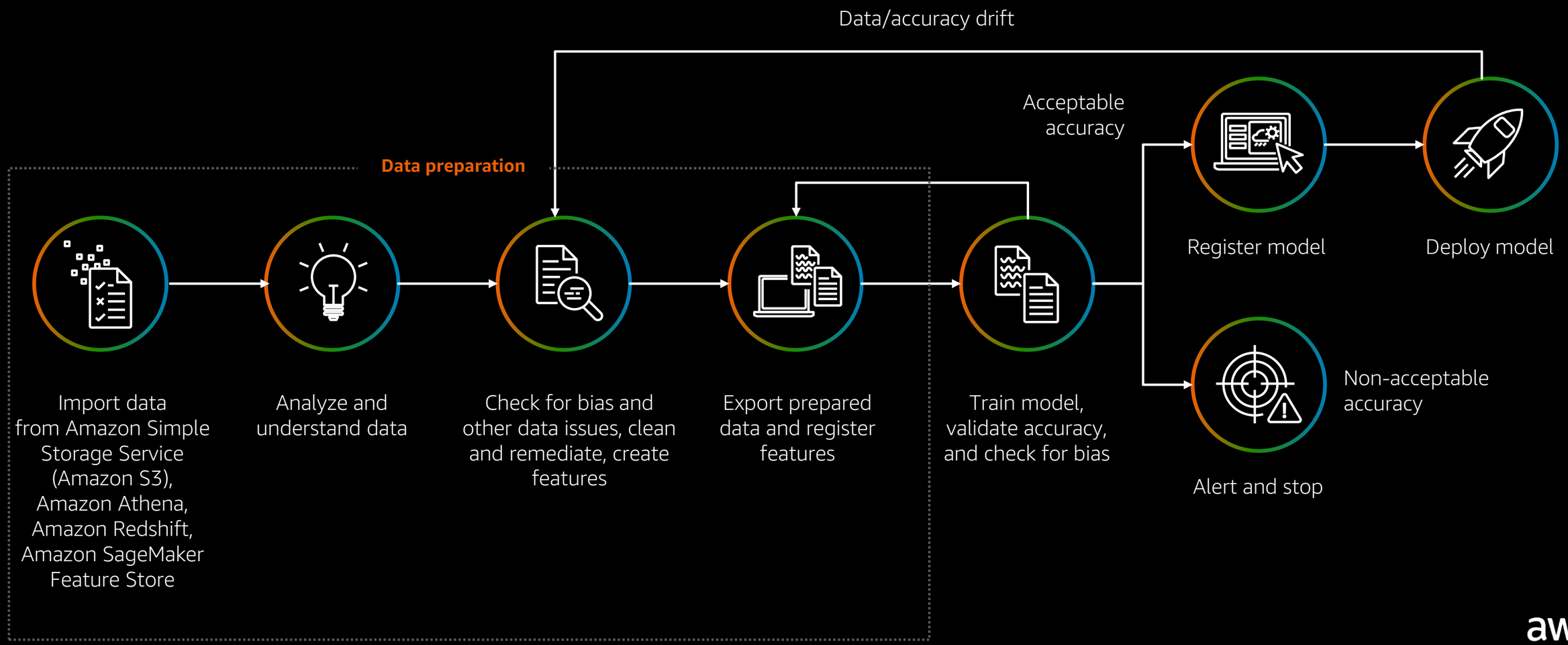
1. Data preparation challenges

2. Amazon SageMaker Data Wrangler

    - Select, query and merge data
    - Data transformation
    - Visualize data and estimate ML model accuracy
    - Deploy data preparation workflow to production

3. Demo

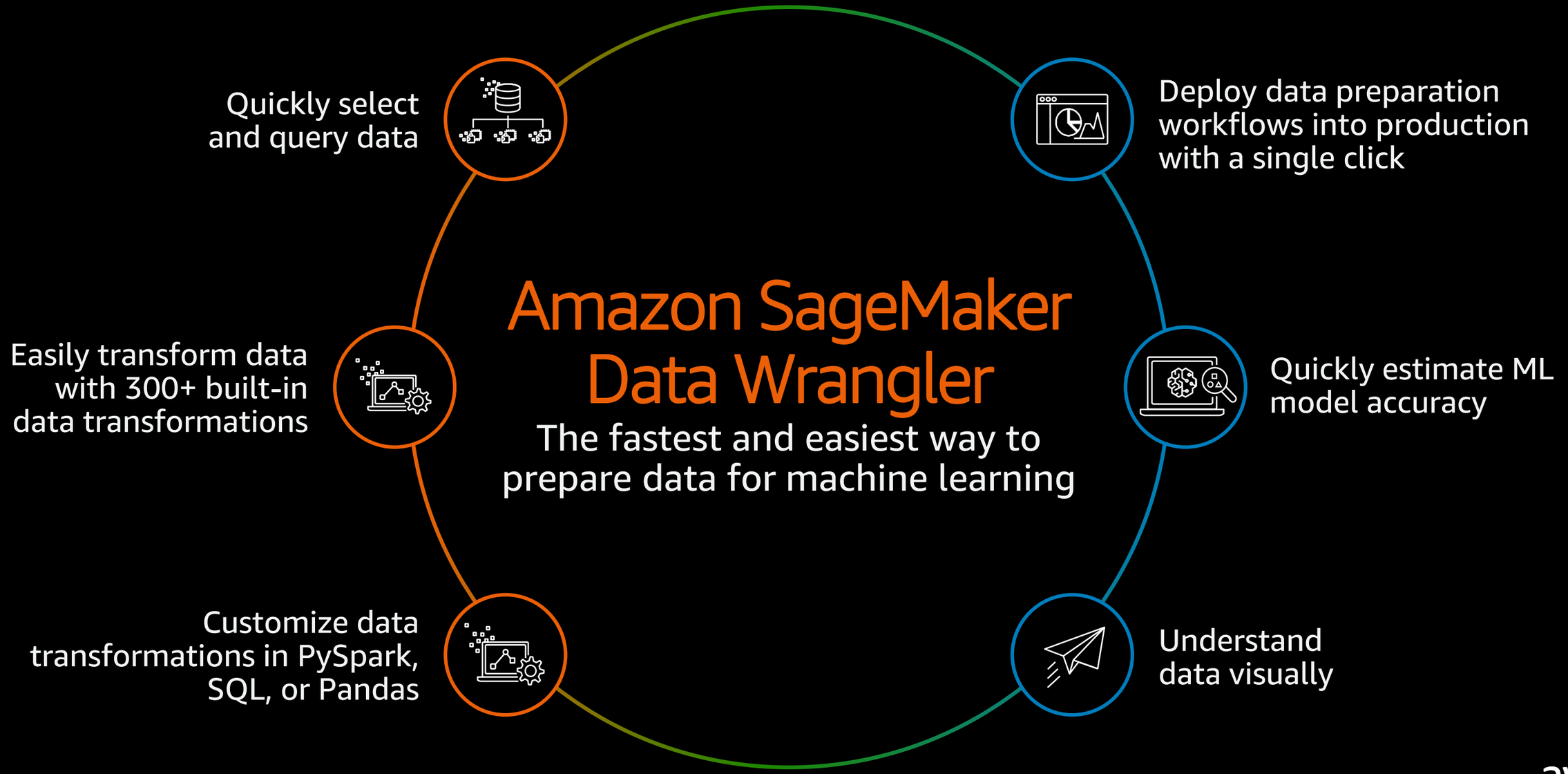
Data preparation is a critical part of the end-to-end ML workflow



Data preparation challenges

- Data preparation is time consuming and requires multiple tools and tasks
- Simple tasks require a lot of code
- Deployment can require a code rewrite, and productionizing can take months

Amazon SageMaker Data Wrangler



Quickly select and query data

SELECT

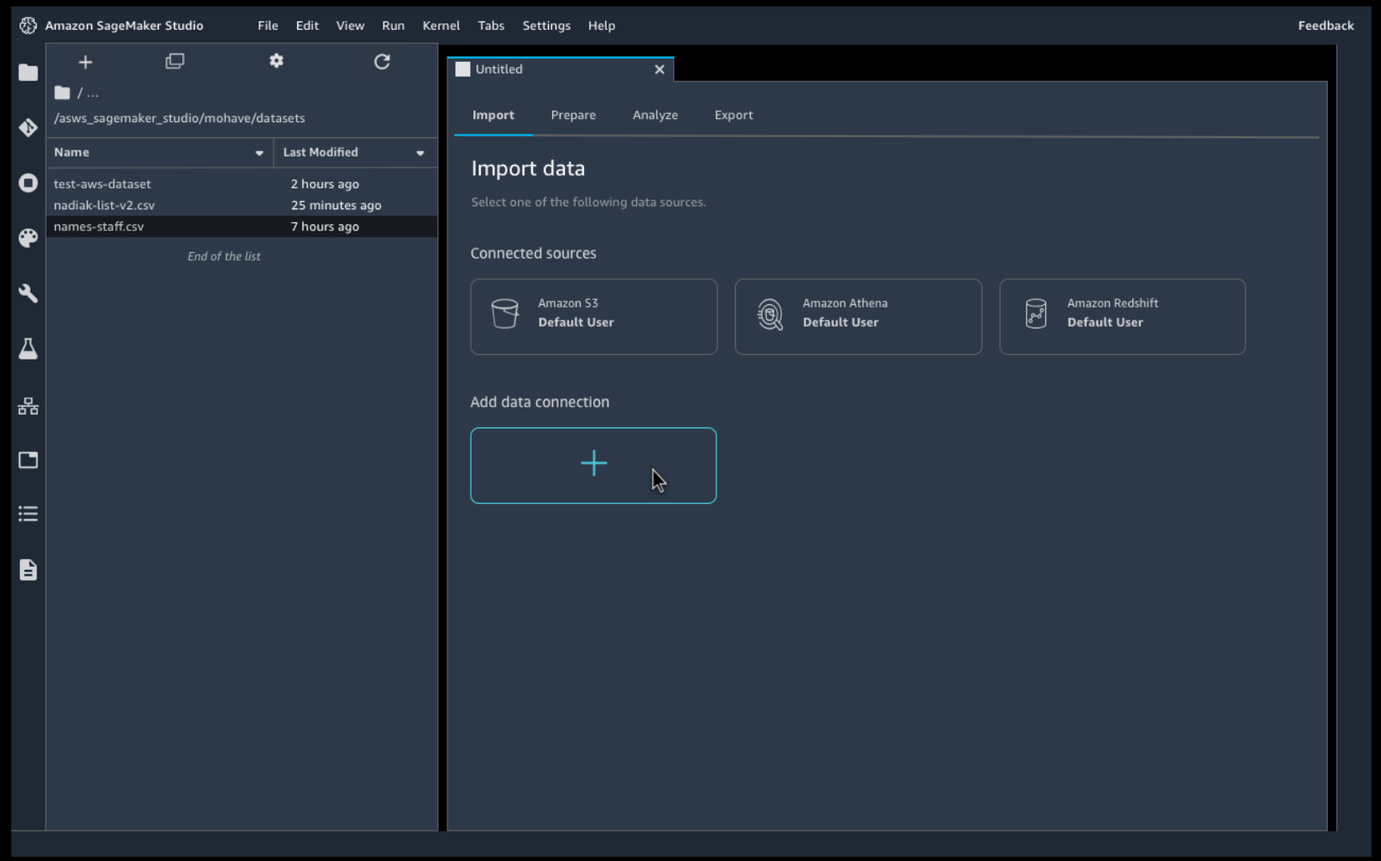
Quickly select data from Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon S3, and features from Amazon SageMaker Feature Store

WRITE

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

IMPORT

Easily import data in various file formats, such as CSV files, parquet files, as well as database tables, directly into Amazon SageMaker



Easily transform data

TRANSFORM

Transform your data without writing a single line of code using 300+ built-in data transforms

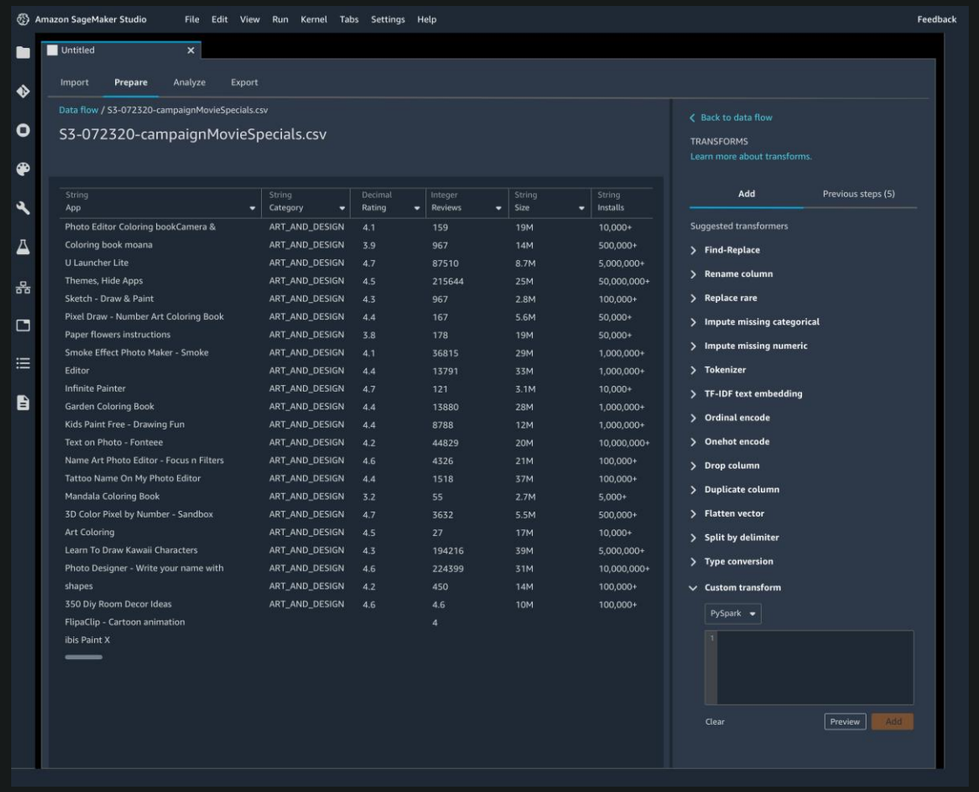
PREPARE

Preconfigured data transforms include:

- Missing value detection and imputation
- Outlier detection and handling
- Featurizers for string and date-time columns
- Column manipulation
- String cleaning and processing tools
- Categorical encoding

AUTHOR

Author custom transforms in PySpark, SQL, and Pandas



Demo – 1

In today's demo we'll get to:

- Import data from Amazon S3, Amazon Redshift, and Amazon Athena
- Check data for missing data and outliers
- Clean our data of possible inconsistencies
- Join and aggregate multiple data sources together

Understand your data visually

UNDERSTAND

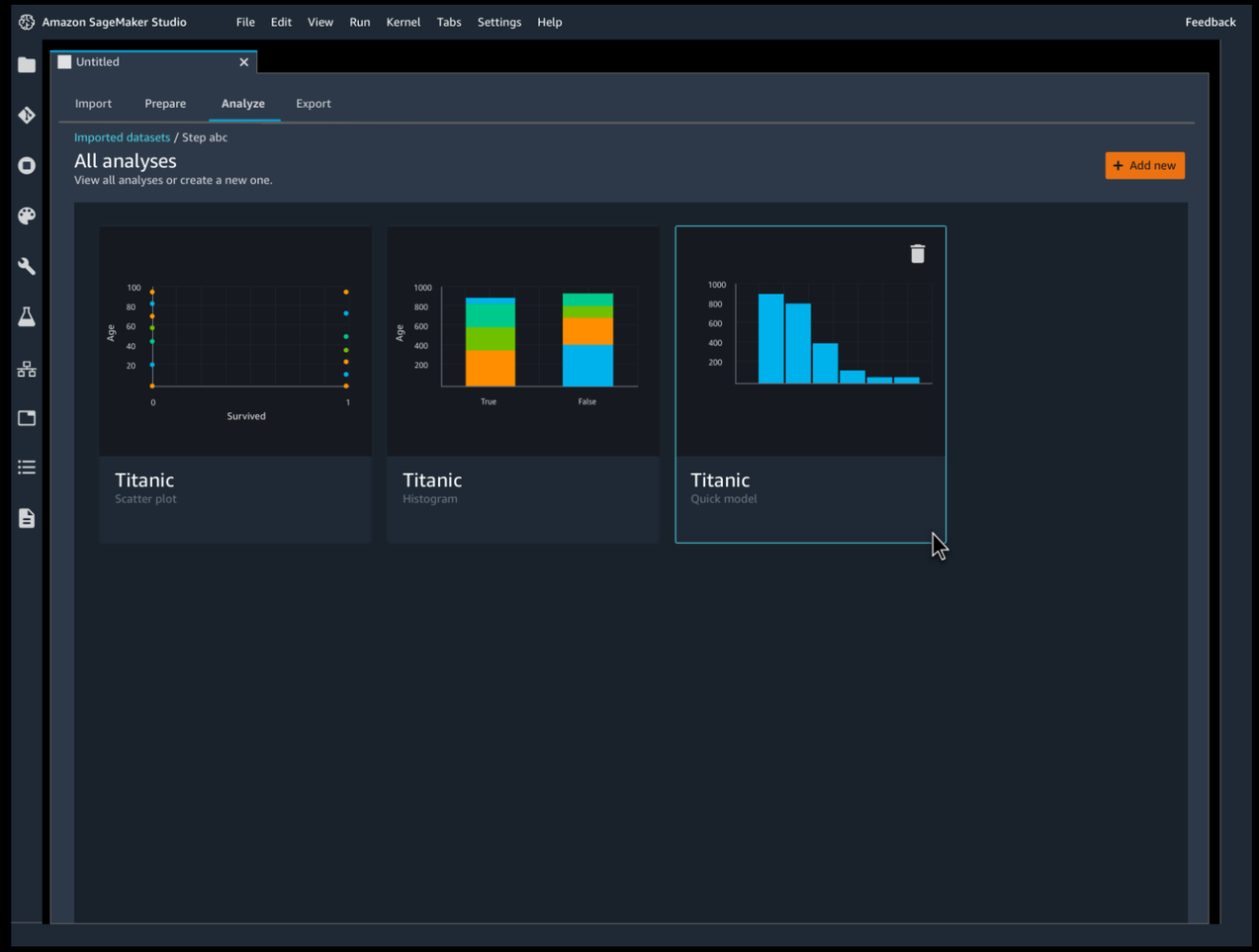
Intuitively understand your data with a set of preconfigured visualization templates which include histograms and scatter plots

CREATE

Create your own custom templates using Altair for data visualization

SAVE

Save and organize your visualizations easily for future reference and reuse



Quickly estimate ML model accuracy

IDENTIFY

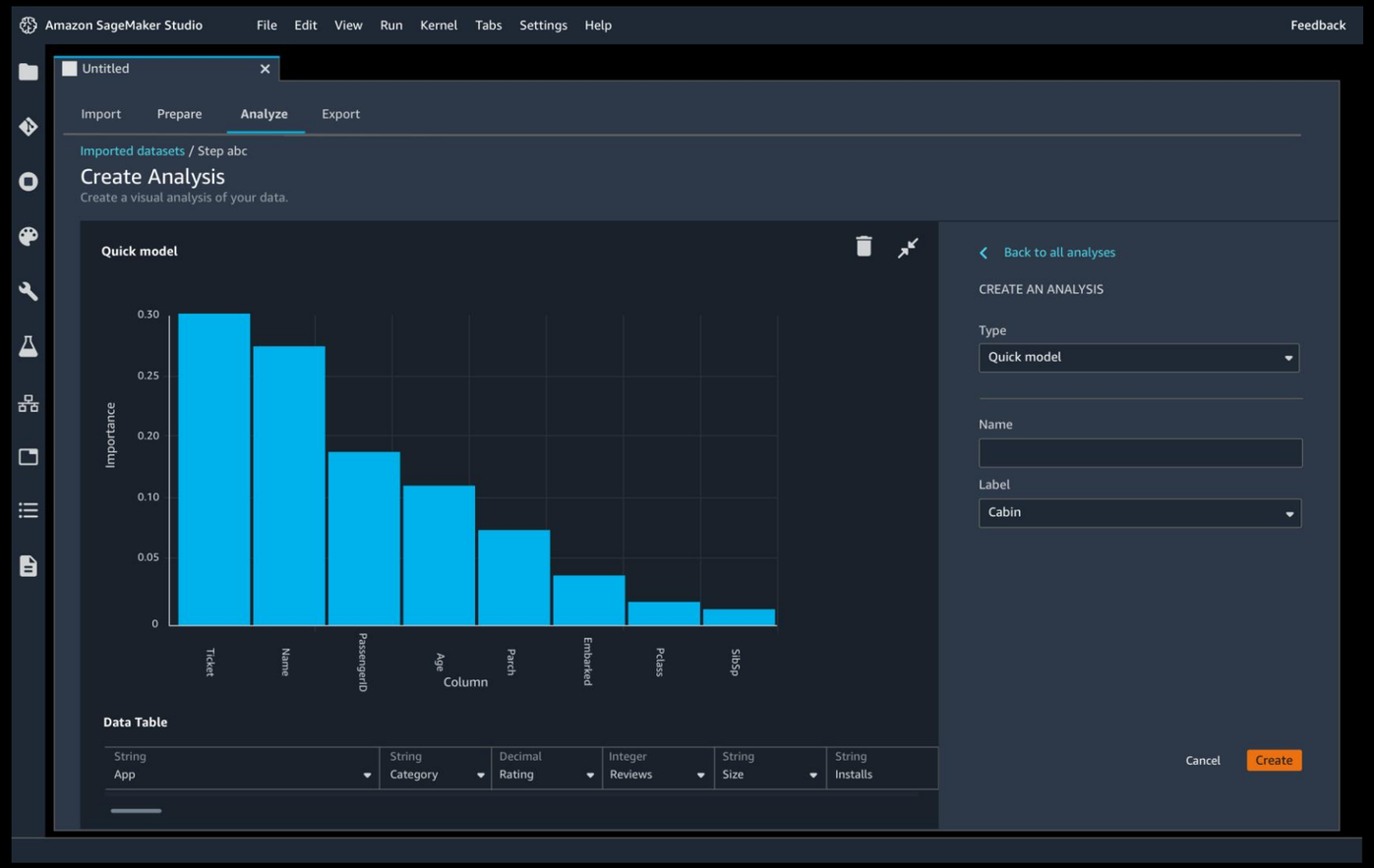
Quickly identify inconsistencies in your data preparation workflow and diagnose issues before ML models are deployed into production

DETECT

Detect which features are contributing to model performance relative to others

DETERMINE

Determine if additional feature engineering is needed to improve model performance



Deploy data preparation workflows into production with a single click

EXPORT

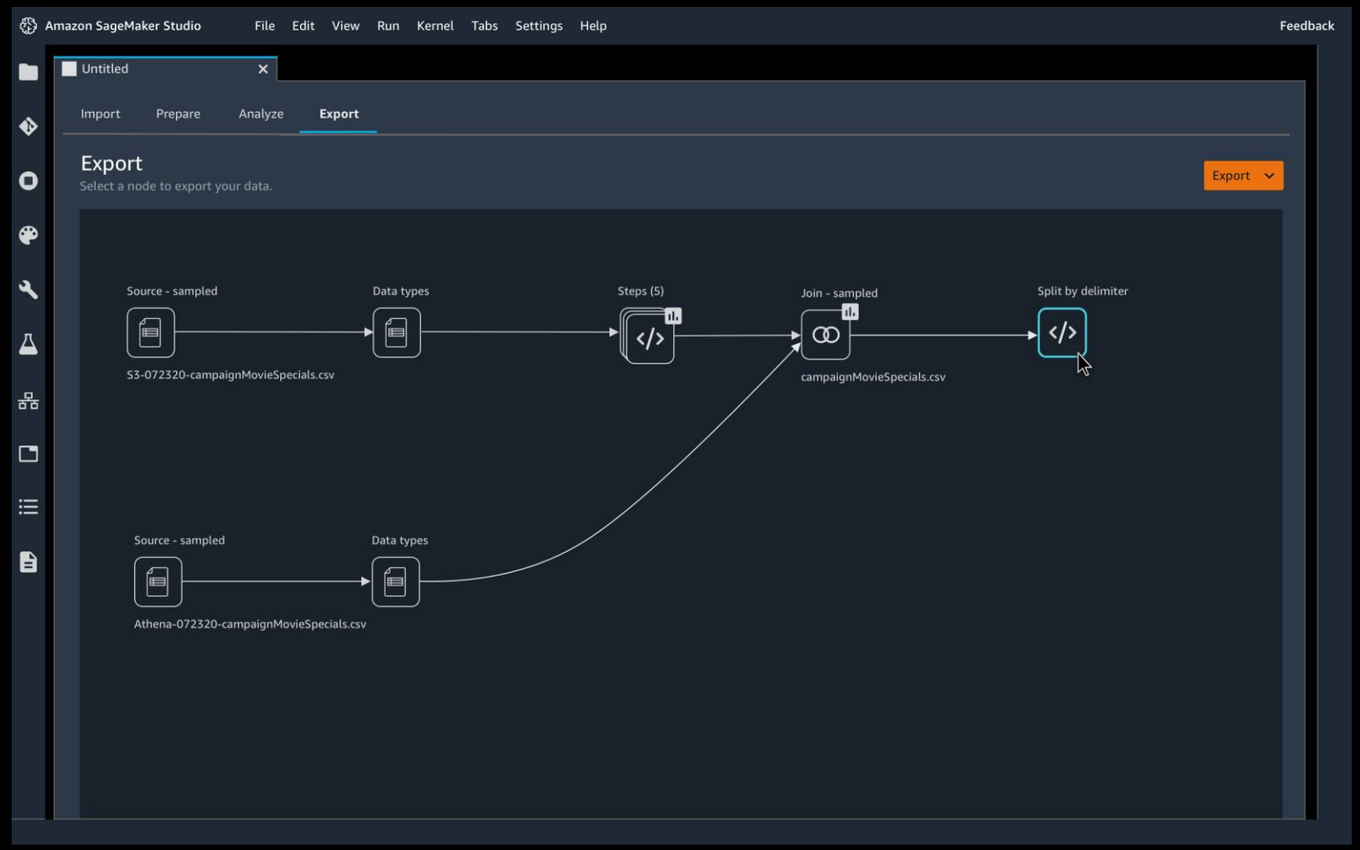
Export data preparation workflow as a processing job notebook or Python code with a single click

INTEGRATE

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

PUBLISH

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



Demo – 2

In today's demo we'll get to:

- Visualize our data using built-in features
- Estimate performance of a model on our dataset
- Export a production-ready data preparation workflow

Get started preparing your data for machine learning

Get started in Amazon SageMaker Studio

Get started with Amazon SageMaker Data Wrangler directly from SageMaker Studio

Learn more about Amazon SageMaker Data Wrangler

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

AWS News Blog

https://aws.amazon.com/blogs/aws/introducing-amazon-sagemaker-data-wrangler-a-visual-interface-to-prepare-data-for-machine-learning

Visit the AI and Machine Learning Resource Hub for more resources

Dive deeper with these resources, get inspired and learn how you can use machine learning to accelerate business outcomes.

- The machine learning journey e-book
- Machine learning enterprise guide
- 7 leading machine learning use cases e-book
- A strategic playbook for data, analytics, and machine learning
- Accelerating ML innovation through security e-book
- … and more!

Visit resource hub »

https://tinyurl.com/aiml-aws

AWS Machine Learning (ML) Training and Certification

Learn like an Amazonian, based on the curriculum we've used to train our own developers and data scientists

AWS is how you build machine learning skills

Courses built on the curriculum leveraged by Amazon's own teams. Learn from the experts at AWS.

Flexibility to learn your way

Learn online with 65+ on-demand digital courses or live with virtual instructor-led training, plus hands-on labs and opportunities for practical application.

Validate your expertise

Demonstrate expertise in building, training, tuning, and deploying machine learning models with an industry-recognized credential.

aws.training/machinelearning

Thank You for Attending AWS Innovate

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apac-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

aws INNOVATE
AI/ML EDITION

Thank you!