
AWS
re:Invent

EMB034

Accelerate data preparation with Amazon SageMaker Data Wrangler

Ajai Sharma

Senior Product Manager, Amazon SageMaker

AWS

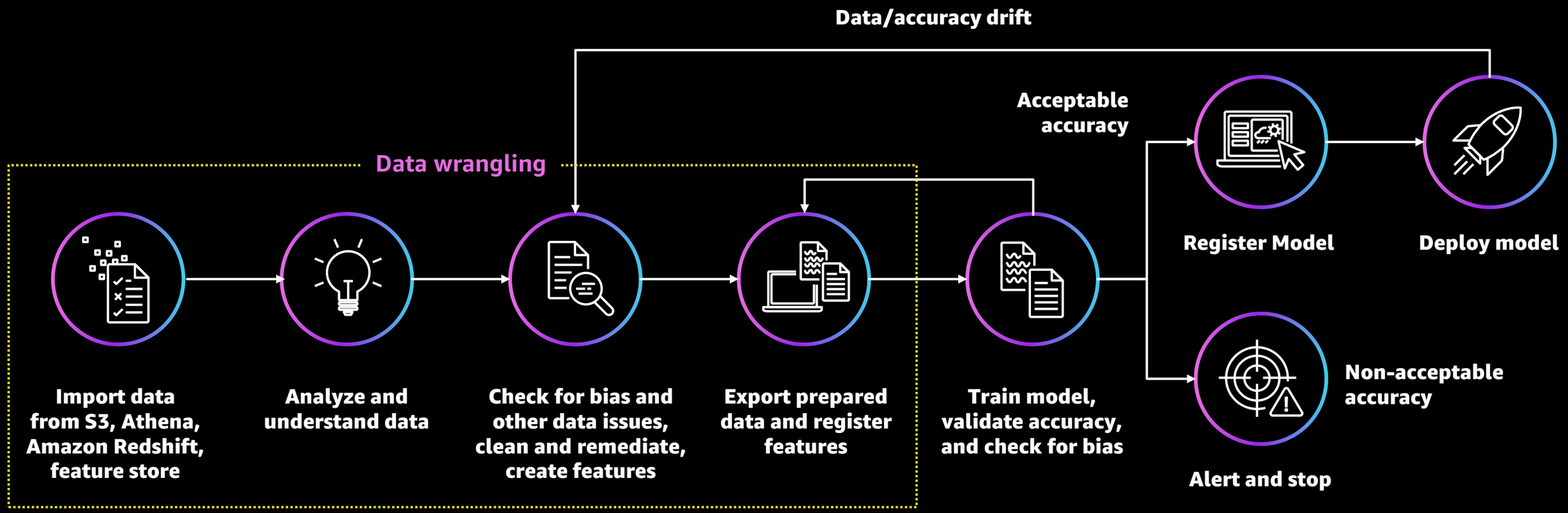
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Data preparation challenges
- Introduction to Amazon SageMaker Data Wrangler
- Demo
- Partnerships & Customers

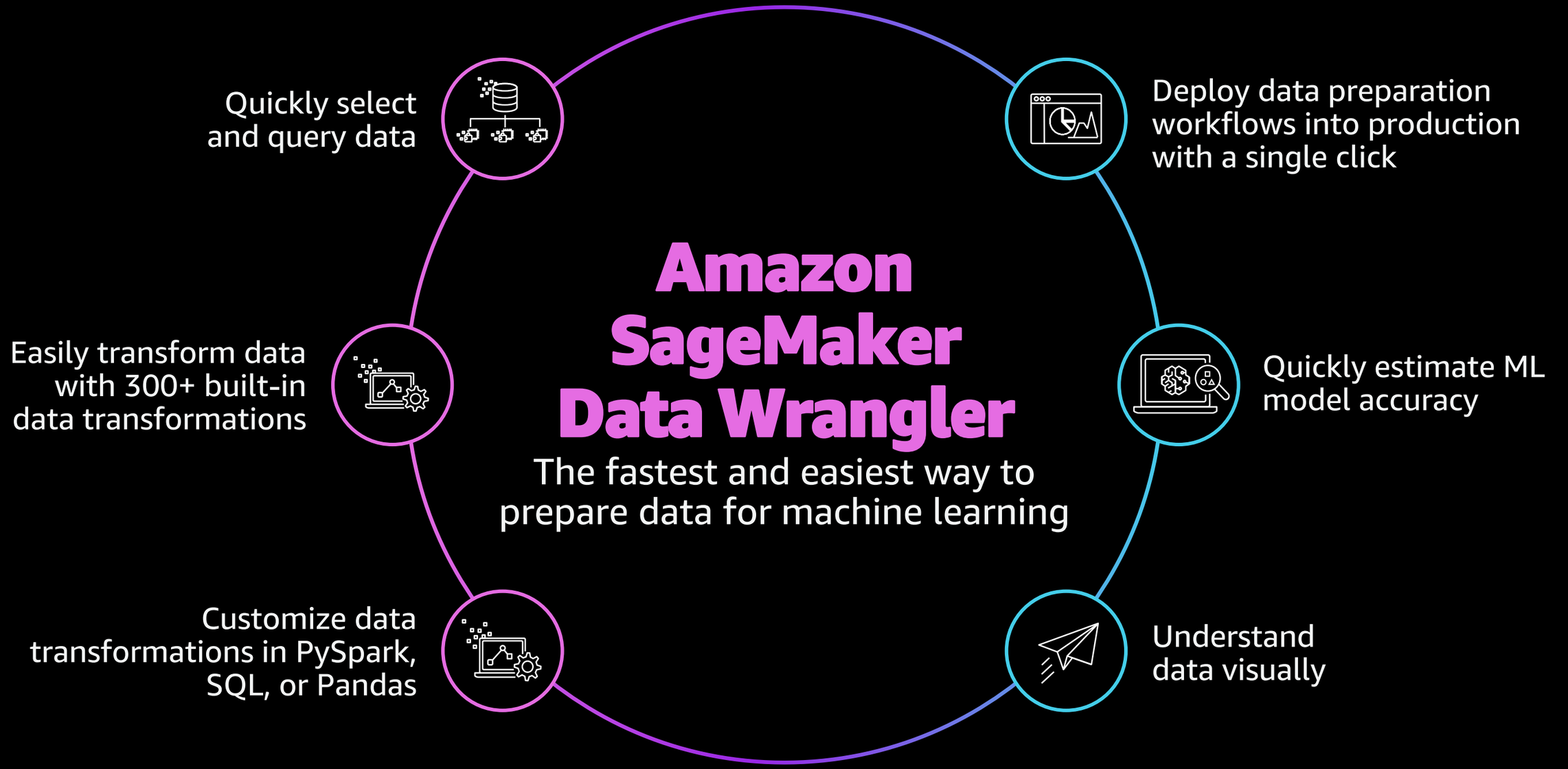
Data preparation is a critical part of the end-to-end ML workflow



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

Introducing Amazon SageMaker Data Wrangler



Quickly select and query data

SELECT

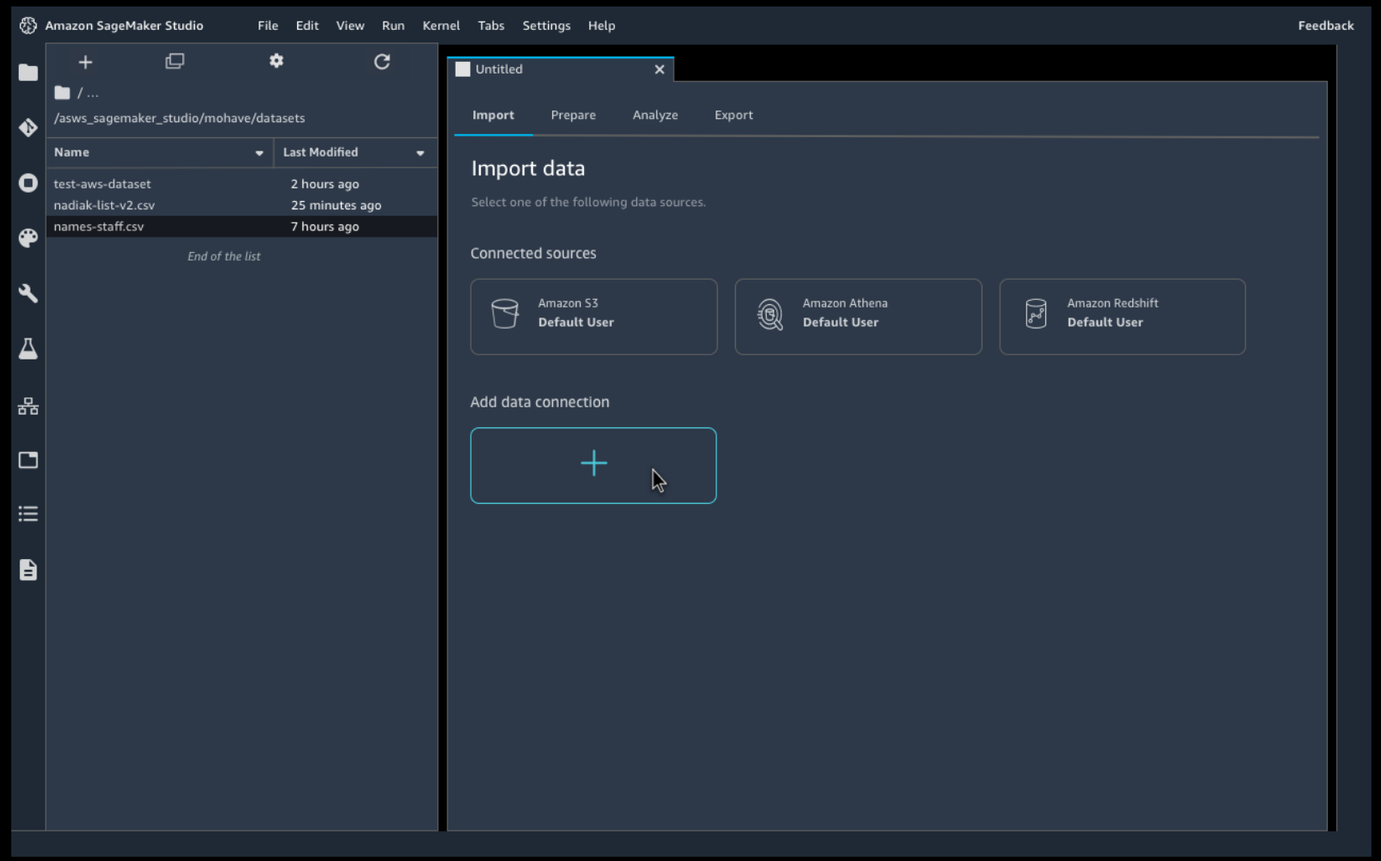
Quickly select data from Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon S3, and features from SageMaker Feature Store

WRITE

Write queries for data sources before importing data over to 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

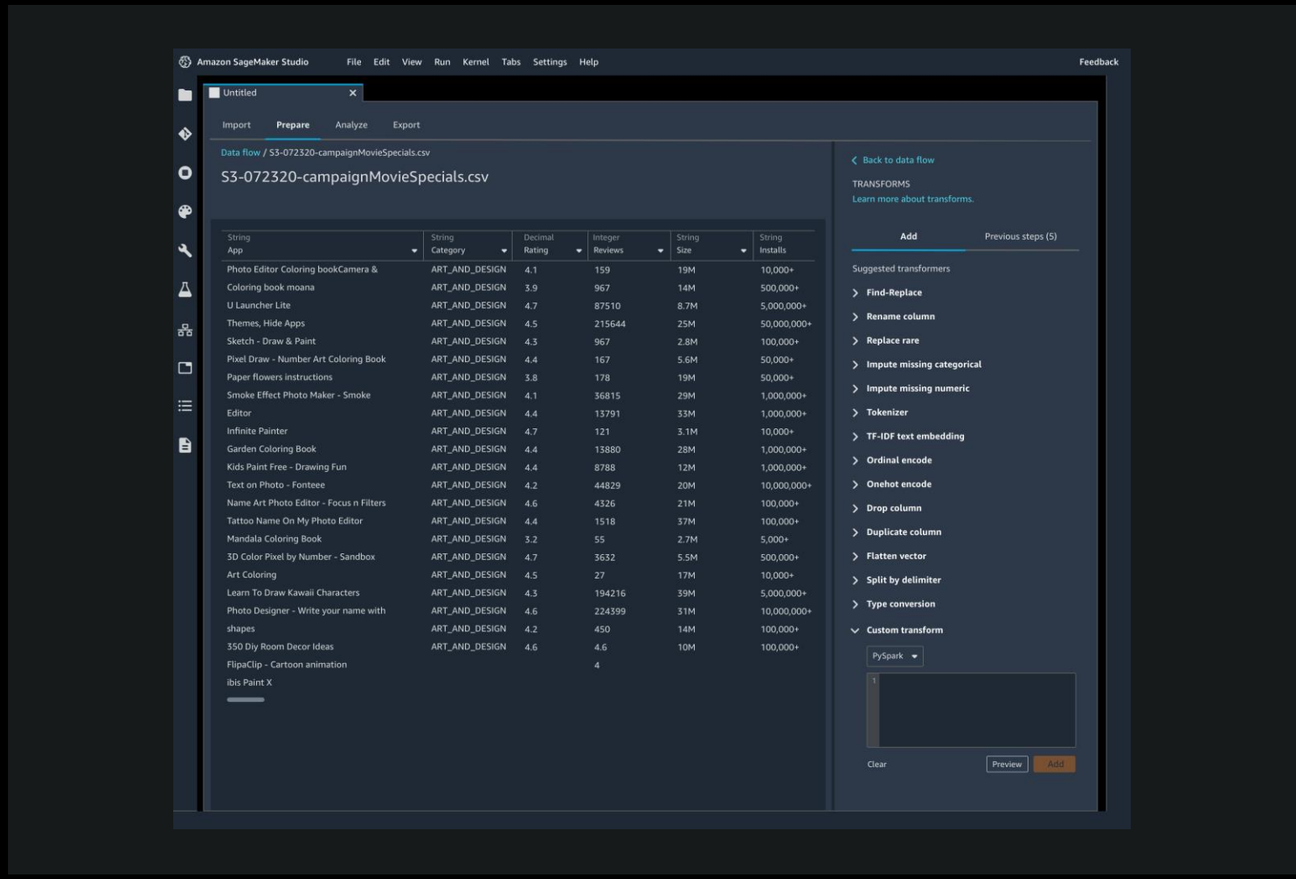
PREPARE

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



Understand your data visually

UNDERSTAND

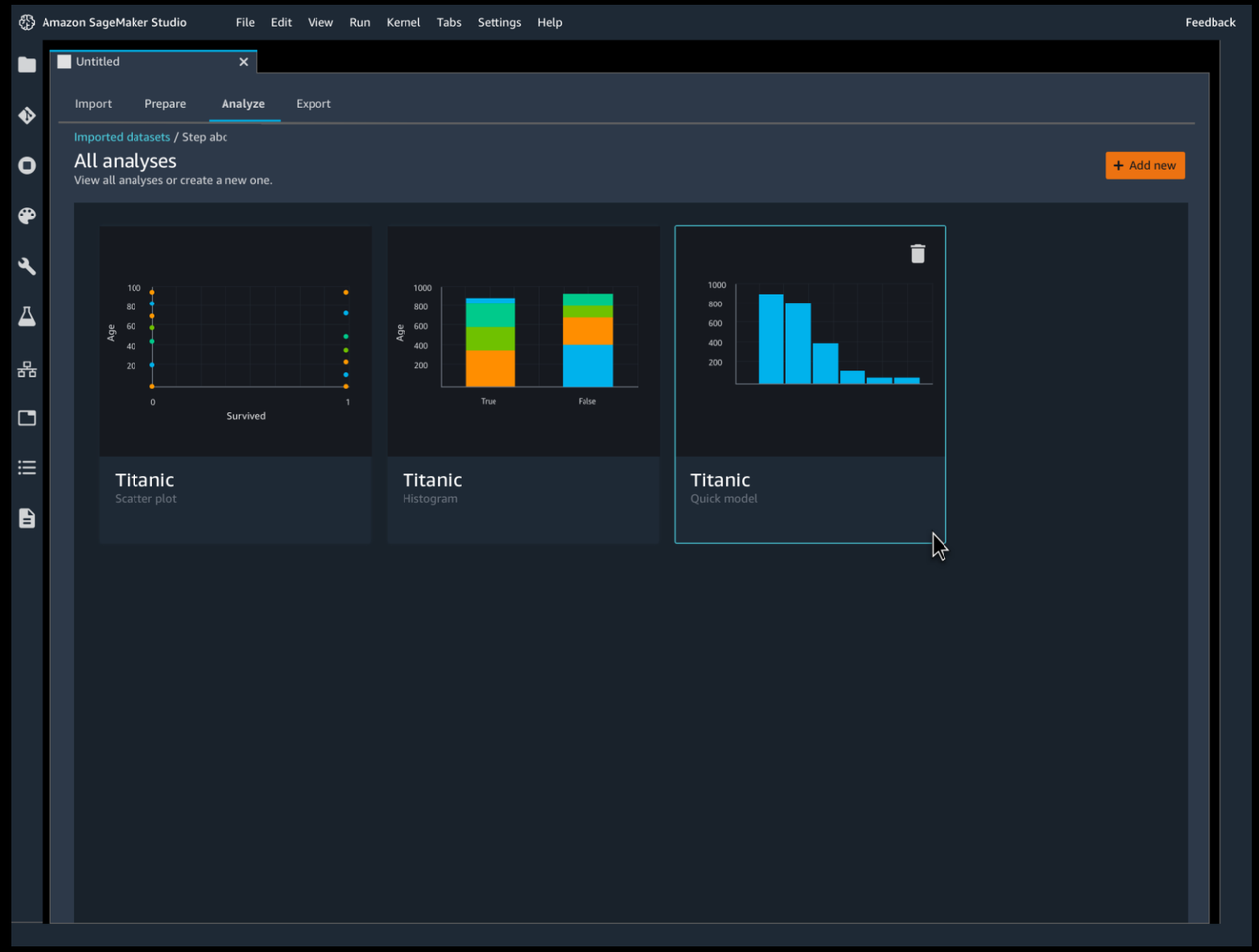
Intuitively understand your data with a set of preconfigured visualization templates which include histograms and scatter plots

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

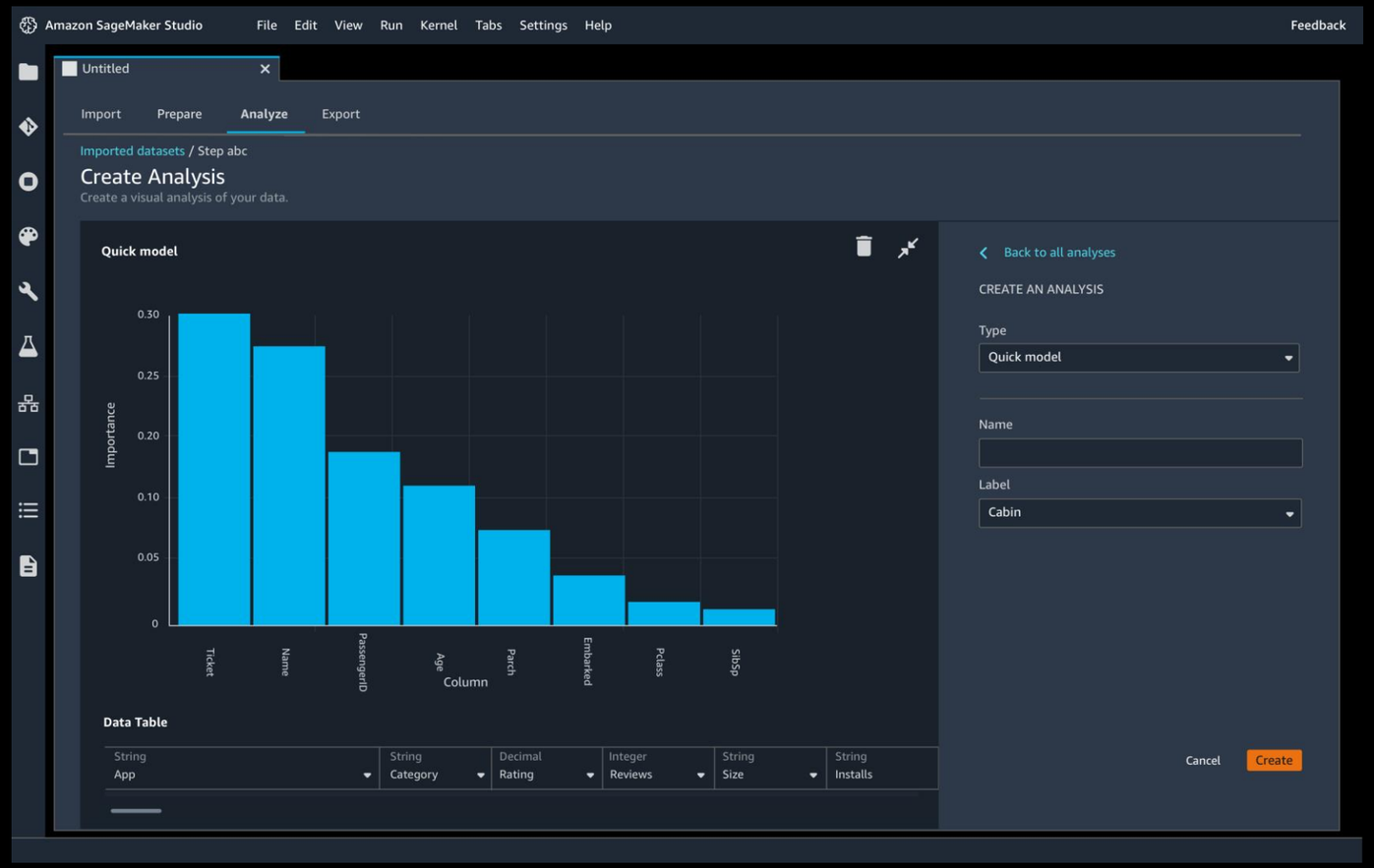
Quickly identify inconsistencies in your data preparation workflow and diagnose issues before ML models are deployed into production

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

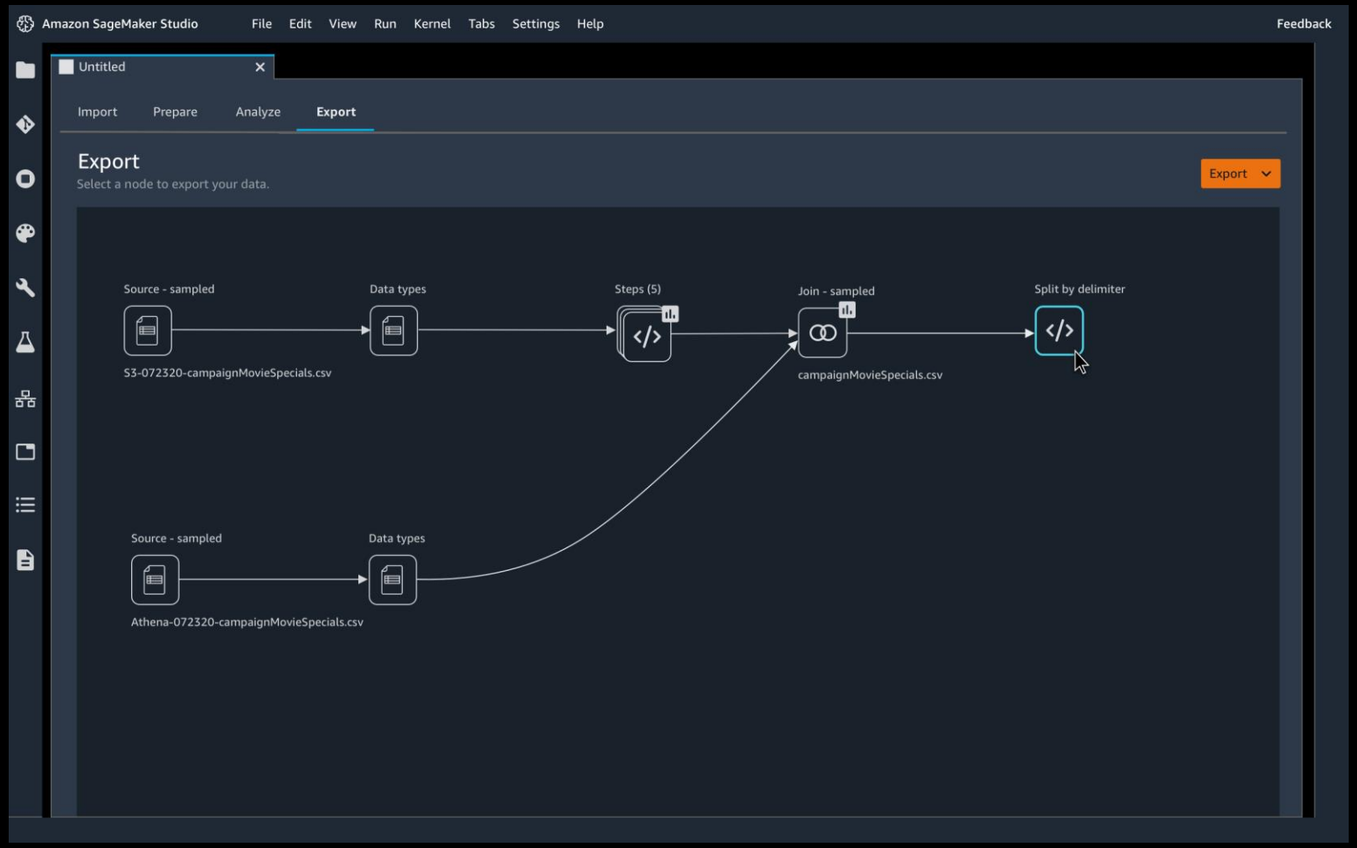
Export data preparation workflow as a processing job notebook or Python code with a single click

INTEGRATE

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

PUBLISH

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



Demo

In today's demo we'll get to:

- Import data from Amazon S3, Amazon Redshift, and Amazon Athena
- Visualize our data and check it for missing data and outliers
- Clean our data of possible inconsistencies
- Join and aggregate multiple data sources together
- Extract features from our data using SageMaker Data Wrangler's built-in feature extraction capabilities
- Estimate performance of a model on our dataset
- Export a production-ready data preparation workflow

...all using SageMaker Data Wrangler!

Partnership and customer spotlight



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker Data Wrangler

Partnerships

The following data sources will be available from within SageMaker Data Wrangler in 2021







Amazon SageMaker Data Wrangler

Customers

Customers using SageMaker Data Wrangler today to accelerate data preparation include:

:DeNA

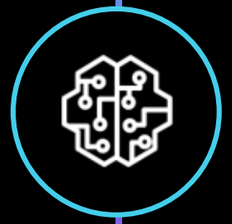
INVISTA™

Deloitte.

3M Science. Applied to Life.

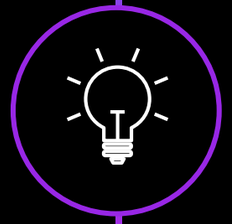
Nomura Research Institute (NRI)

Get started preparing your data for machine learning



Get started in 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

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.