

Dominando o Amazon SageMaker: Coleta de Dados e Preparação com o Amazon SageMaker

Carolina Junqueira Ferreira

Startup Solutions Architect
AWS



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Desafios de dados em ML
- Amazon Sagemaker
- Amazon Sagemaker Canvas
- Amazon Sagemaker Studio Notebooks

Tour de recursos do Amazon SageMaker

PREPARE DADOS E CRIE, TREINE E IMPLANTE MODELOS DE ML PARA QUALQUER CASO DE USO

PREPARAR

Geoespacial

Visualize dados geoespaciais

Ground Truth

Crie conjuntos de dados de alta qualidade para ML

Data Wrangler

Agregue e prepare dados para ML

Spark integrado

Integração integrada com o Amazon EMR e o AWS Glue

Processing

Python embutido, BYO R/Spark

Feature Store

Armazene, catalogue, pesquise e reutilize features

Clarify

Detecte viés e entenda as previsões do modelo

CONSTRUIR

Notebooks Studio e instâncias de notebooks

Notebooks Jupyter totalmente gerenciados com computação elástica

Editor de código

Editor de código baseado no VS Code Open Source (Code-OSS)

RStudio

FRStudio totalmente gerenciado

Studio Lab

Ambiente de desenvolvimento de ML gratuito

Algoritmos integrados

Algoritmos tabulares, de PNL e de visão integrados

Jump Start

Descoberta, treinamento e implantação de modelos, soluções e exemplos baseados em UI

Autopilot

Crie automaticamente modelos de ML com visibilidade total

Traga o seu

Traga seu próprio contêiner e algoritmos

Modo local

Teste e crie protótipos em sua máquina local

TREINAR E SINTONIZAR

Treinamento totalmente gerenciado

Amplas opções de hardware, fáceis de configurar e escalar

Bibliotecas de treinamento distribuídas

Treinamento de alto desempenho para grandes conjuntos de dados e modelos

Compilador de treinamento

Treinamento mais rápido em modelos de aprendizado profundo

Ajuste automático do modelo

Otimização de hiperparâmetros

Treinamento local gerenciado

Reduza o custo de treinamento em até 90%

Depurador e Profiler

O treinamento de depuração e perfil é executado

Experiments

Rastreie, visualize e compartilhe artefatos do modelo entre equipes

Suporte de personalização

Integre-se com estruturas populares de código aberto e bibliotecas

IMPLANTAR E GERENCIAR

Implantação totalmente gerenciada

Latência ultrabaixa e inferência de alto rendimento

Inferência em tempo real

Para padrões de tráfego estáveis

Inferência sem servidor

Para padrões de tráfego intermitente

Inferência assíncrona

Para cargas úteis grandes para longos tempos de processamento

Transformação em lote

Fou inferência off-line em lotes de grandes conjuntos de dados

Endpoints de vários modelos

Reduza os custos hospedando vários modelos por instância

Endpoints de vários contêineres

Reduza os custos hospedando vários contêineres por instância

Teste de sombra

Valide o desempenho do modelo na produção

Recomendador de inferência

USelecione automaticamente uma instância de computador e configuração

Model monitor

Mantenha a precisão dos modelos implantados

Operadores e componentes do Kubernetes

Gerencie e monitore modelos em dispositivos periféricos

Gerente de borda

Gerencie e monitore modelos em dispositivos periféricos

MLOPs: Pipelines|Project| Model Registry

Automação do fluxo de trabalho, CI/CD para ML, catálogo central de modelos

Amazon CodeWhisperer

Acelere a construção com Sugestões de código baseadas em IA

Canvas

Gere previsões precisas de aprendizado de máquina, sem necessidade de código

Studio

Plataforma com vários ambientes de desenvolvimento integrados (IDE) para ML e ciência de dados

Governança

Cartões modelo | Painel | Permissões

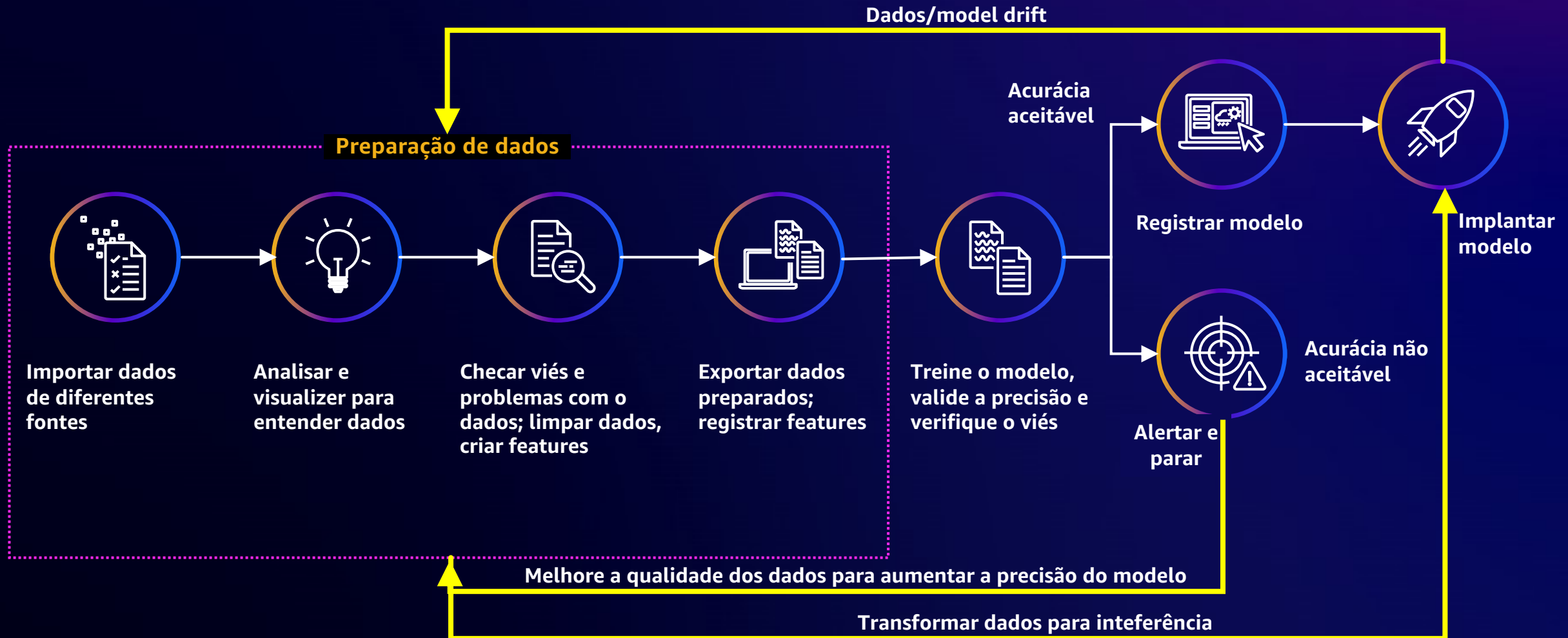


Amazon SageMaker ajuda as organizações a aproveitar o ML



A preparação de dados ocupa de 60% a 80% do tempo do projeto de ML

A TAREFA MAIS CRÍTICA E DEMORADA



Superando as barreiras do ML



Construtores de ML
insuficientes



Ferramentas de ML no-code

Faça previsões de ML independentemente da experiência
em ML



Processamento e rotulagem de
grandes volumes de dados
para ML



Ferramentas de preparação de dados específicas

Rotule e prepare dados para ML



Ferramentas diferentes de
ciência de dados



Ferramentas de ML integradas em uma única interface

Crie, treine e implante modelos usando IDEs



Operações manuais de ML



Recursos integrados de MLOps

Implemente práticas de MLOps para agilizar o ciclo de
vida do ML

O Amazon SageMaker traz ferramentas para cada etapa do ciclo de vida do ML em uma interface de usuário visual unificada



Amazon Sagemaker Canvas



A maneira mais rápida e fácil de preparar dados para ML no SageMaker Canvas



Preparação de dados ponta a ponta

Importação de dados, análise, limpeza, engenharia de recursos e produção



No code/low code

Análises visuais de dados, mais de 300 transformações integradas, biblioteca de códigos personalizados



Fácil de escalar

Apoiado por trabalhos de processamento Apache Spark sem servidor



Fácil de produzir

Treine diretamente no Canvas, agende trabalhos e automatize fluxos de trabalho de preparação de dados com SageMaker Pipelines

Como funciona a preparação de dados no Canvas



Amazon SageMaker Canvas

A maneira mais fácil e rápida de preparar visualmente os dados para o aprendizado de máquina



Importe dados

de mais de 50 fontes de dados usando interface visual ou SQL



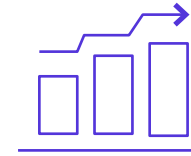
Analise

Explore e analise dados com visualizações e análises integradas



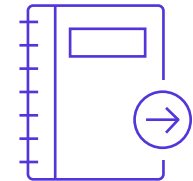
Limpe e enriqueça

Execute a engenharia de recursos com mais de 300 transformações de dados integradas



Melhore a precisão

Estime o desempenho e a precisão do modelo e diagnostique possíveis problemas



Treine e automatize

Treine o modelo com o AutoML ou integre-o aos componentes do SageMaker

Produção interativa

Interface unificada para agregar dados

The screenshot displays the AWS Glue console's 'Import Data' interface. At the top, there's a search bar labeled 'Search data source' and a filter dropdown set to 'All (55)'. Below this, a list of data sources is shown, including 'Accident-detection-CCTV' and 'Accident'. The main section is titled 'Import Data' and features buttons for 'Upload', 'S3', 'Snowflake Crystal 1', and 'Redshift Crystal 1', along with an 'Add Connection' link. A search bar for connections is present, and a list of connections (database1, database2, database3, database4, schema1, schema2, table1) is shown. The central area displays a visual data flow diagram with 'table1.csv' connected to 'table2.csv'. The bottom section, 'Import preview', shows a table of data with columns: Sold, Price, Region, Discount, Fabric, and Age. The table contains 5 rows of data. At the bottom, there are 'Close' and 'Import data' buttons.

<input checked="" type="checkbox"/> Sold	ABC	<input type="checkbox"/> Price	123	<input checked="" type="checkbox"/> Region	ABC	<input checked="" type="checkbox"/> Discount	123	<input checked="" type="checkbox"/> Fabric	ABC	<input checked="" type="checkbox"/> Age	123
Yes		29.99		Southwest		23		Cotton		27	
Yes		29.99		Southwest		23		Silk		35	
Yes		29.99		Southwest		23		Silk		32	
Yes		29.99		Southwest		23		Silk		32	
Yes		29.99		Southwest		23		Cotton		30	

- Conecte-se a várias fontes de dados
- Amazon DocumentDB, Salesforce Data Cloud, Amazon S3, Amazon Athena, Amazon Redshift, Amazon RDS, Snowflake, and 50+ data sources
- **Multimodal**
 - CSV, Parquet, tabelas de banco de dados
 - JPG ou PNG
- Explore e selecione dados com o visual, constructor de consultas ou consultas SQL

Limpe dados e recursos de engenharia com facilidade

- Mais de 300 transformações apoiadas pelo Spark
- Transformações comuns e avançadas de ML
 - Séries temporais, divisão de trem/teste, redução de dimensionalidade e muito mais
 - Transformações de texto e imagem
- Flexibilidade e potência de trechos de código personalizados

The screenshot displays the Amazon Data Wrangler interface. The top navigation bar shows the path: Data Wrangler: Data flow > House-loan.flow > canvas-sample-loans-part-1.csv. The main panel is divided into 'Data' and 'Analyses' tabs. The 'Data' tab is active, showing a table with columns: id_0 (long), loan_status (string), and loan_amount (long). The table contains 12 rows of data. Above the table, there are three histograms for each column. The 'loan_status' column has a histogram showing 3 categories. The 'loan_amount' column has a histogram showing a range from 1000 to 35000. On the right side, a 'Custom transform' panel is open. It has a title bar with a back arrow, 'Custom transform', and a close button. Below the title bar, there is a text input field with the value 'Generate text embedding with Bedrock'. Below that, there is a dropdown menu with the value 'Python (PySpark)'. At the bottom of the panel, there is a section titled 'Example code snippet' with a search bar and a list of snippets: 'Generate text embedding with Bedrock', 'Query Bedrock model', 'Chunk text (characters)', 'Chunk text (sentence)', 'Prompt engineering with the context', 'Mask PII', 'Extract txt from PDF', and 'Uniform sampling'.

Data Wrangler: Data flow > House-loan.flow > canvas-sample-loans-part-1.csv

Data Analyses

Data Wrangler: Data flow > House-loan.flow > canvas-sample-loans-part-1.csv

Step 1. Join

id_0 (long)	loan_status (string)	loan_amount (long)
1077501	fully paid	5000
1077430	charged off	2500
1077175	fully paid	2400
1076863	fully paid	10000
1075358	current	3000
1075269	fully paid	5000
1069639	fully paid	7000
1072053	fully paid	3000
1071795	charged off	5600
1071570	charged off	5375
1070078	fully paid	6500

Custom transform

Use Pyspark, Pandas, or Pyspark (SQL) to define custom transformations. [Learn more.](#)

Name

Generate text embedding with Bedrock

Python (PySpark)

Example code snippet

Search for an example snippet

Select...

- Generate text embedding with Bedrock
- Query Bedrock model
- Chunk text (characters)
- Chunk text (sentence)
- Prompt engineering with the context
- Mask PII
- Extract txt from PDF
- Uniform sampling

Demo – Amazon Sagemaker Canvas



Amazon Sagemaker Studio Notebooks



Notebooks SageMaker Studio

TOTALMENTE GERENCIADO
Jupyter Lab



Configuração com um clique

Configure identidade, acesso, controle, rede e armazenamento no clique



Prepare dados em grande escala

Crie, navegue e conecte-se aos clusters do Amazon EMR; Crie, teste e execute aplicativos interativos de preparação e análise de dados com sessões interativas do AWS Glue



Notebooks de início rápido

Lançamento totalmente gerenciado Jupyter Lab em segundos; redimensione o computação e armazenamento em qualquer lugar



SageMaker Distribution

Imagens pré-configuradas que incluem estruturas populares para aprendizado de máquina, ciência de dados e visualização



Personalizável

Traga seu próprio ambiente de notebook para o SageMaker Studio usando uma imagem personalizada do Docker

Use o Amazon SageMaker Studio para atualizar modelos e ver imediatamente o impacto na qualidade dos modelos

The screenshot displays the Amazon SageMaker Studio interface. The main window shows a Jupyter notebook titled 'xgboost_customer_churn.ipynb' with the following content:

- Have the predictor variable in the first column
- Not have a header row

But first, let's convert our categorical features into numeric features.

```
[ ]: model_data = pd.get_dummies(churn)
model_data = pd.concat([model_data['Churn?_True.'], model_data.drop(['Churn?_True.'], axis=1)], axis=1)
```

And now let's split the data into training, validation, and test sets. This will help prevent us from overfitting the model, and allow us to test the models accuracy on data it hasn't already seen.

```
[ ]: train_data, validation_data, test_data = np.split(model_data.sample(frac=1, random_state=123), [int(0.33 * len(model_data)), int(0.33 * len(model_data))])
train_data.to_csv('train.csv', header=False, index=False)
validation_data.to_csv('validation.csv', header=False, index=False)
```

Now we'll upload these files to S3.

```
[ ]: boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'train.csv')).upload_file(train_data.to_csv(index=False).getvalue(), bucket, os.path.join(prefix, 'train.csv'))
boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'validation.csv')).upload_file(validation_data.to_csv(index=False).getvalue(), bucket, os.path.join(prefix, 'validation.csv'))
```

On the right side, there are two panels:

- Trial Component Chart:** A line chart showing 'trainloss_last' (y-axis, 0.0 to 0.4) versus 'period' (x-axis, 0 to 6). The chart displays four lines (blue, green, orange, red) representing different trials, showing a general downward trend in loss over time.
- Trial Component List:** A table listing trial components. It shows 10 rows selected. The table has columns: Status, Experiment, Type, Trial, and Trial component. All four rows shown are 'Completed' training jobs for 'customer-churn-predi...'.

The bottom status bar indicates 'Mode: Command', 'Ln 1, Col 1', and the file 'xgboost_customer_churn.ipynb'.

Preparação de dados em notebooks Studio



Demo - Amazon Sagemaker Studio





Amazon Sagemaker Feature Store

Desafios de separar features stores



Feature drift



Feature duplicados



**Desenvolvimento/implantação
lento de modelo**

Amazon SageMaker Feature Store

Armazene, descubra e compartilhe
com segurança recursos para ML



Online e off-line



Latência de milisegundo



Features consistentes



Busca visual

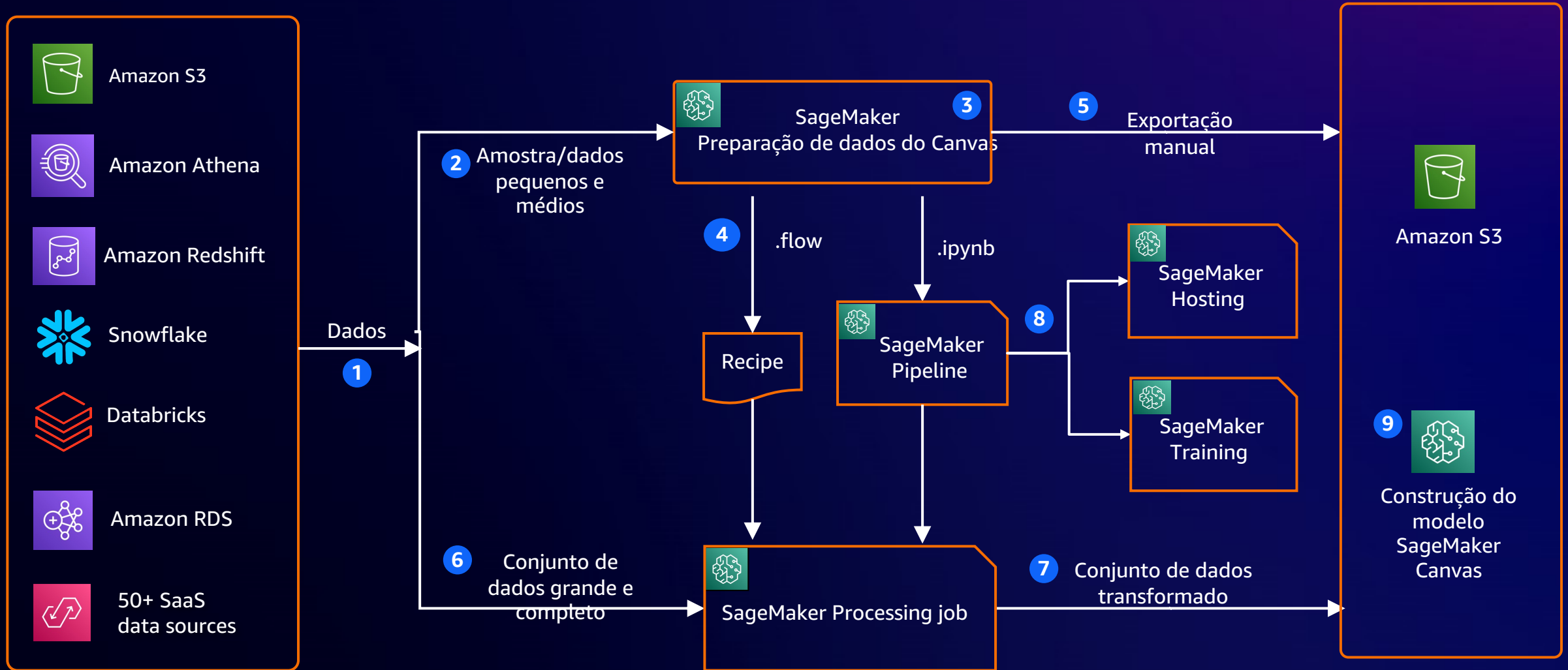


Compartilhar e colaborar

Resumo



Acelere a preparação de dados com o SageMaker Canvas



Perguntas?



Interessado em encurtar seu ciclo de inovação?



Visão geral do SageMaker Canvas



Workshop SageMaker Canvas

Obrigada!

Carolina Junqueira Ferreira

caroljf@amazon.com

