



# AWS LATAM Black Belt 2023

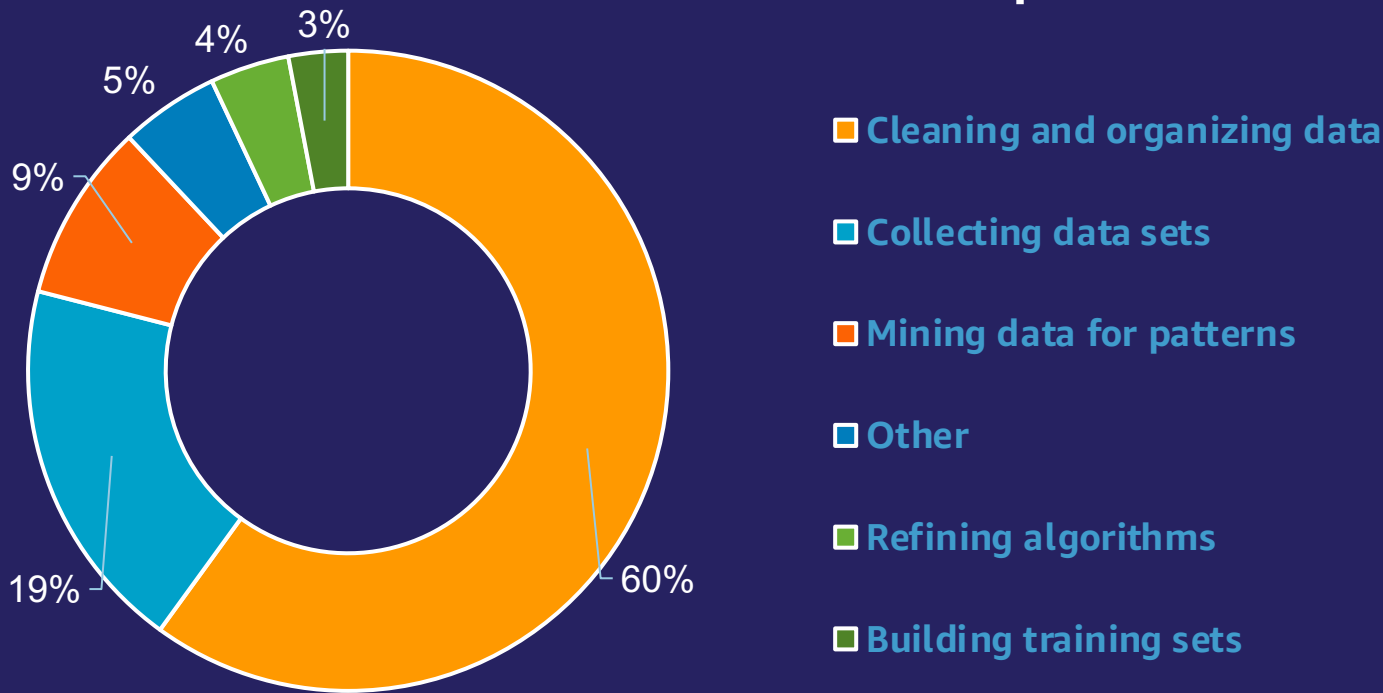


## SageMaker DataWrangler

Matheus Oliveira

# 80% of time spent on data prep

What data scientists spend the most time doing



# Amazon SageMaker Data Wrangler

EXPLORE, PREPARE, AND PROCESS  
DATA WITH LITTLE TO NO CODE



Import data from multiple sources



Get insights on data and data quality



Visually explore, analyze, and prepare data



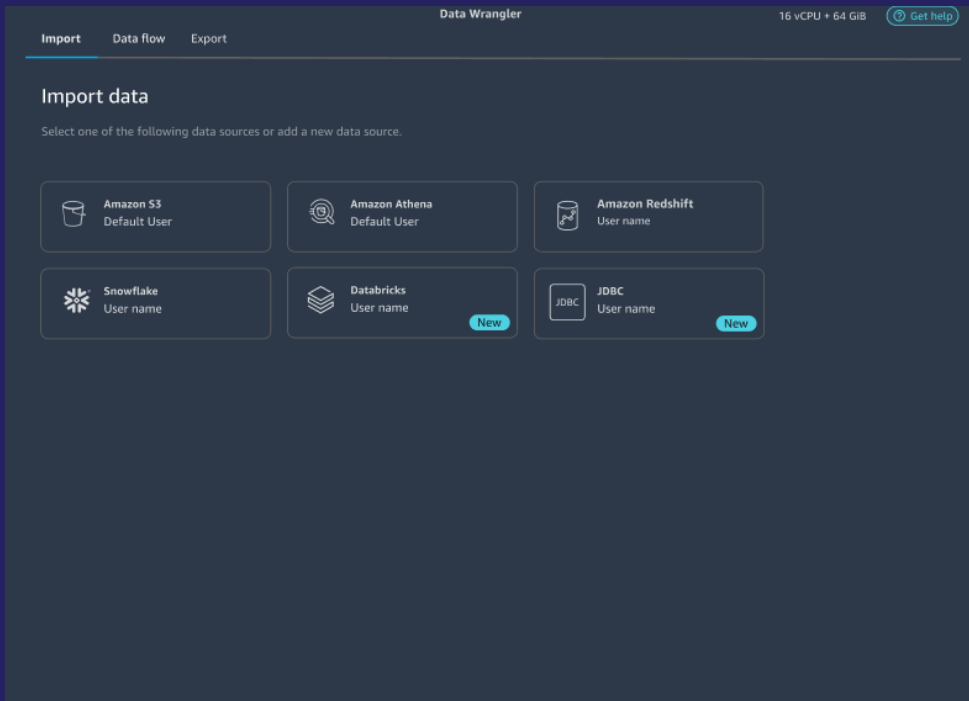
Quickly perform feature engineering



Automate ML data preparation workflows



# Quickly connect and query data



## Connect to multiple data sources:

Amazon S3

Amazon Athena

Amazon Redshift

Snowflake

Databricks DeltaLake (NEW)

SageMaker Feature Store & more coming

## Support for common file formats:

CSV files

Parquet files

JSON & JSONL (NEW)

ORC file (NEW)

Database tables



# Quickly connect and query data

[Back to Import](#)

## Import a dataset from S3

Enter the S3 URL of a file or prefix (folder) in the text box, or use the following table to browse S3

[S3 / sagemaker-us-east-2-562522975874 / airports](#)

Object name	Size	Last modified
airports.csv	209.09KB	2020-11-05 17:17:30+00:00

[Previous](#) [Next](#)

Displaying 1 - 1

PREVIEW • airports.csv (first 100 rows shown)

lata	airport	city	state	country	lat	long
00M	Thigpen	Bay Springs	MS	USA	31.95376472	-89.23450472
00R	Livingston Municipal	Livingston	TX	USA	30.68586111	-95.01792778
00V	Meadow Lake	Colorado Springs	CO	USA	38.94574889	-104.5698933
01G	Perry-Warsaw	Perry	NY	USA	42.74134667	-78.05208056
01J	Hilliard Airpark	Hilliard	FL	USA	30.6880125	-81.90594389
01M	Tishomingo County	Belmont	MS	USA	34.49166667	-88.20111111
02A	Gragg-Wade	Clanton	AL	USA	32.85048667	-86.61145333
02C	Capitol	Brookfield	WI	USA	43.08751	-88.17786917
02G	Columbiana County	East Liverpool	OH	USA	40.67331278	-80.64140639

DETAILS [Get help](#)

Name [Preview on](#)

airports.csv

File type

csv

parquet

json

jsonl

orc

Visually browse data sources like **objects on S3**, or **database, schema, tables and objects in Snowflake**

**Preview & sample top rows**

Join data from multiple sources

Support for VPC, KMS, CMK, and AWS Secrets Manager



# Easily transform data for ML with 300+ built-in transforms

300+ built-in data transformations (no code) for common data prep needs and ML specific needs

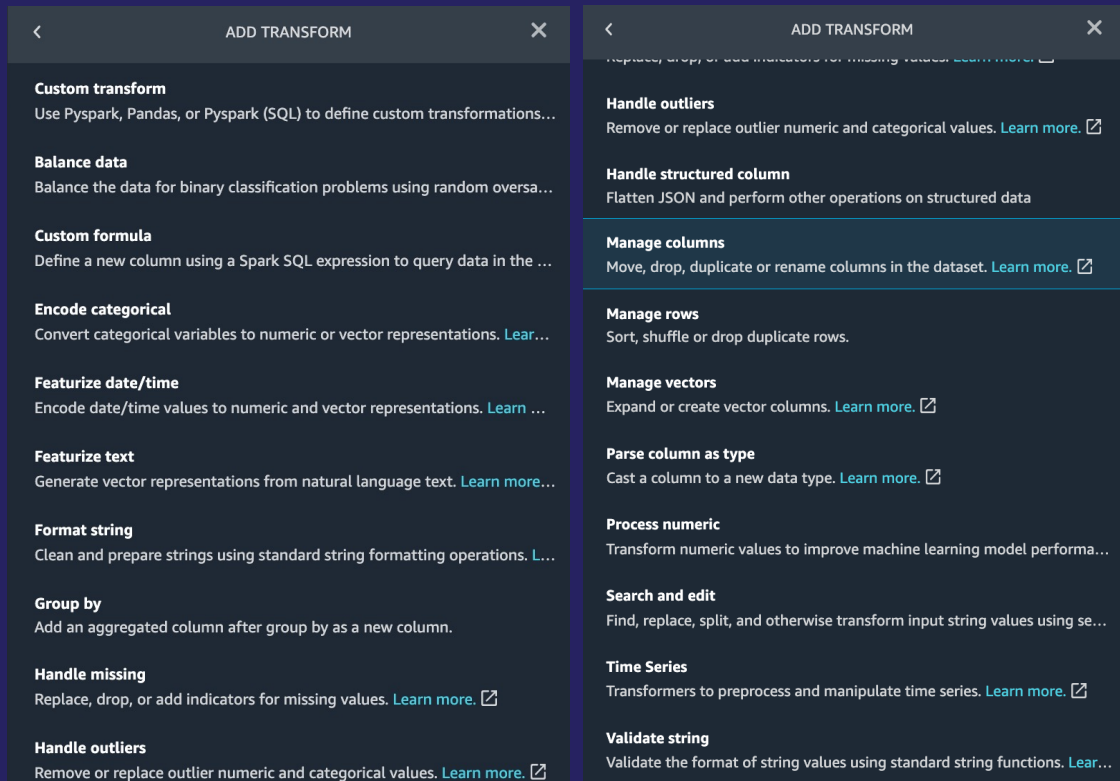
Built by data scientists for data scientists

**ML specific transforms such as:**

One hot encoding

Balance data

Time series transforms

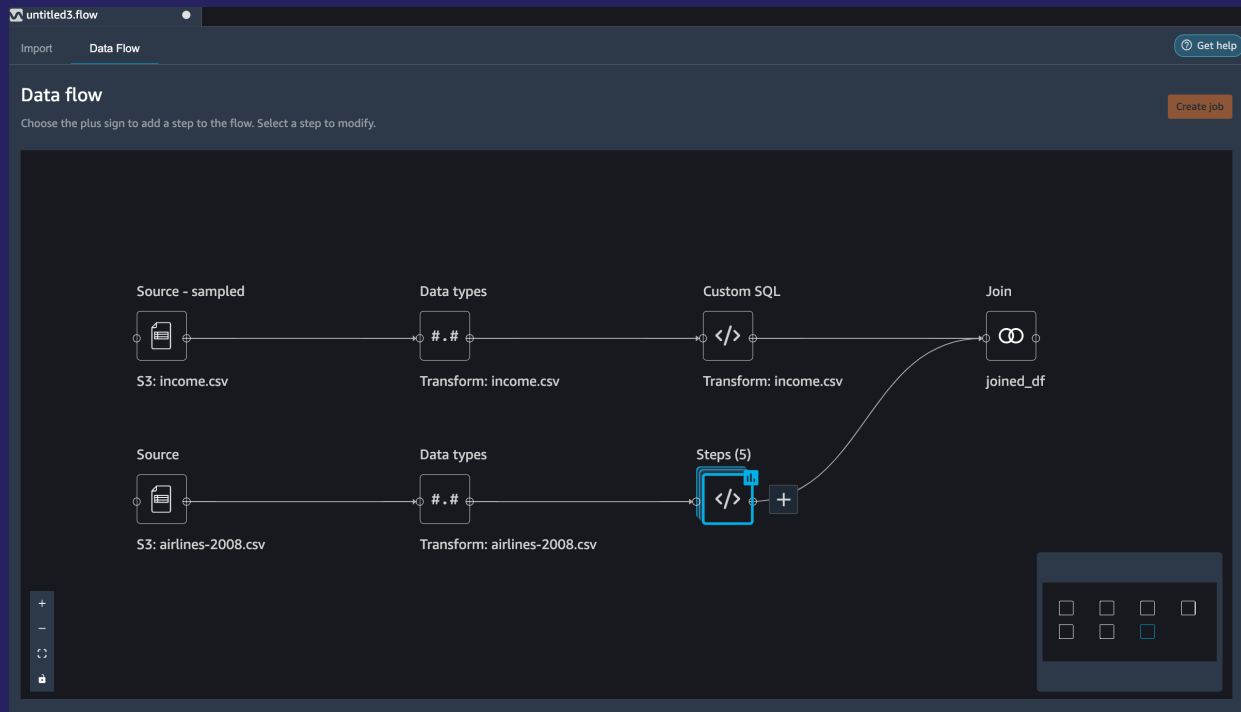


# Easily visualize the steps of your data processing pipeline

Data Wrangler records all the steps of data prep workflow in a data flow graph

Visualize the order of transformations, join and concatenate operators

Easily navigate data transformation flow, and modify and delete steps iteratively





# Sagemaker Data Wrangler Lab





# Questions?

Matheus Oliveira

kmsilvam@amazon.com