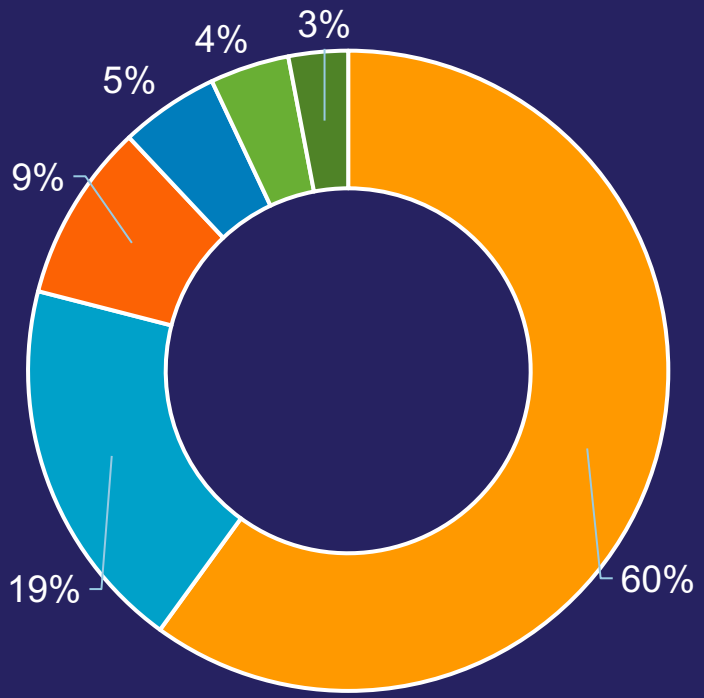
AWS LATAM Black Belt 2023

SageMaker DataWrangler

Matheus Oliveira

80% of time spent on data prep

What data scientists spend the most time doing

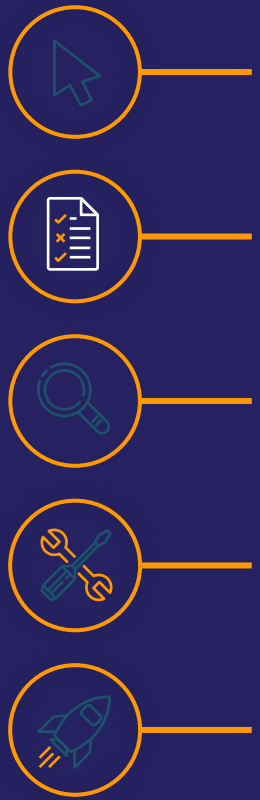


- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Other
- Refining algorithms
- Building training sets

Source: Forbes survey of 80 data scientists, March 2016

Amazon SageMaker Data Wrangler

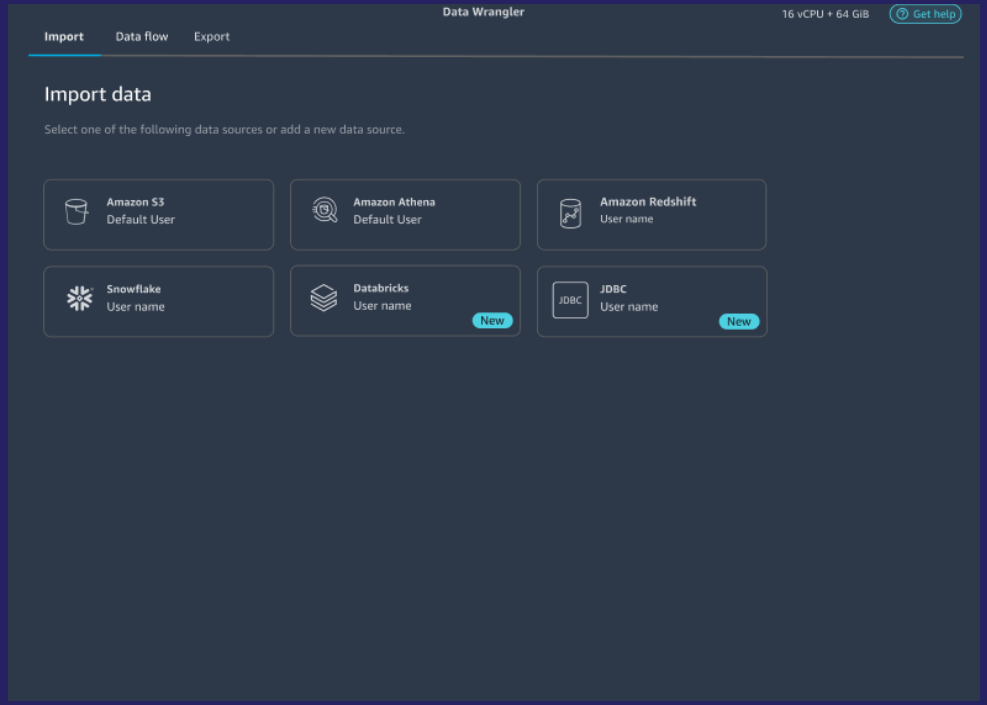
EXPLORE, PREPARE, AND PROCESS DATA WITH LITTLE TO NO CODE



- 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

aws Black Belt

Quickly connect and query data



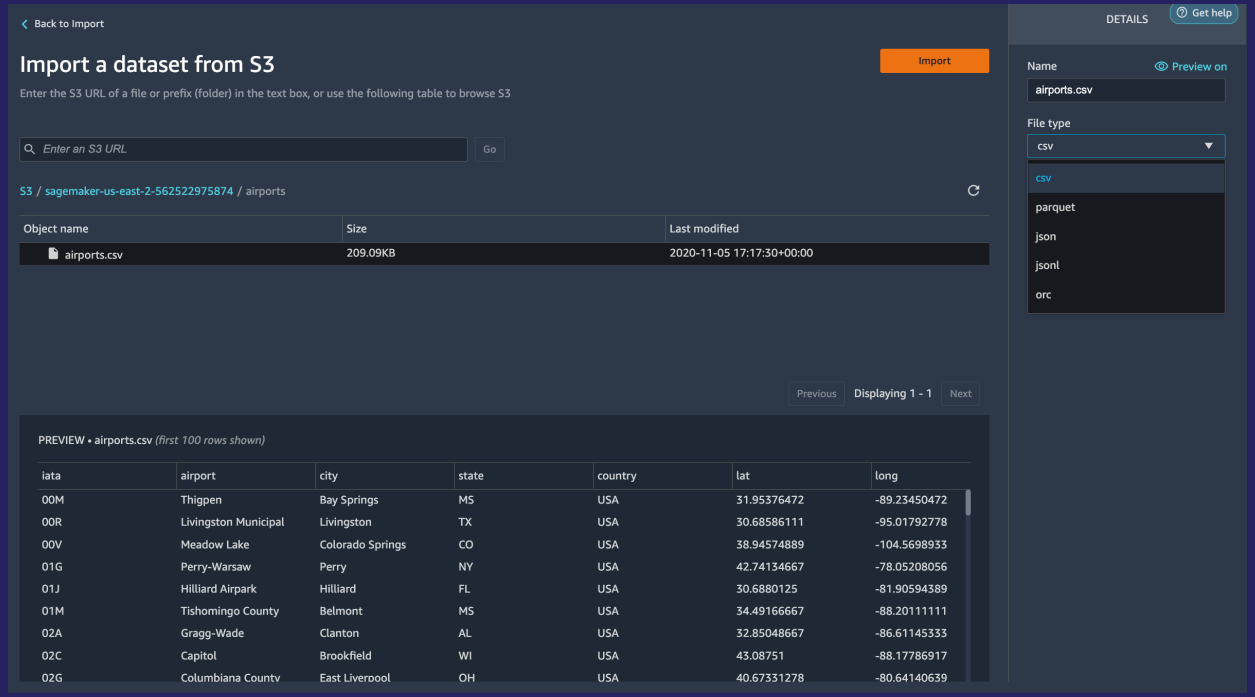
Connect to multiple data sources:

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



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

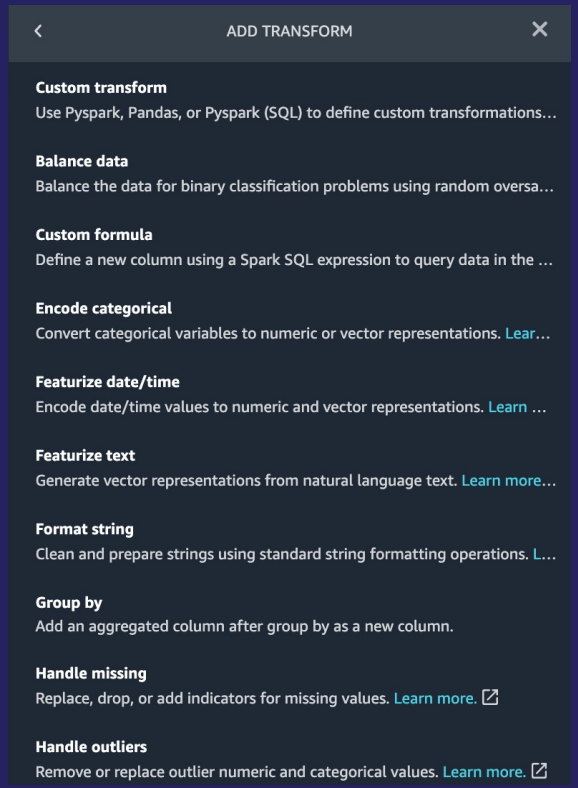
Support for VPC, KMS, CMK, and AWS Secrets Manager

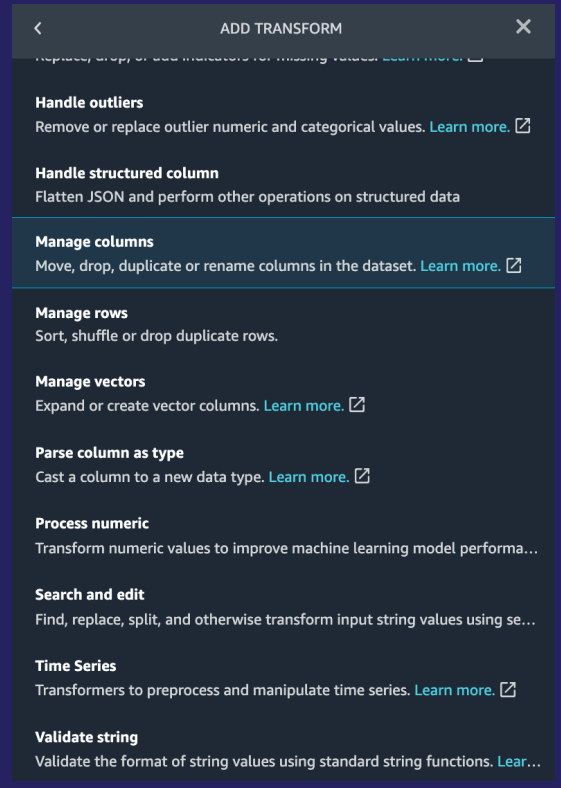
Easily transform data for ML with 300+ built-in transforms

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



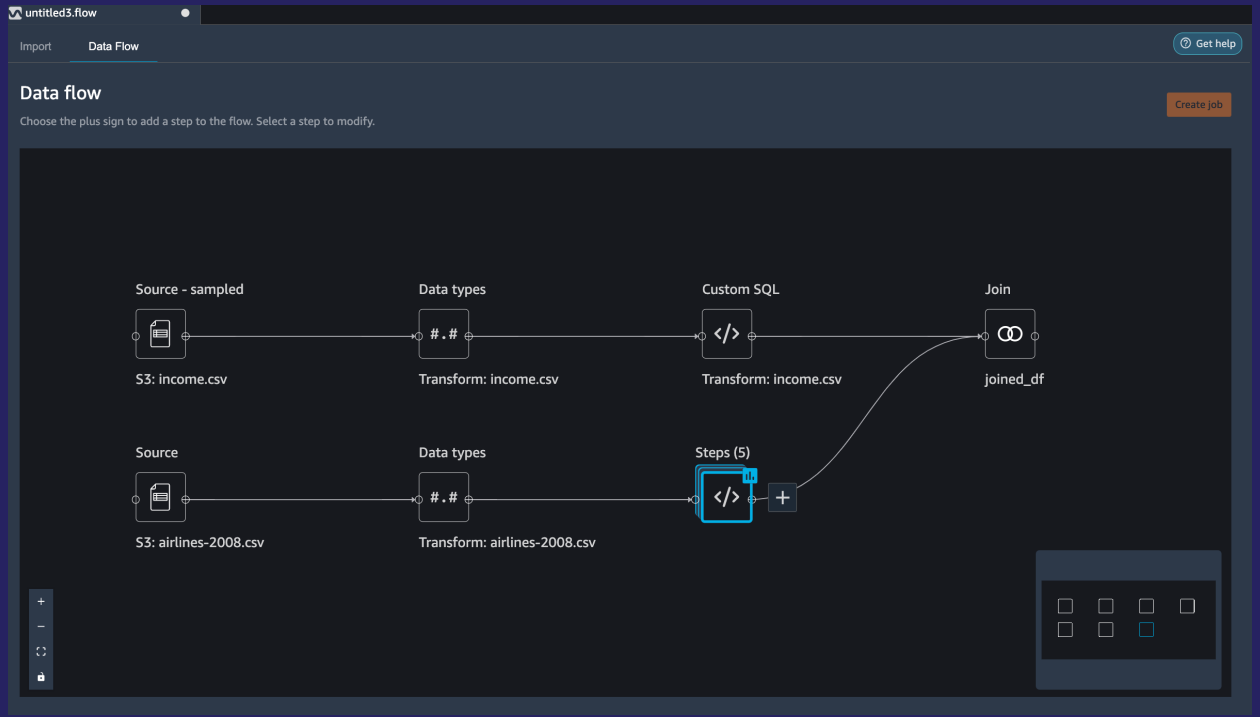


Easily visualize the steps of your data processing pipeline

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