AWS Glue DataBrew

Visual Data Preparation

Eren Akbaba

Agenda

- Data integration challenges and trends
- Introduction to AWS Glue DataBrew
- Popular use cases
- DataBrew highlighted features
- Pricing
- Q&A

aws

Trend: data is...

Growing Exponentially

From new sources

Increasingly diverse

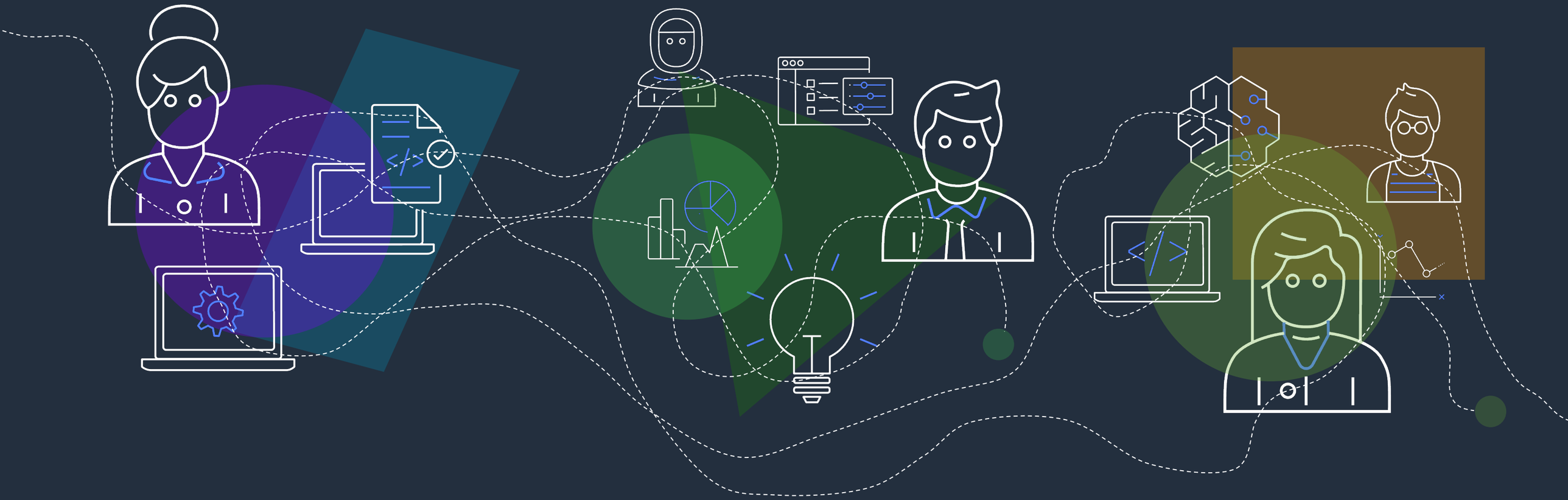
Used by people with diverse skillsets

Accessed by many applications

aws

As much as 80% of time is spent preparing data today

Data engineers | **ETL developers** | **Data analysts** | **Data scientists**



Needs the right tool for the right persona

aws

Companies spend as much as 80% time to prepare data today

Time-consuming

- Multi-step process to extract, clean, normalize, and load data at scale
- Needs right tools for the right persona that are integrated

Expensive

- Costly user licenses and siloed tools causing rework
- Often requires moving large amounts of data into silos, at times out of VPCs

Manual

- Hard to operationalize and build repeatable workflows
- Needs a lot of code-based heavy-lifting for it to work at scale

aws

The AWS analytics portfolio

Data, visualization, engagement, & machine learning

NEW

Data Exchange


QuickSight


Pinpoint


SageMaker


Comprehend


Lex


Polly


Rekognition


Translate

+ many more

Analytics


Redshift


EMR (Spark & Hadoop)


AWS Glue (Spark & Python)

NEW

AWS Glue DataBrew


Athena


Elasticsearch Service


Kinesis Data Analytics

Data lake infrastructure & management


S3/Glacier


Lake Formation

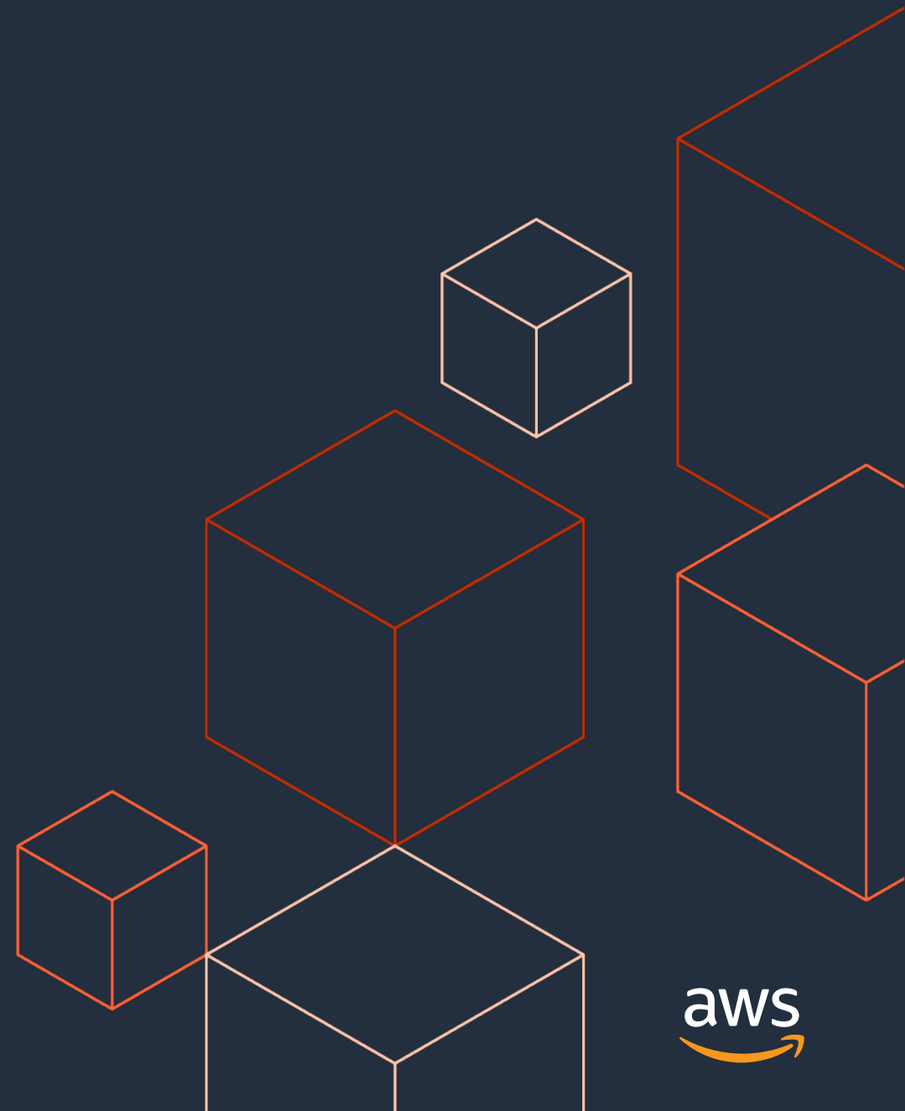

AWS Glue

Data movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Streams | Kinesis Data Firehose | Managed Streaming for Apache Kafka

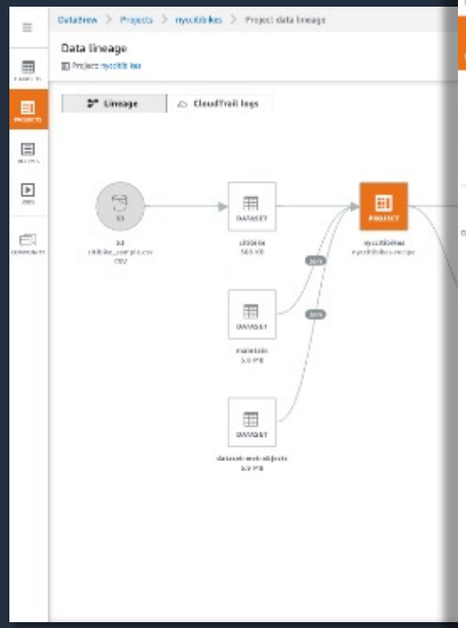
aws

AWS Glue DataBrew

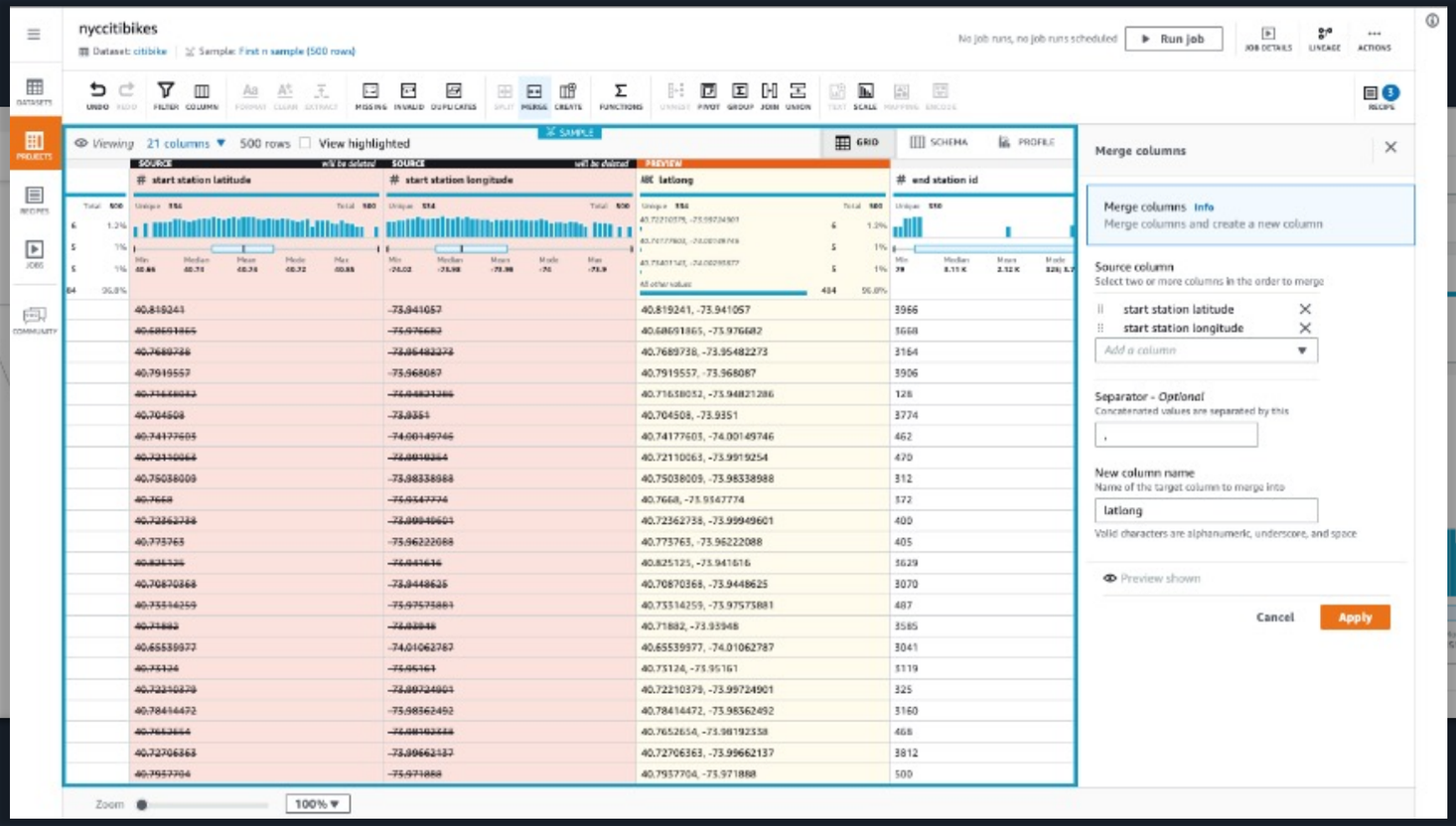


DataBrew – serverless, no code data preparation

for data analysts and data scientists



Visual lineage



Data quality dashboard

Easy access from the following interfaces:

AWS Management Console **Plugin for Jupyter Notebooks** **Plugin for SageMaker Studio**

Built for data analysts and data scientists

Understand data quality

Understand patterns and detect anomalies using profiles

Clean and normalize data

Over 250 built-in transformations

Visually map data lineage

Understand steps that the data has been through

Automate at scale

Save transformations and apply to new data as it comes in

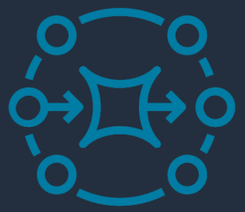
Data preparation made easy

aws

DataBrew Functionality

Connect to data sources

Amazon S3, Amazon Redshift, Amazon RDS Aurora or AWS Glue Data Catalog



250 + Built in Transformations

Join, tokenize, split, merge, extract, remove, group, pivot, normalize, one hot or label encode, and more

For users of all technical levels

Visually interact with your data and schema as you build a Recipe without writing any code

Advanced data profiling

Run a profile job in the background to get detailed statistics on your data

Visual Data lineage

View the various stages of data transformation from start to end

Integrate with data pipelines

SDK/API access to integrate with existing data pipelines

Serverless usage at scale

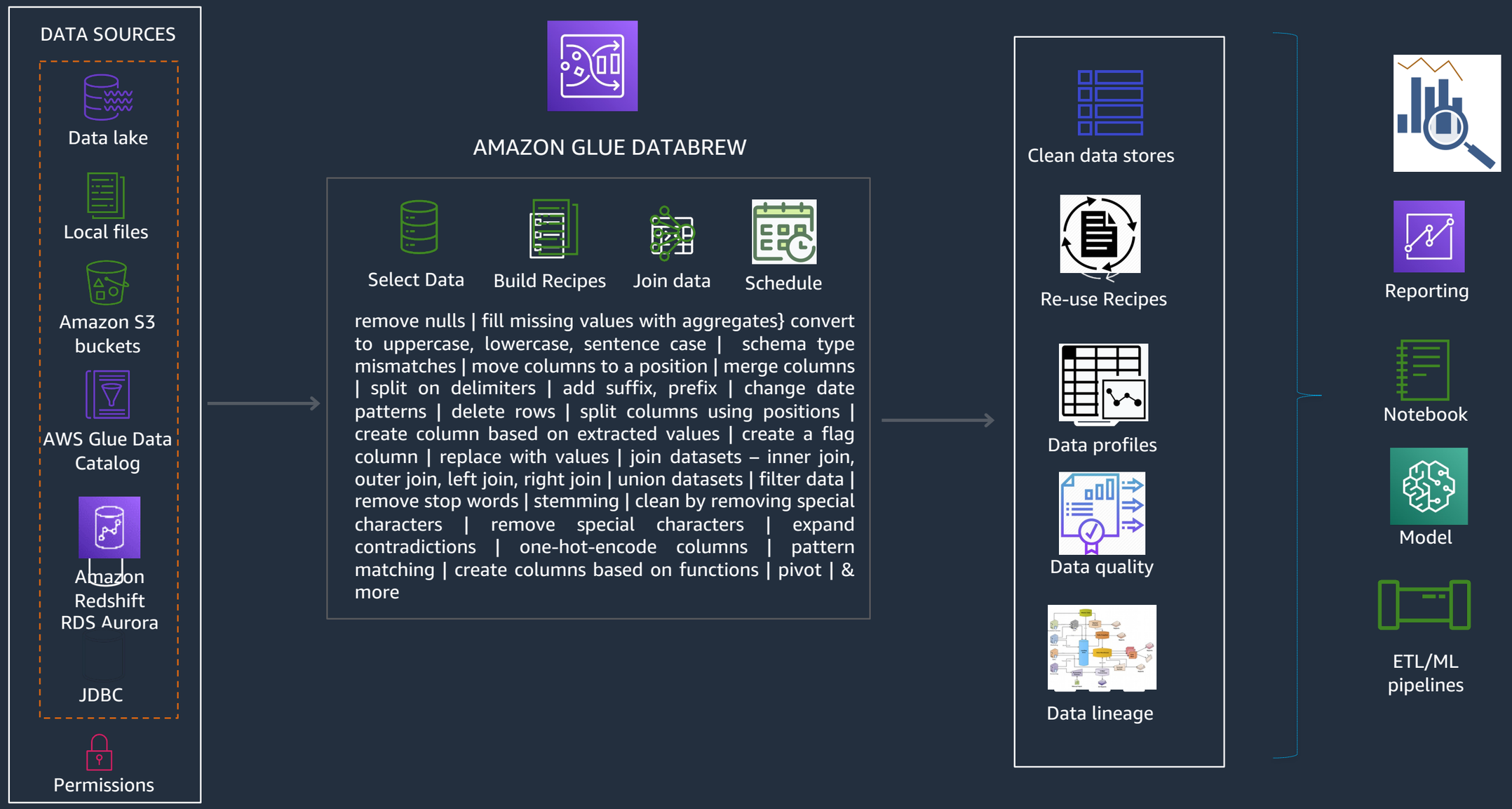
Operate at massive scale in a serverless capacity. Pay only for what you use

Advanced use cases

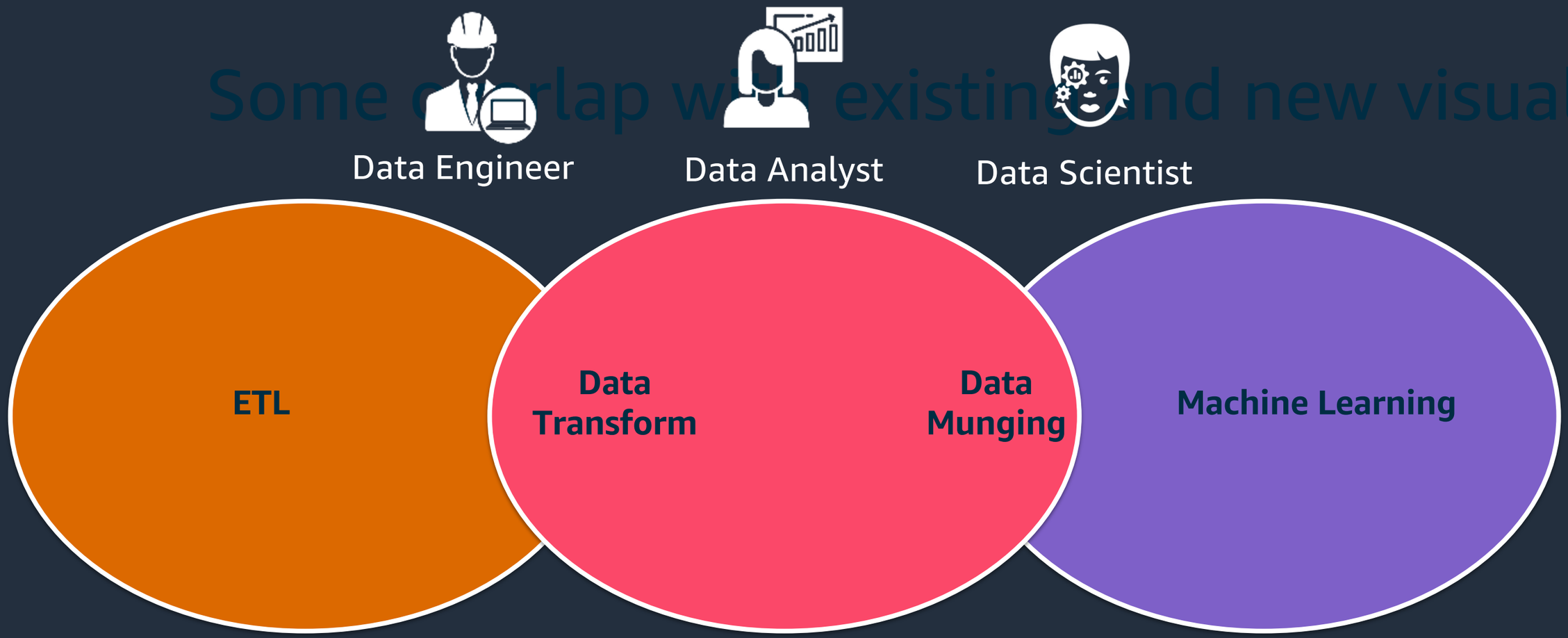
Use for traditional Data Preparation, Machine Learning, Analytics/Reporting and more

aws

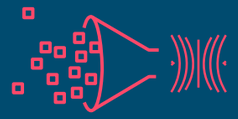
DataBrew: How does it work?



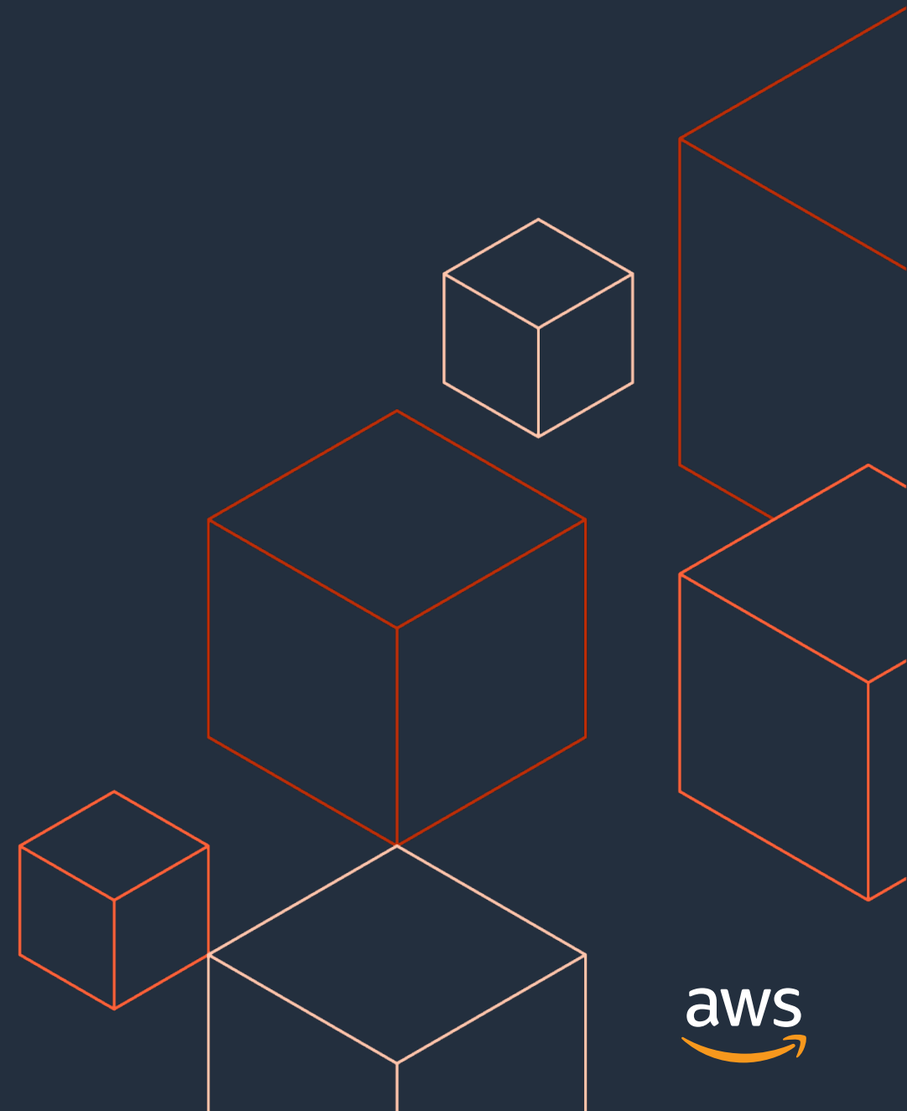
Glue DataBrew– Where does it fit?



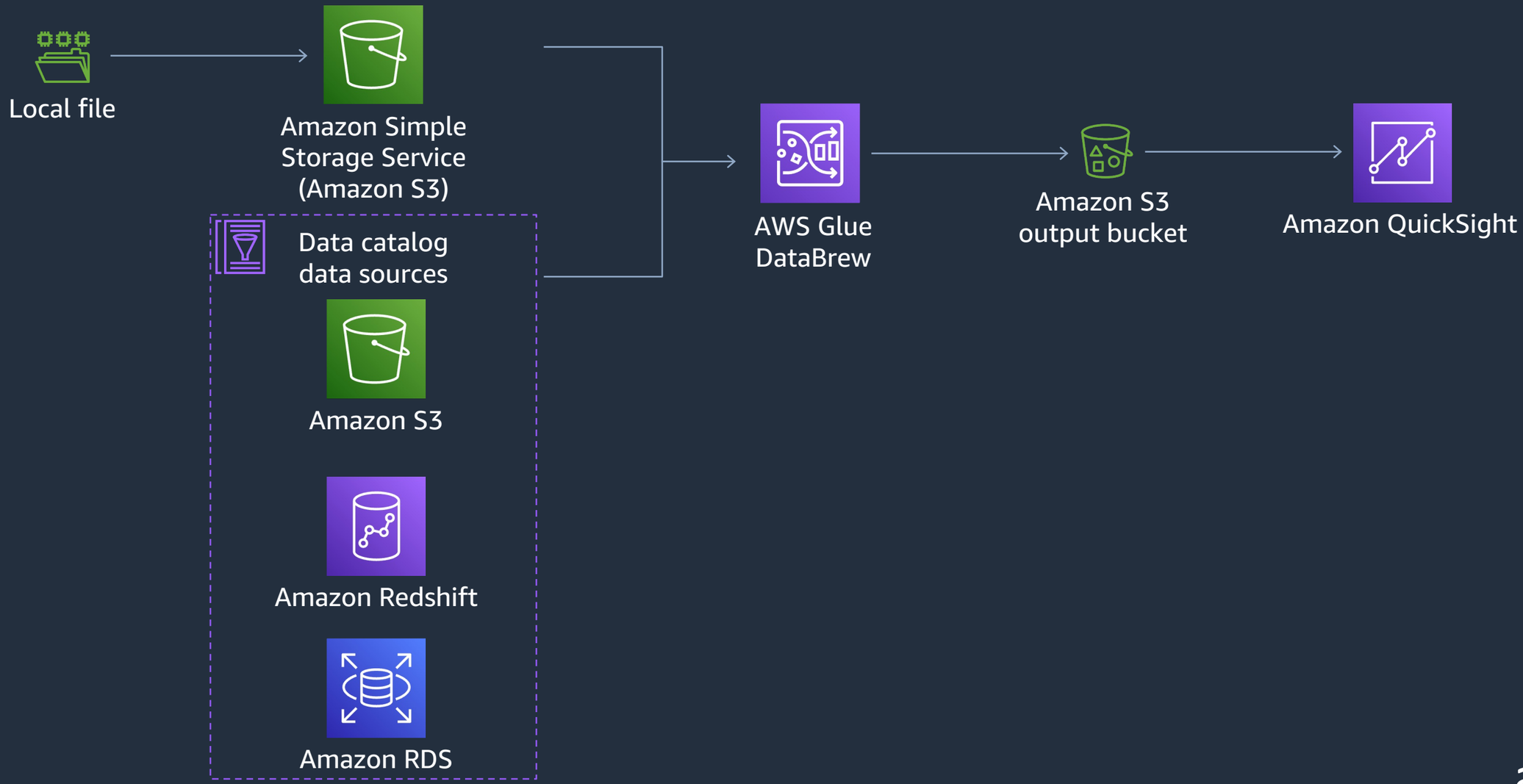
Clean and prepare real time and batch data

Develop complex data transformations	 **Glue Studio -** Visual ETL IDE for non-Spark experts	 **Dev Endpoints -** interactive development for Glue jobs for Spark Experts	 Data Engineer	ETL
Process IoT streams in real time	 **Glue Studio -** Develop streaming jobs visually	 **Glue Streaming -** processes real time data streams from IoT devices	 Data Engineer	Streaming
Replicate data across purpose built data stores	 **Glue Elastic Views**[1] **-** replicate data across multiple data stores without code		 Data Engineer	Data Replication
Transform data without writing code	 **Glue Data Brew -** build data without coding		 Data Analyst Data Scientist	Data Prep

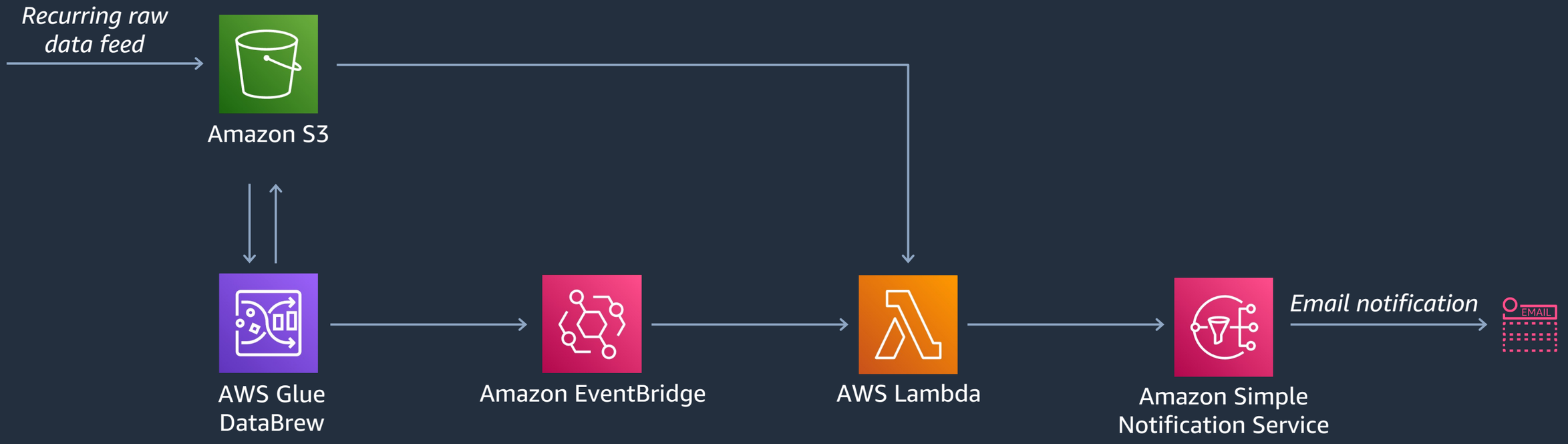
Popular use cases



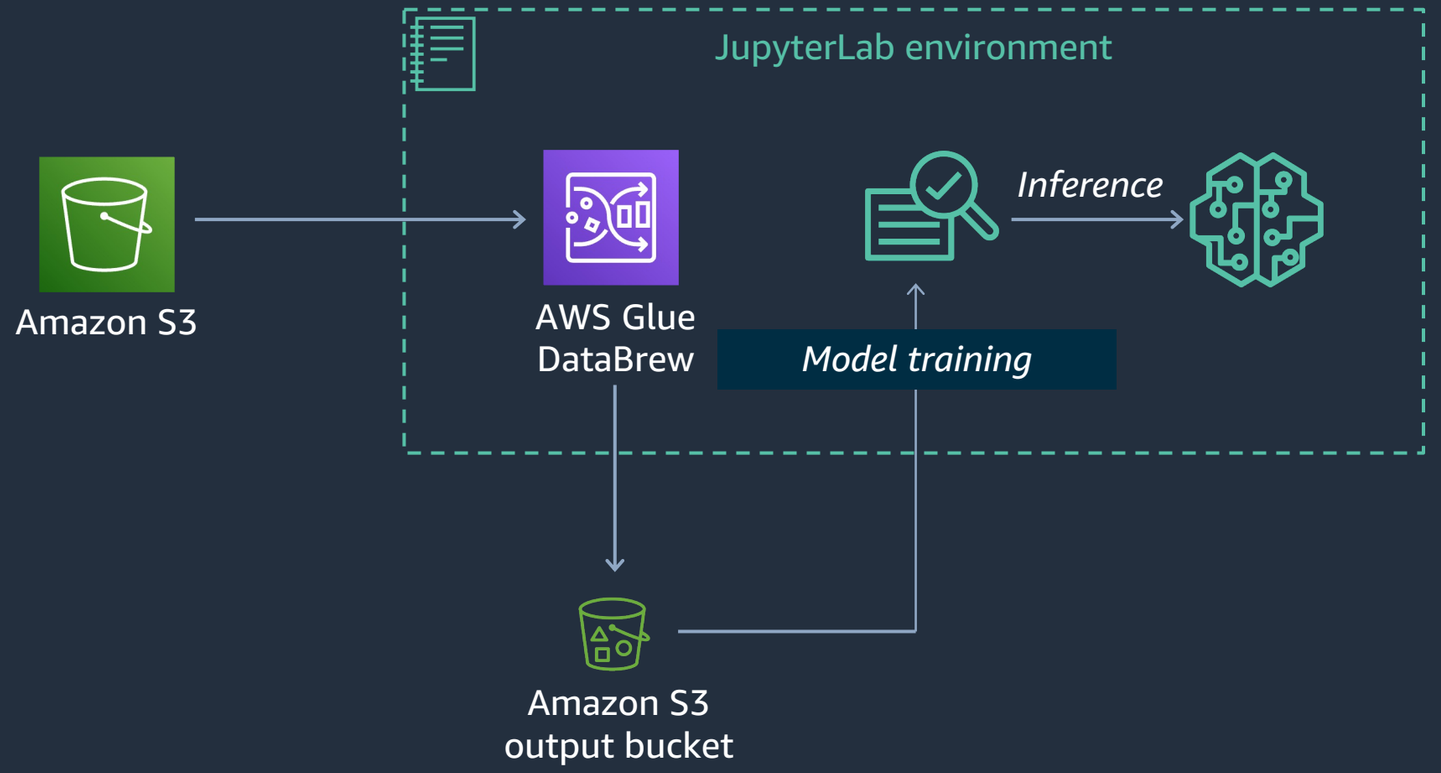
One-time data analysis for business reporting



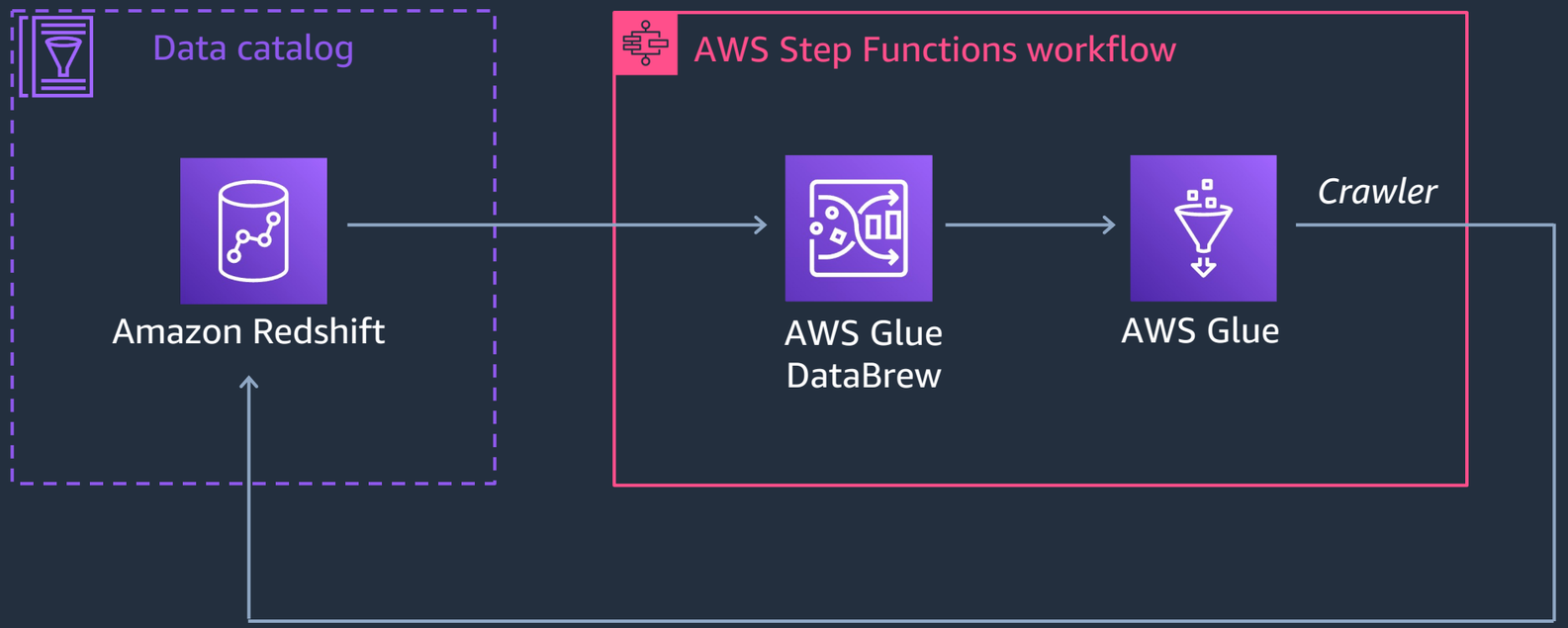
Set up data quality rules with AWS Lambda



Data preprocessing for machine learning

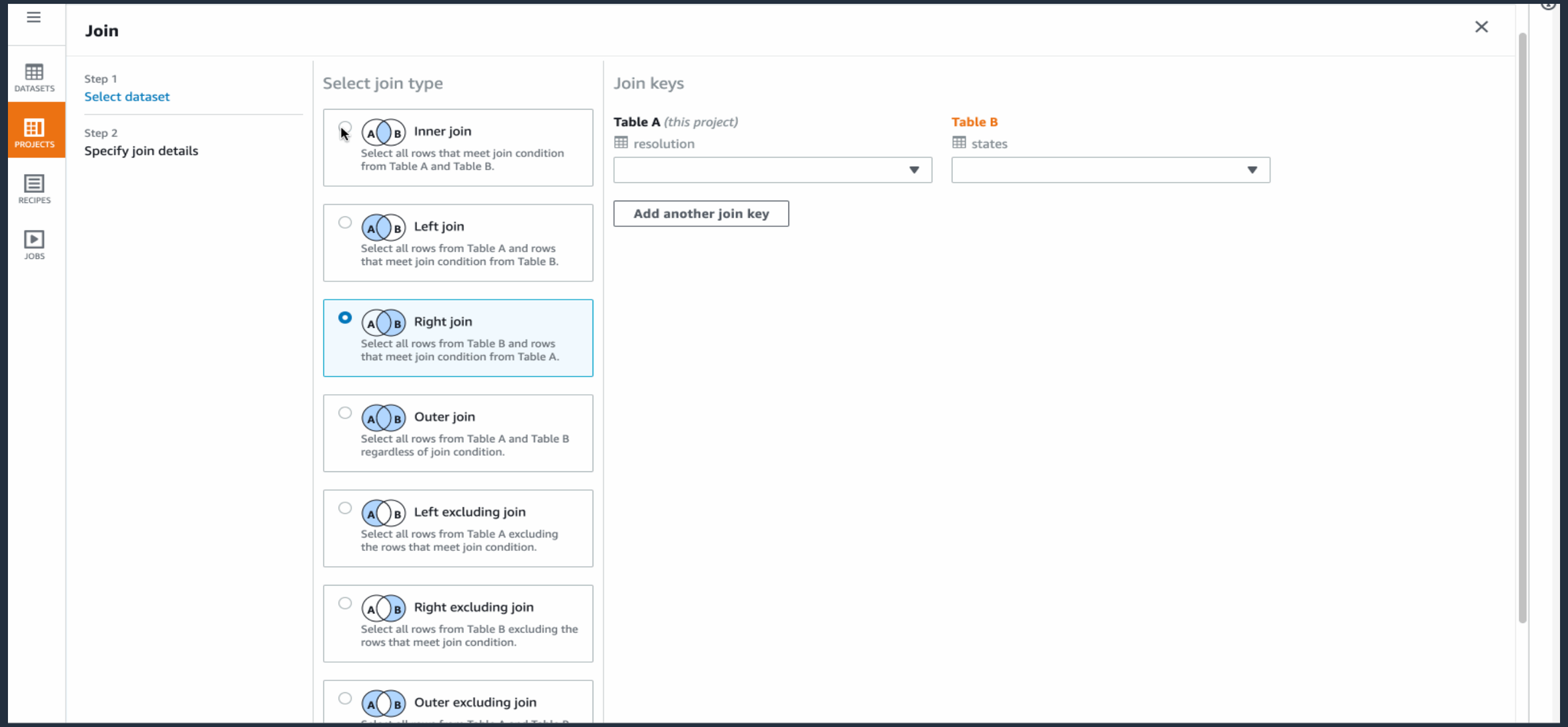


Orchestrating data preparation in workflows

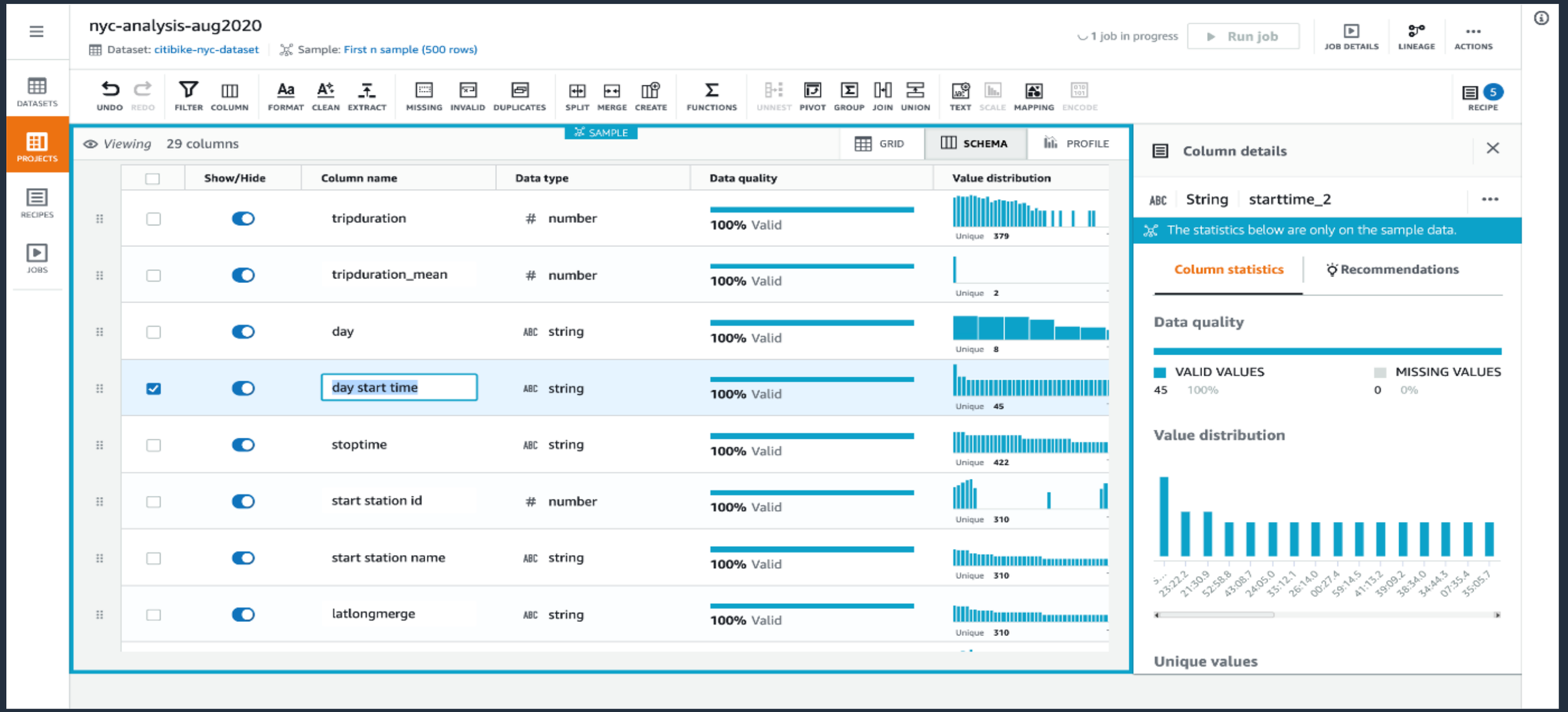


Some Highlighted Features

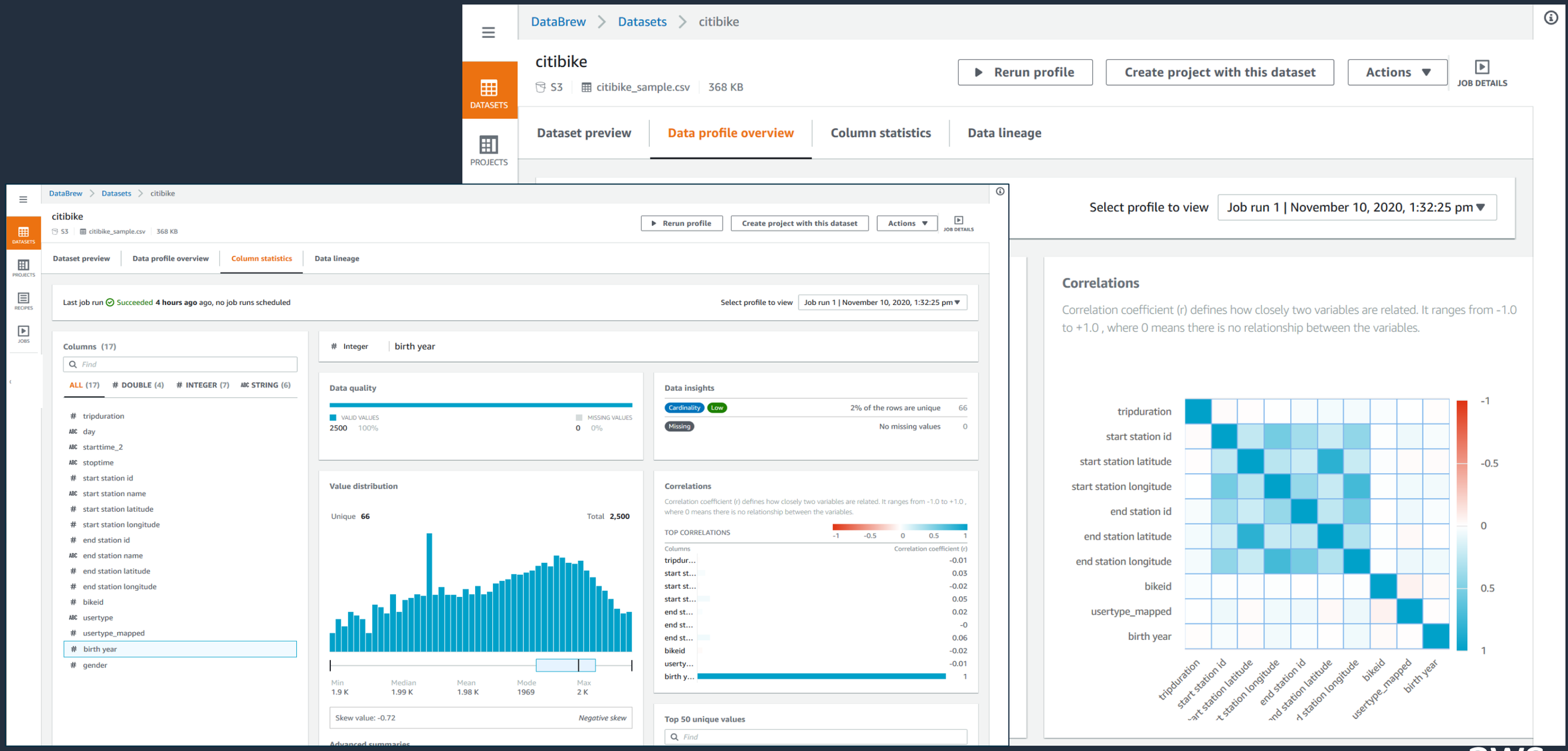
Enrich data with unions and joins



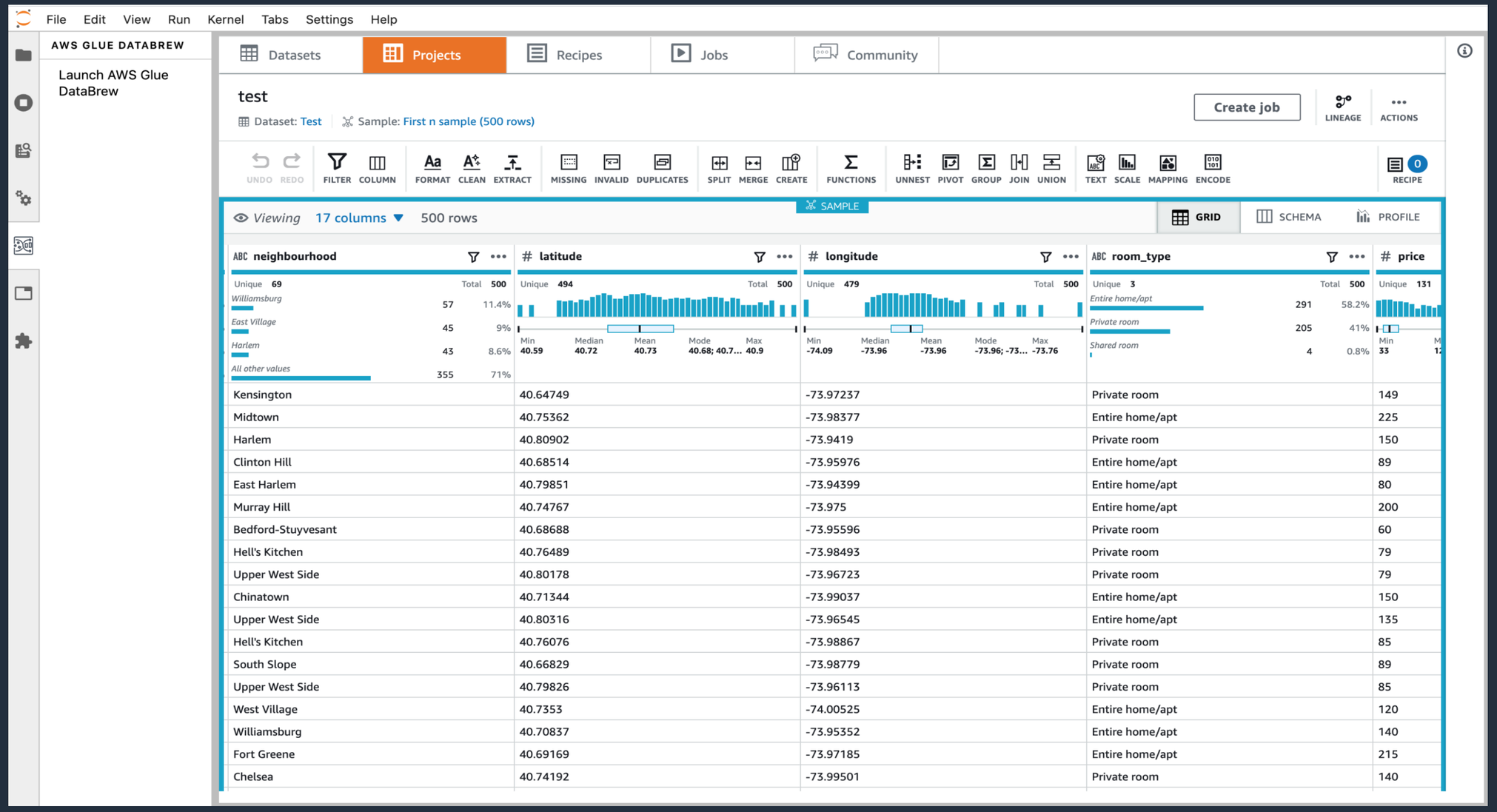
Update the schema



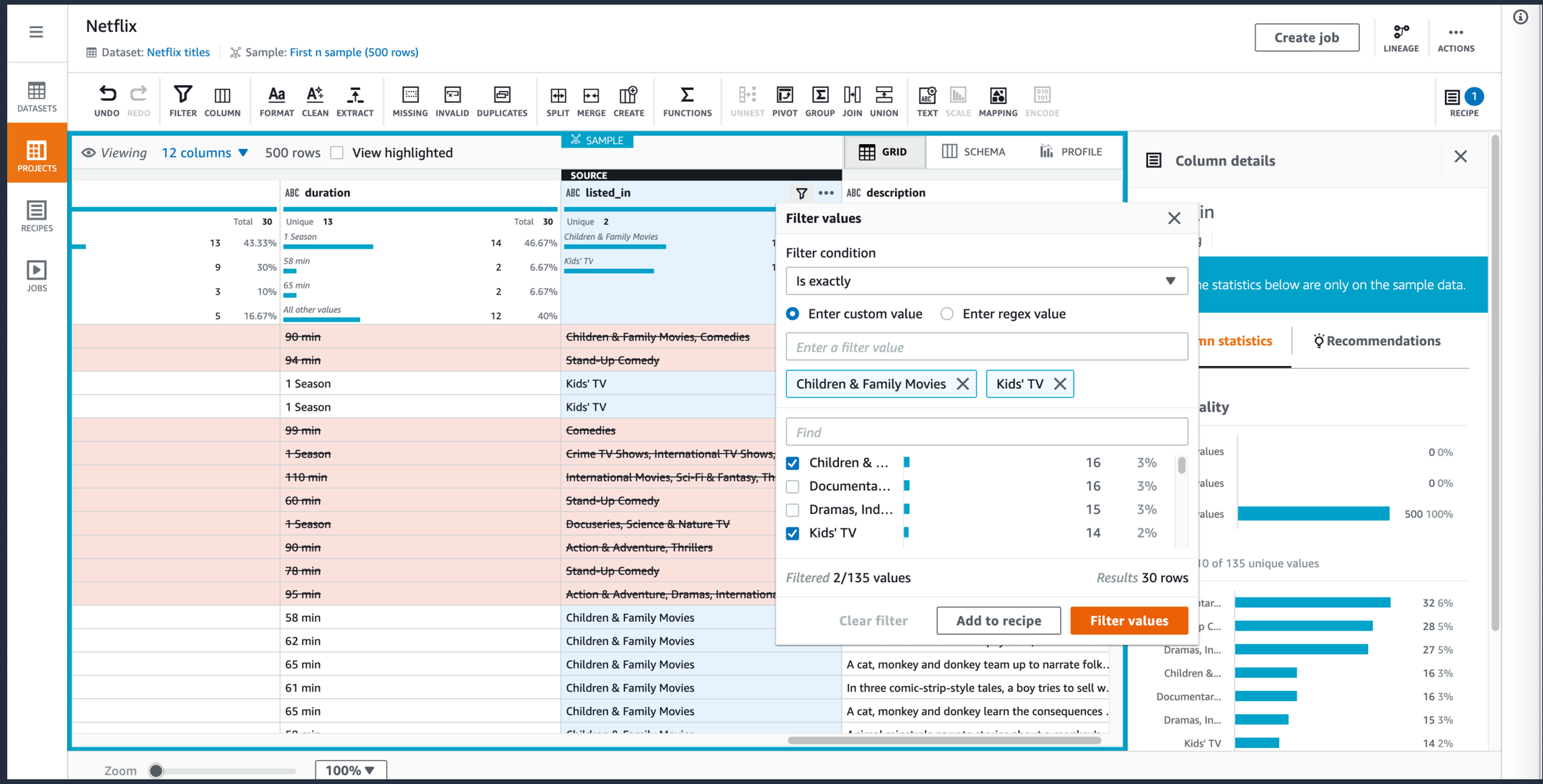
Generate data profiles



Feature engineering – plugin in Notebooks!



Remove outliers



Analyze aggregated data