



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

Trend: data is...



Growing
Exponentially



From new
sources



Increasingly
diverse

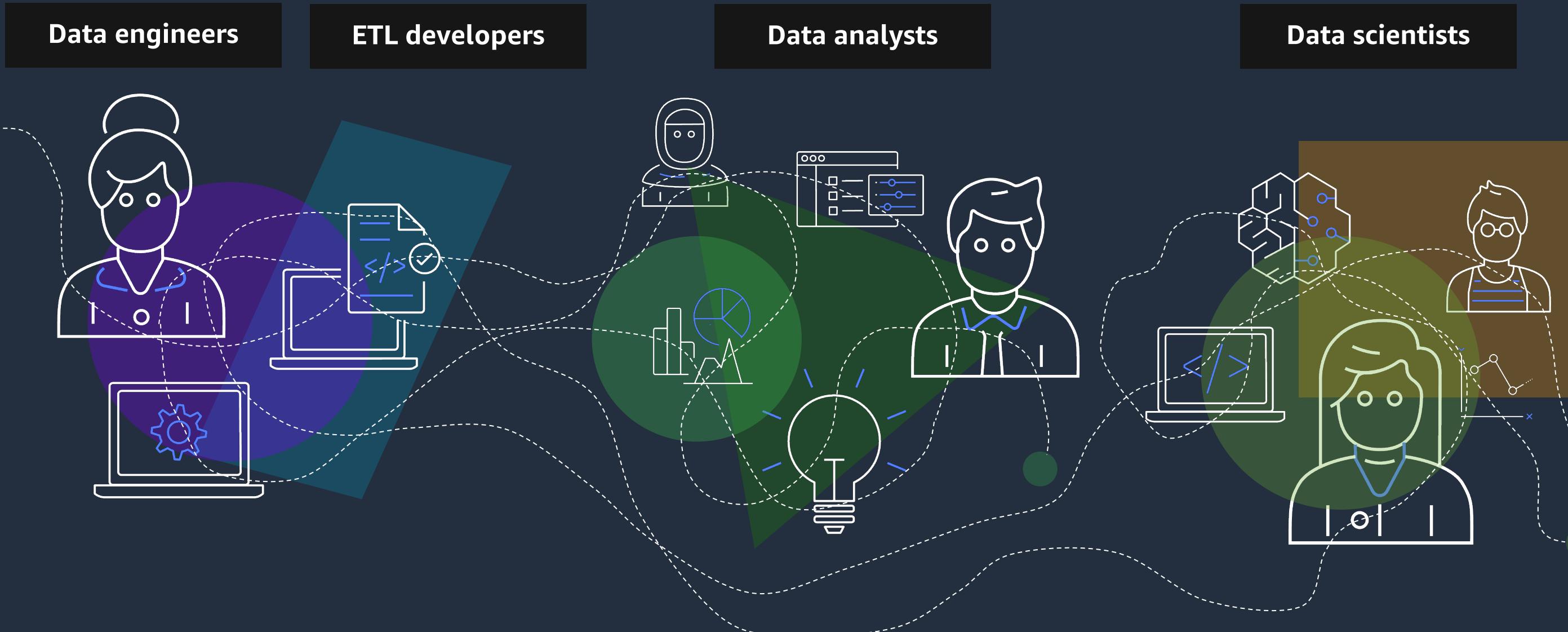


Used by people with
diverse skillsets



Accessed by many
applications

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



Needs the right tool for the right persona

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

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

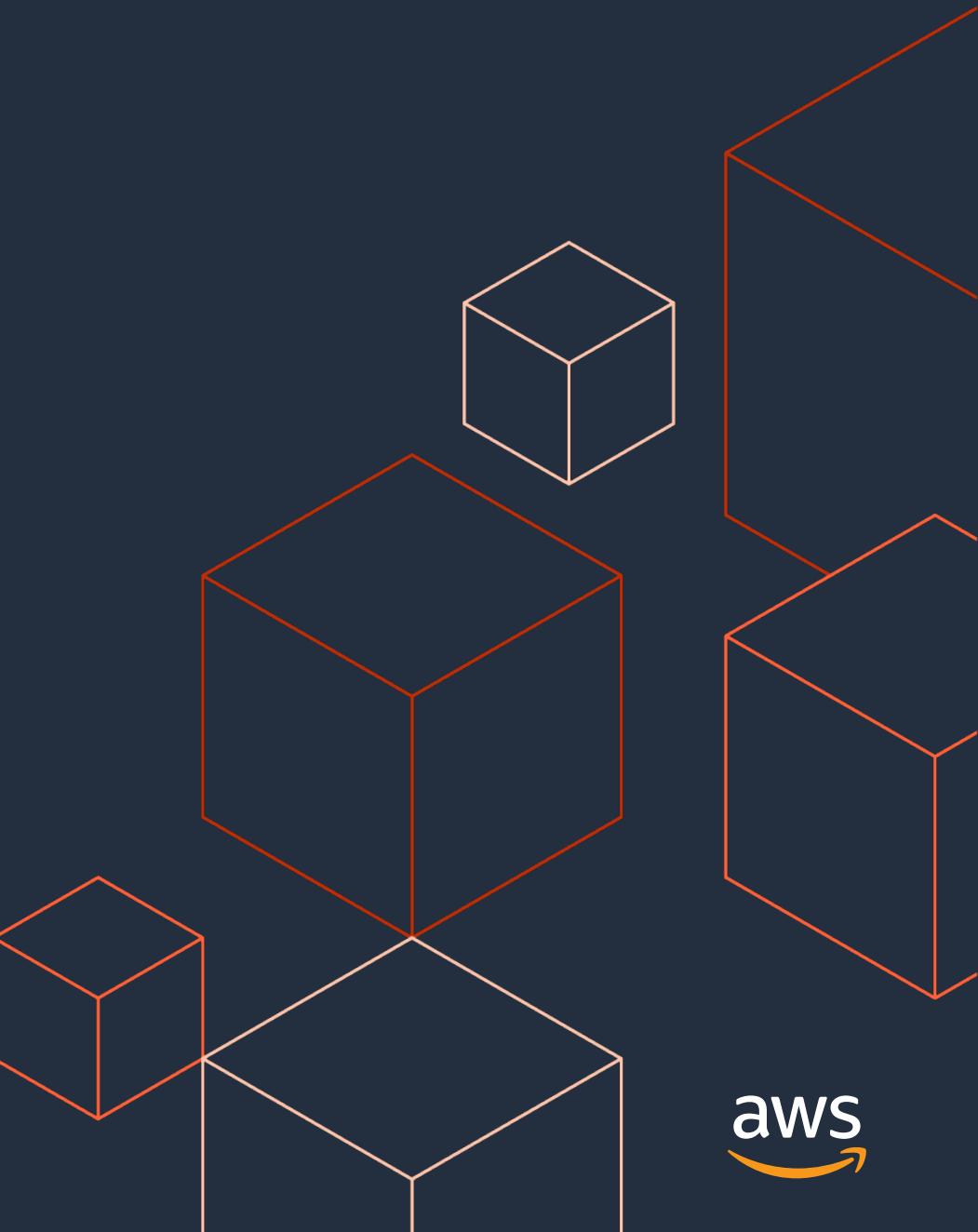


AWS Glue

Data movement

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

AWS Glue DataBrew



DataBrew – serverless, no code data preparation for data analysts and data scientists

The screenshot displays three main components of the AWS DataBrew interface:

- Visual lineage:** On the left, a diagram shows the flow of data from raw CSV files through a dataset to a project. It highlights a "visualized lineage scope" area.
- Data preview:** The central part shows a preview of the "nycitibikes" dataset with 21 columns and 500 rows. It includes histograms for numerical columns like start station latitude and longitude, and a grid view of the data.
- Merge columns:** A modal window titled "Merge columns" allows users to merge two source columns ("start station latitude" and "start station longitude") into a new column named "latlong".
- Data quality dashboard:** On the right, a dashboard provides insights such as "25% of the rows are unique" and "No missing values". It also shows correlations between variables and top 50 unique values.

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



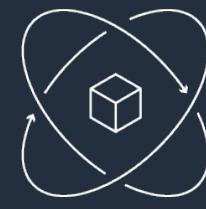
Clean and normalize data

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

DataBrew Functionality



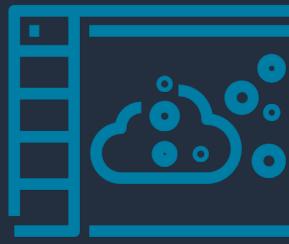
Connect to data sources

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



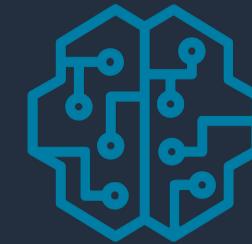
Visual Data lineage

View the various stages of data
transformation from start to
end



Serverless usage at scale

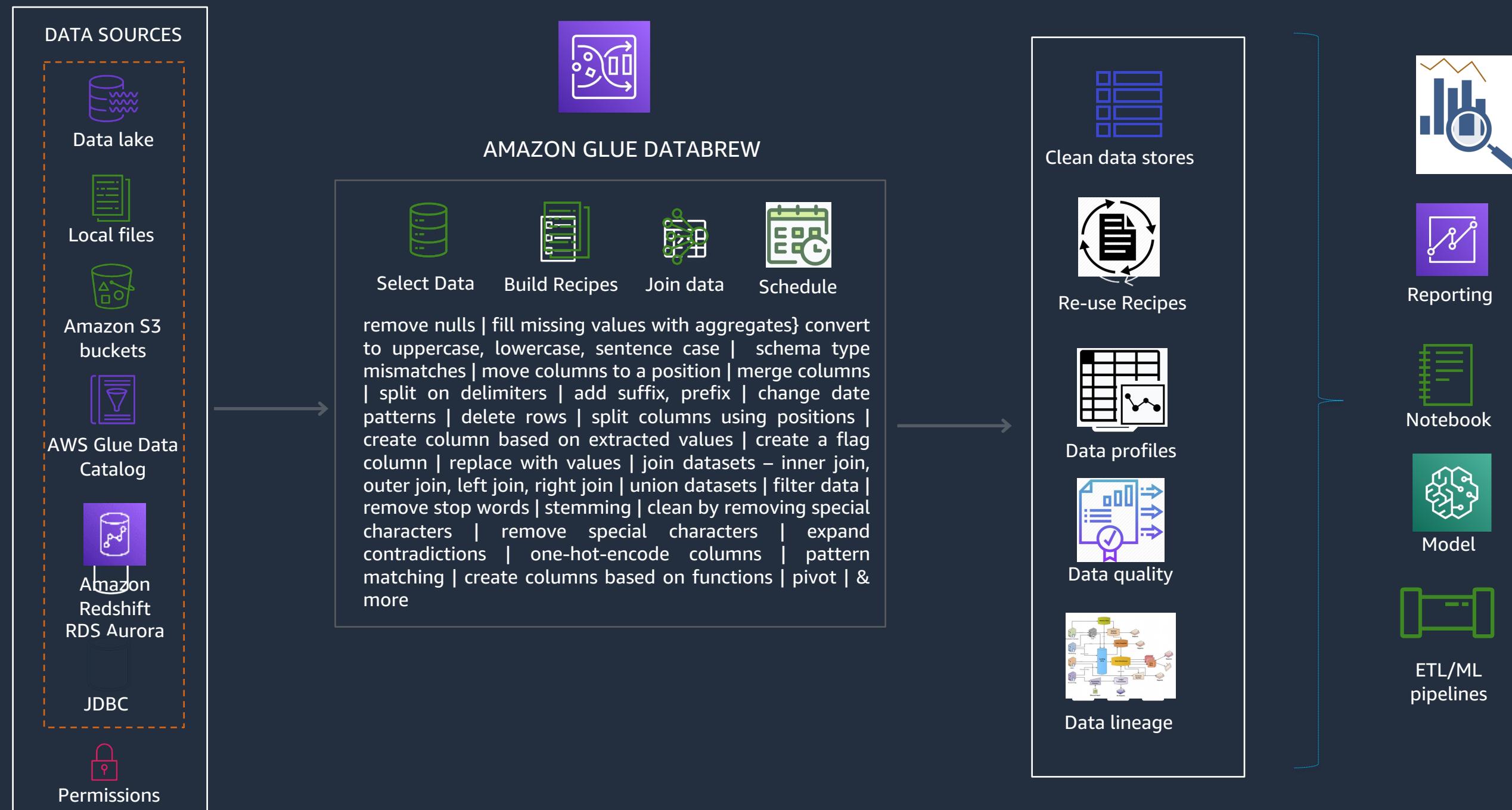
Operate at massive scale in a
serverless capacity. Pay only for what
you use



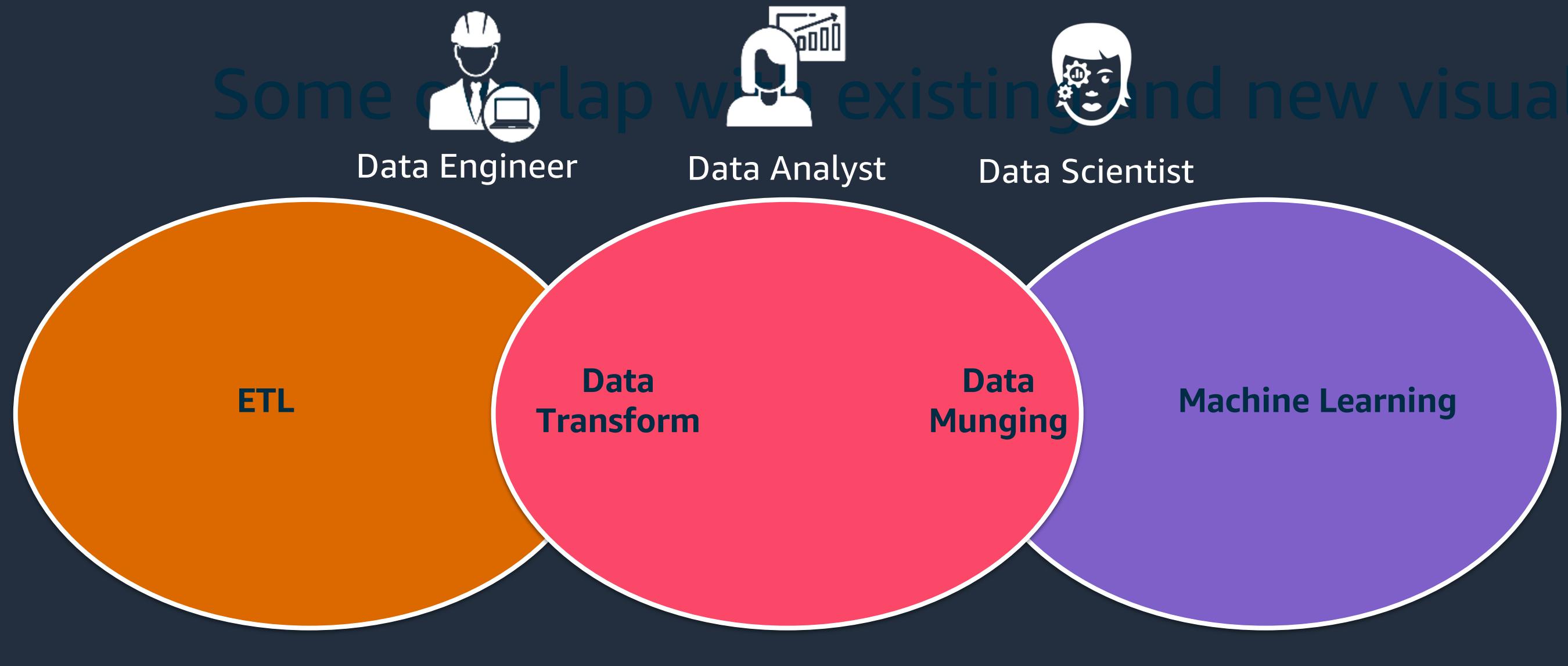
Advanced use cases

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

DataBrew: How does it work?

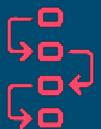


Glue DataBrew— Where does it fit?



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

Process IoT streams in real time



Glue Studio - Develop streaming jobs visually



Glue Streaming - processes real time data streams from IoT devices



Data Engineer

Replicate data across purpose built data stores



Glue Elastic Views¹ - replicate data across multiple data stores without code



Data Engineer

Transform data without writing code



Glue Data Brew - build data without coding



Data Analyst
Data Scientist

ETL

Streaming

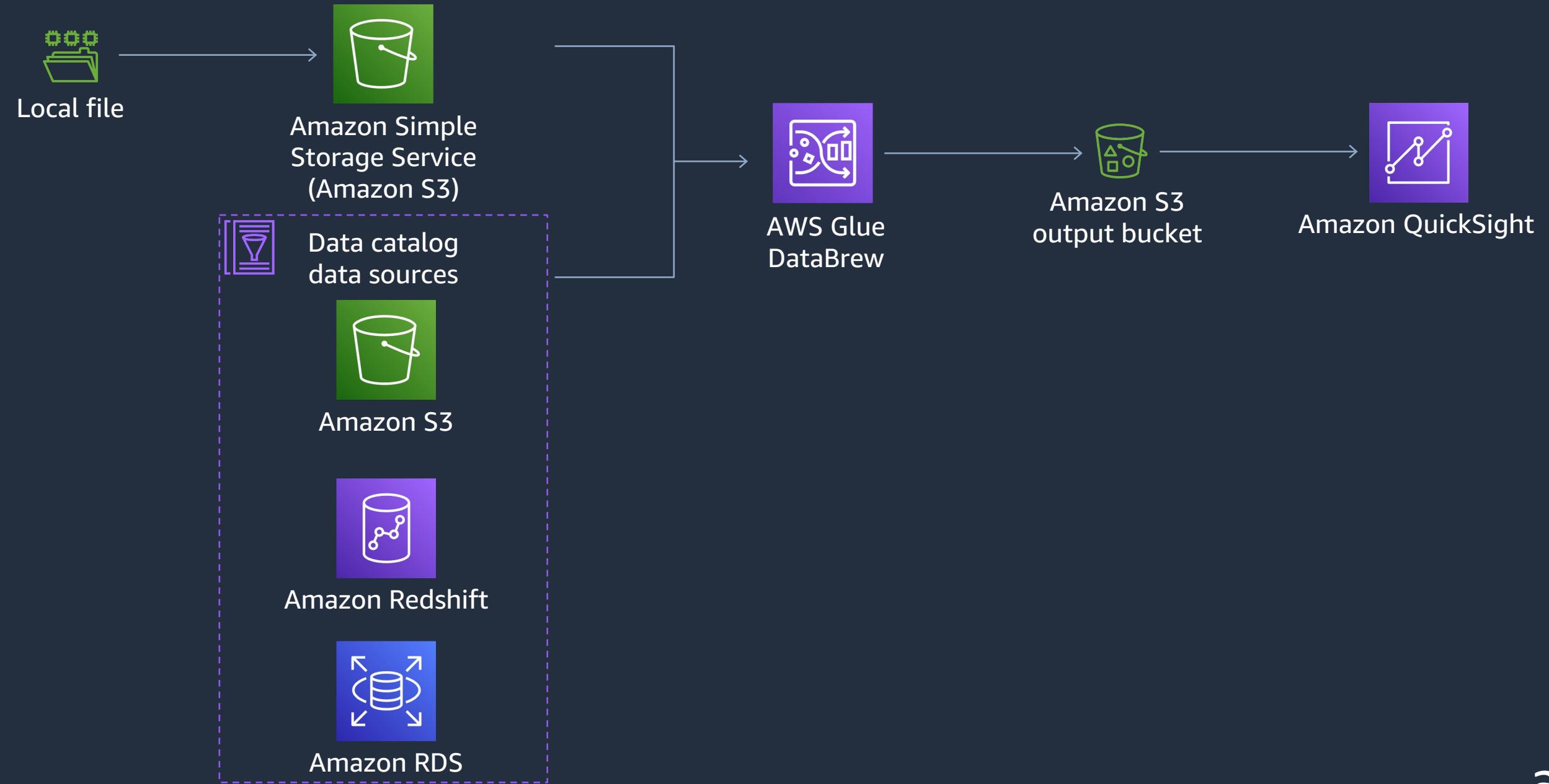
Data Replication

Data Prep

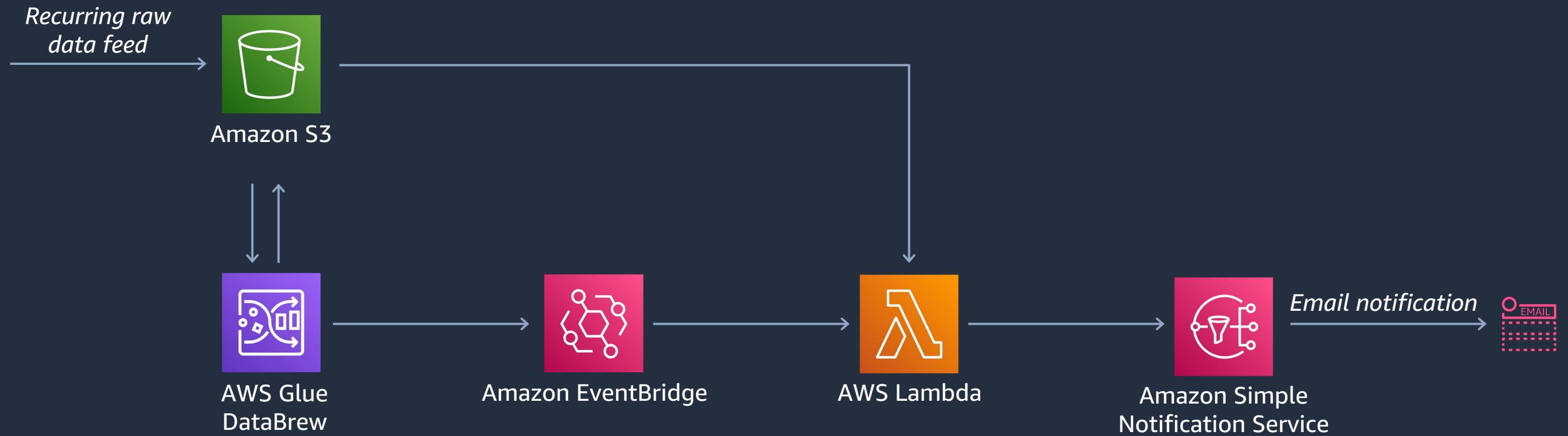
Popular use cases



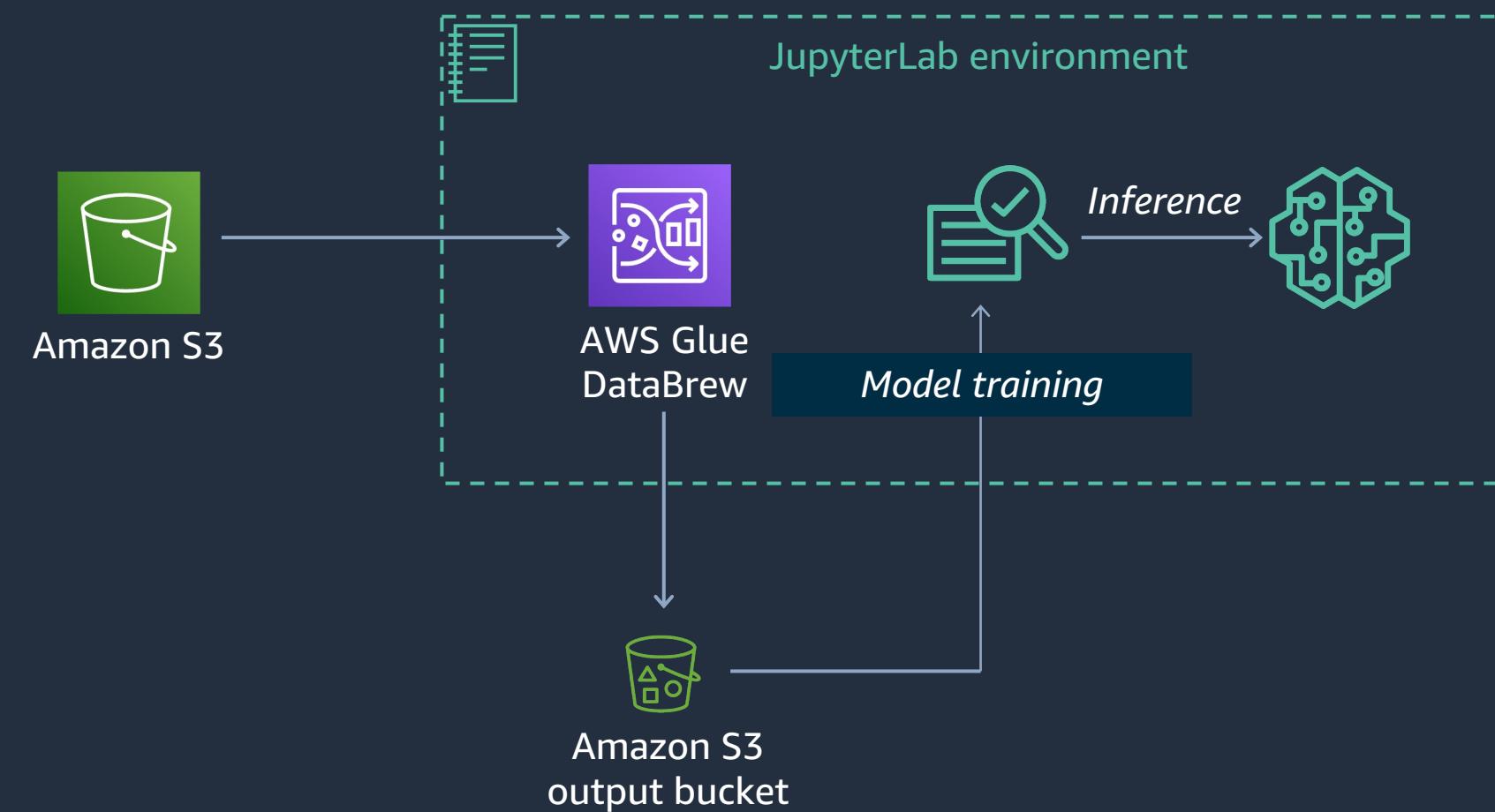
One-time data analysis for business reporting



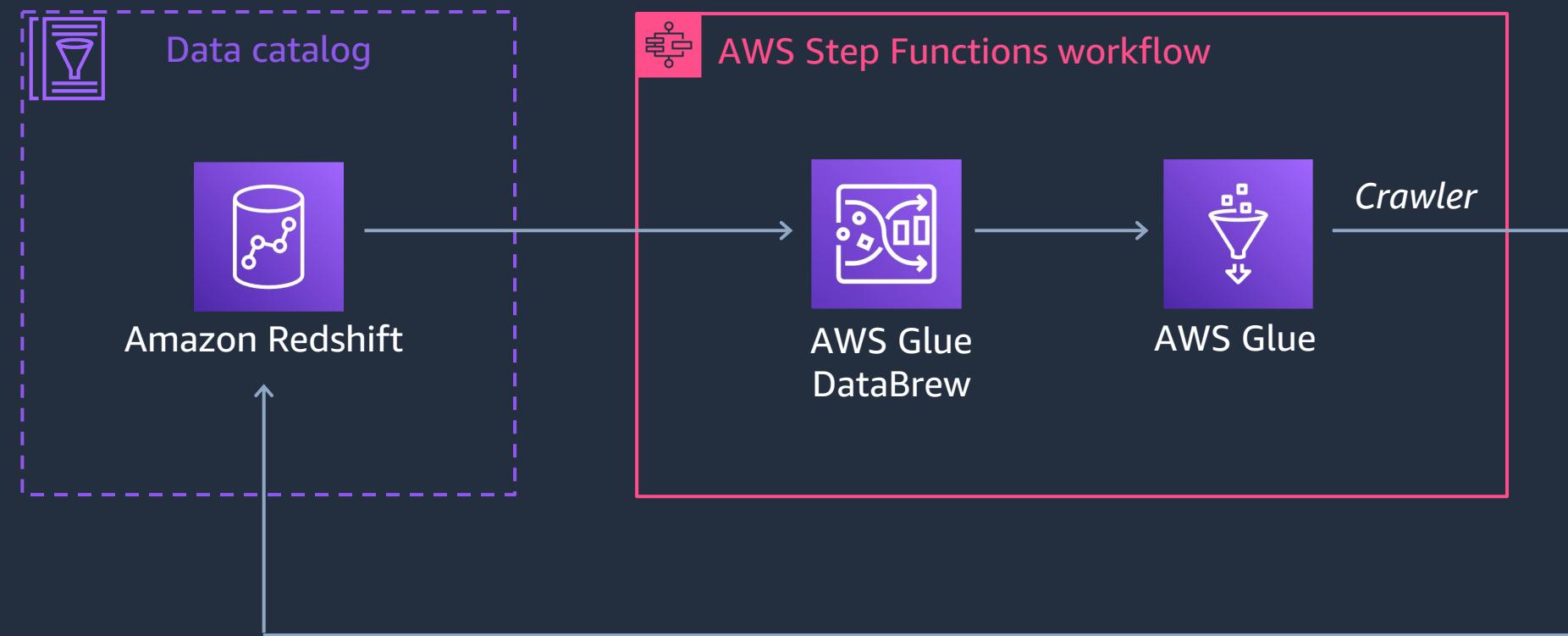
Set up data quality rules with AWS Lambda



Data preprocessing for machine learning



Orchestrating data preparation in workflows



Some Highlighted Features



Enrich data with unions and joins

Join

Step 1 Select dataset

Step 2 Specify join details

Select join type

-  Inner join
Select all rows that meet join condition from Table A and Table B.
-  Left join
Select all rows from Table A and rows that meet join condition from Table B.
-  Right join
Select all rows from Table B and rows that meet join condition from Table A.
-  Outer join
Select all rows from Table A and Table B regardless of join condition.
-  Left excluding join
Select all rows from Table A excluding the rows that meet join condition.
-  Right excluding join
Select all rows from Table B excluding the rows that meet join condition.
-  Outer excluding join

Join keys

Table A (this project)
resolution

Table B
states

Add another join key

Update the schema

Screenshot of the AWS Glue Data Catalog interface showing the schema for the "nyc-analysis-aug2020" dataset.

The dataset contains 29 columns:

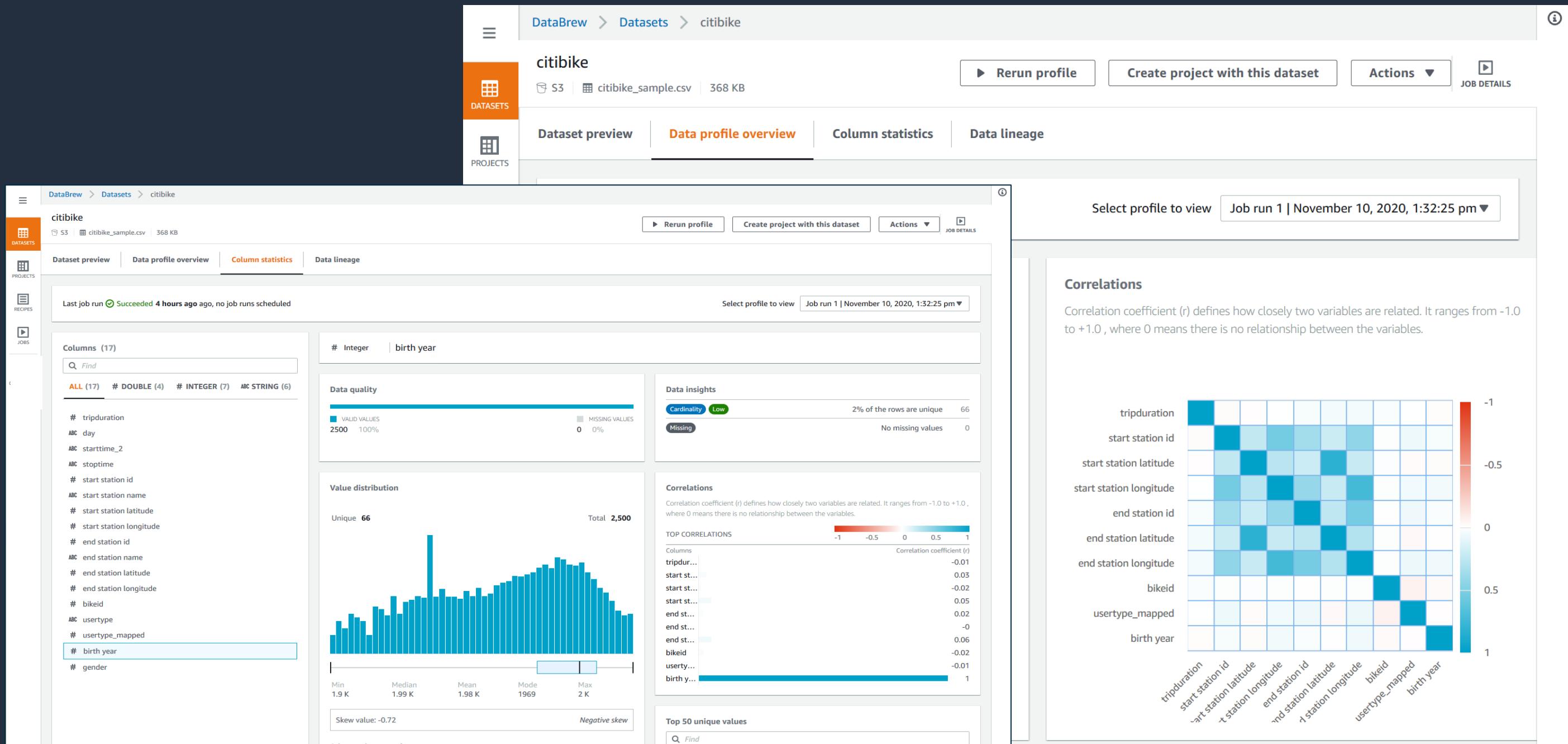
Column name	Data type	Data quality	Value distribution
tripduration	# number	100% Valid	Unique 379
tripduration_mean	# number	100% Valid	Unique 2
day	ABC string	100% Valid	Unique 8
day start time	ABC string	100% Valid	Unique 45
stoptime	ABC string	100% Valid	Unique 422
start station id	# number	100% Valid	Unique 310
start station name	ABC string	100% Valid	Unique 310
latlongmerge	ABC string	100% Valid	Unique 310

A tooltip for the "day start time" column indicates: "The statistics below are only on the sample data."

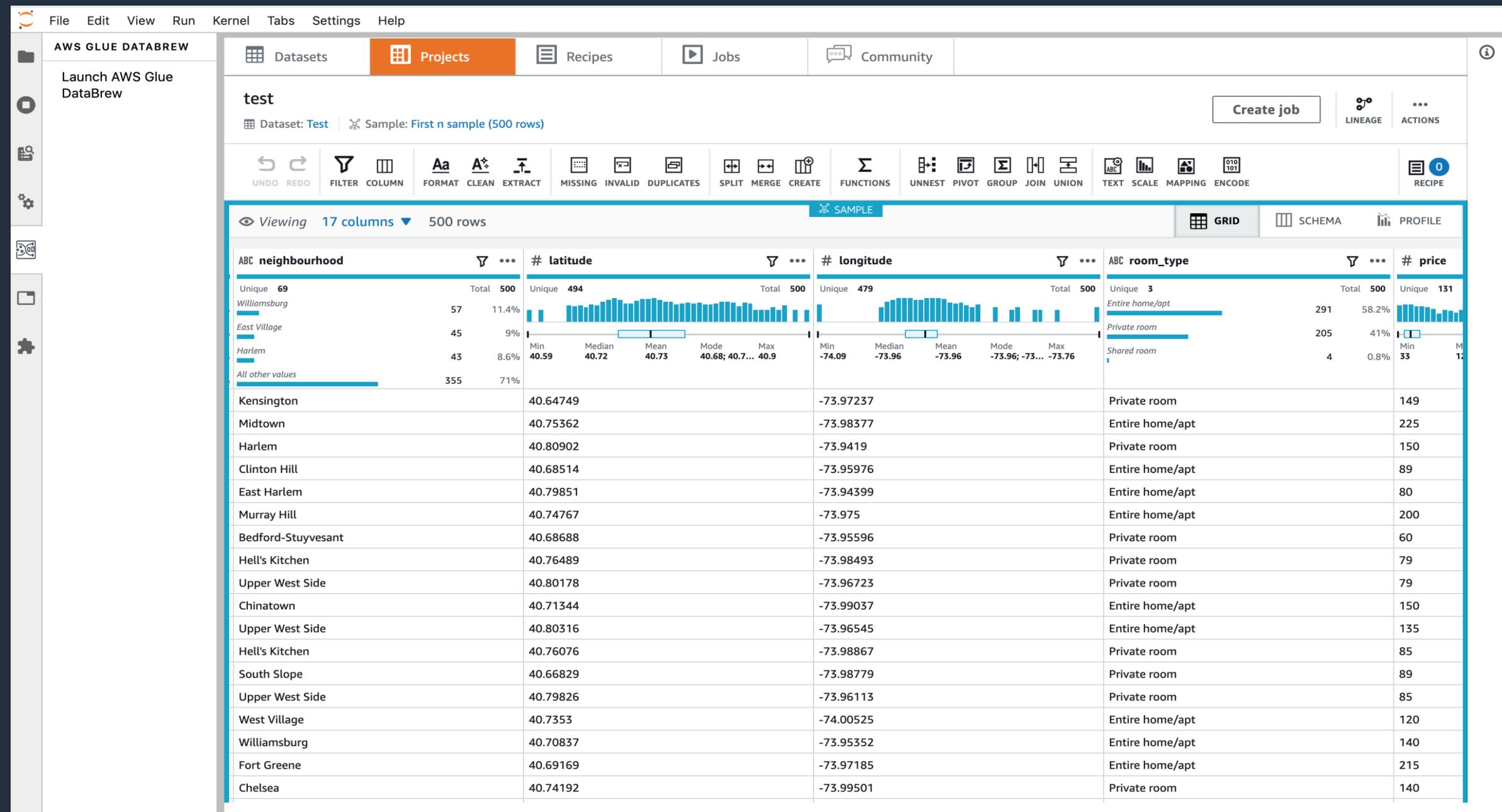
The "Column details" panel for "day start time" shows:

- Column statistics:** VALID VALUES (45, 100%), MISSING VALUES (0, 0%)
- Data quality:** 100% Valid
- Value distribution:** A histogram showing unique values from 23:22.2 to 07:35.4.
- Unique values:** A list of 45 unique values.

Generate data profiles



Feature engineering – plugin in Notebooks!



Remove outliers

The screenshot shows the AWS Glue Data Catalog interface for a dataset named "Netflix". The left sidebar has tabs for "PROJECTS" (selected), "RECIPES", and "JOBS". The main area displays a sample of 500 rows from the "Netflix titles" dataset. A context menu is open over the "ABC duration" column for the value "90-min". The menu includes options like "FORMAT", "CLEAN", "EXTRACT", "MISSING", "INVALID", "DUPLICATES", "SPLIT", "MERGE", "CREATE", "FUNCTIONS", "UNNEST", "PIVOT", "GROUP", "JOIN", "UNION", "TEXT", "SCALE", "MAPPING", and "ENCODE". The "SAMPLE" tab is selected. A tooltip states: "The statistics below are only on the sample data." The "Column details" panel on the right shows the distribution of "ABC listed_in" values: "Children & Family Movies" (16 rows, 3%) and "Kids' TV" (14 rows, 2%).

Analyze aggregated data

