



19 August 2021

Data preparation made easy with AWS Glue DataBrew

Vikas Omer

Senior Analytics Specialist Solutions Architect
Amazon Web Services



Agenda

- Data preparation: State of union
- Introducing AWS Glue DataBrew
- Key features
- Demo
- Use cases

“ Our teams spend too much time on the undifferentiated, repetitive, and mundane tasks associated with data preparation.

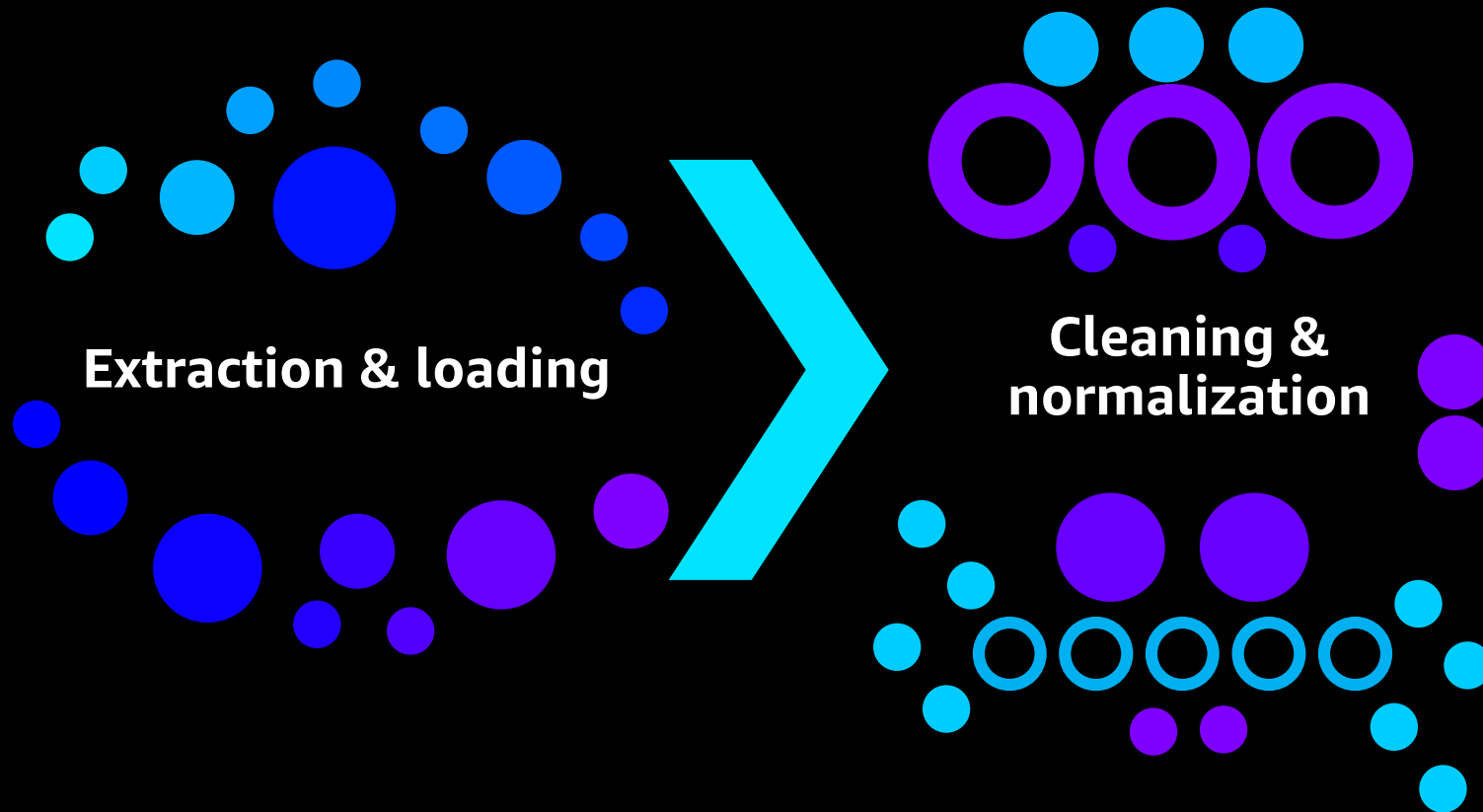
”

Preparing data involves several complex tasks

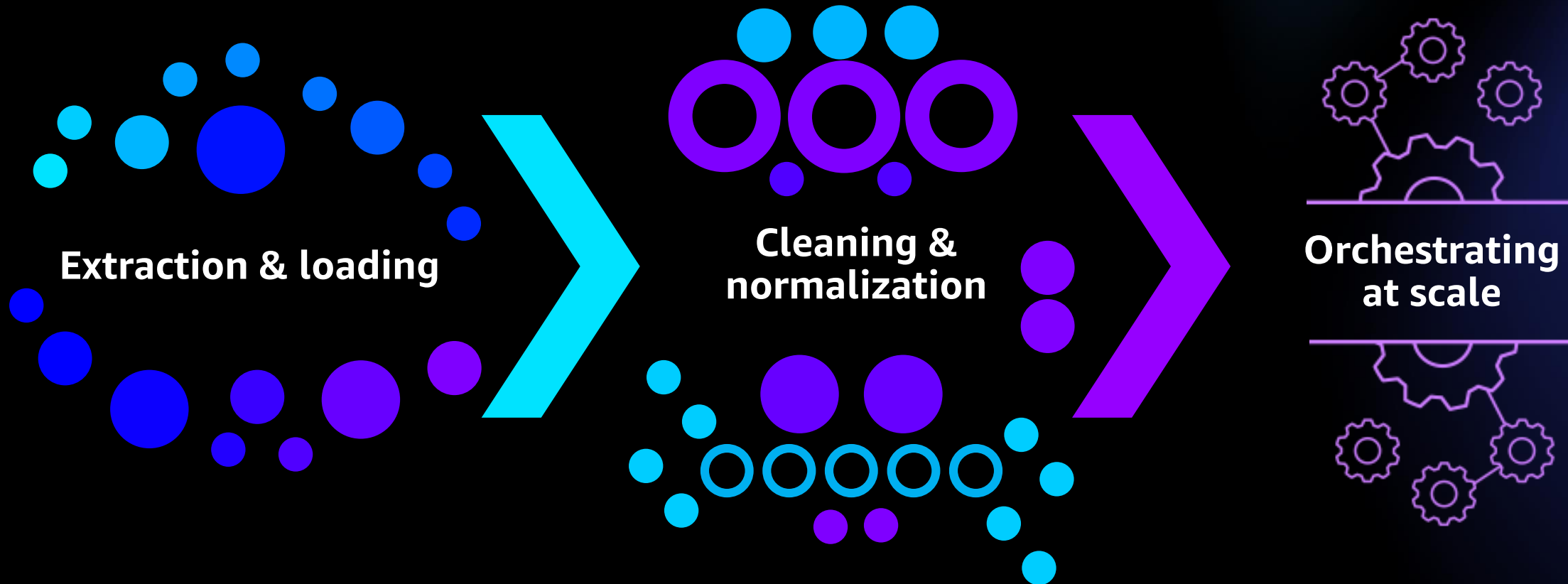
Preparing data involves several complex tasks



Preparing data involves several complex tasks



Preparing data involves several complex tasks



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

Data engineers

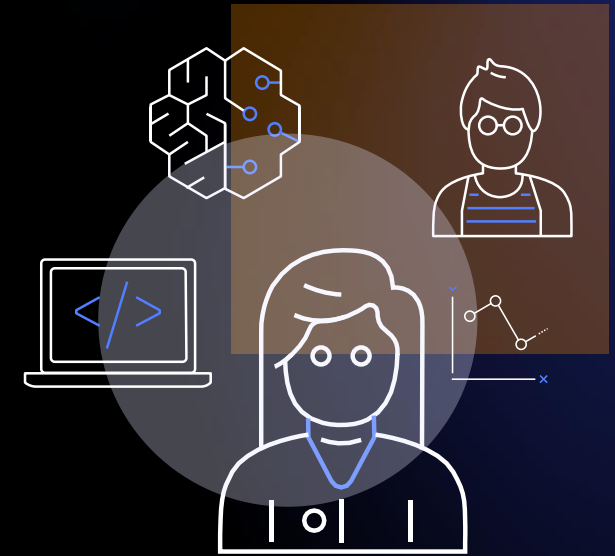
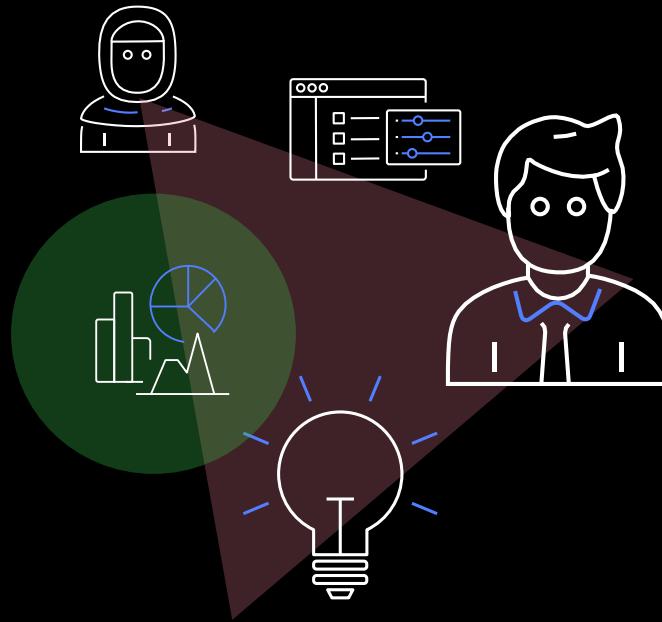


ETL developers

Data analysts

Business analysts

Data scientists



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

Data engineers

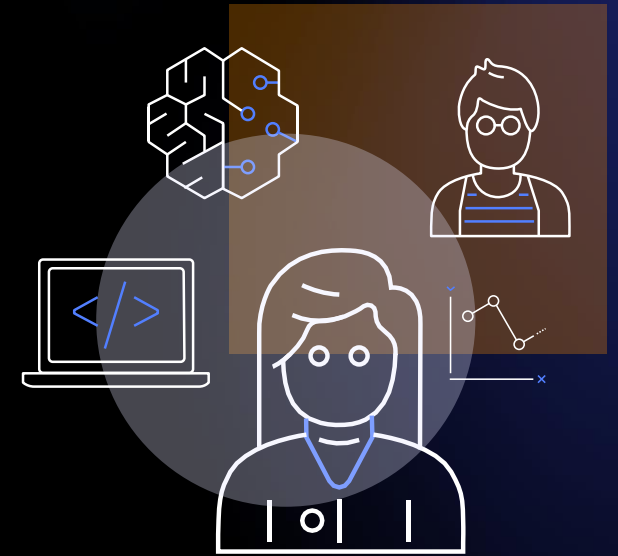
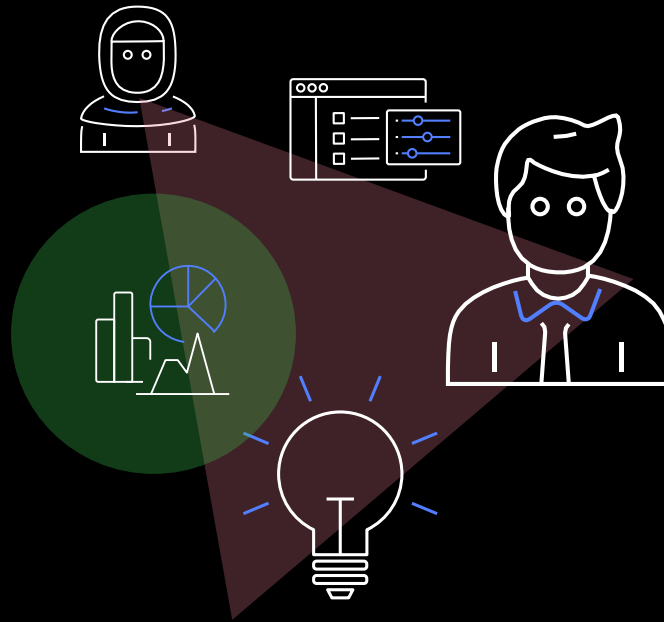


ETL developers

Data analysts

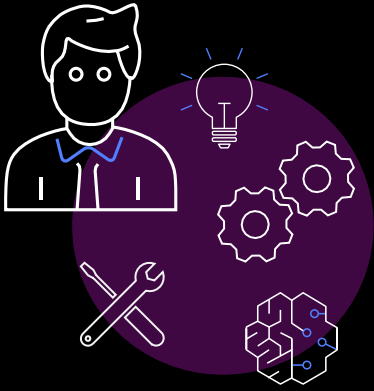
Business analysts

Data scientists



Needs the right tool for the right persona

Challenges with traditional data preparation



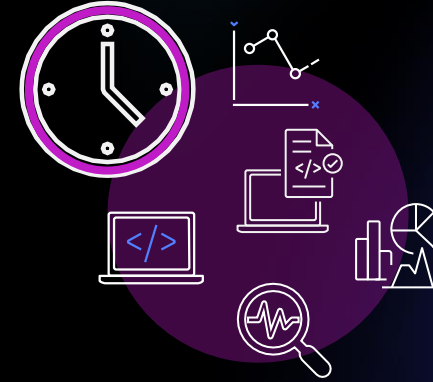
Manual

Needs a lot of code-based heavy-lifting for it to work at scale



Siloed

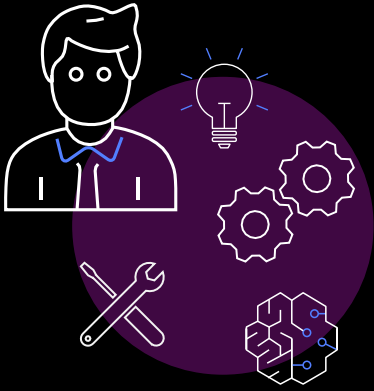
Often requires moving large amounts of data from siloed data sources



Time consuming

Needs the right tools for the right persona that are integrated

Challenges with traditional data preparation



Manual

Needs a lot of code-based heavy-lifting for it to work at scale



Siloed

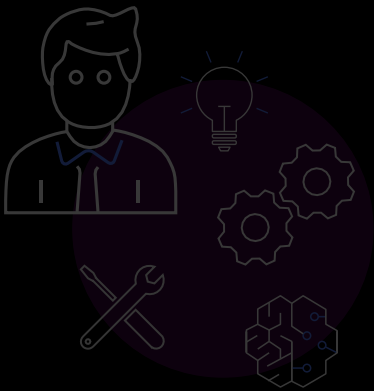
Often requires moving large amounts of data into silos, at times out of VPCs



Time consuming

Needs the right tools for the right persona that are integrated

Challenges with traditional data preparation



Manual

Needs a lot of code-based heavy-lifting for it to work at scale



Siloed

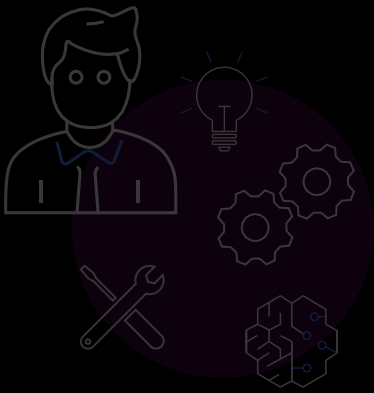
Often requires moving large amounts of data into silos, at times out of VPCs



Time consuming

Needs the right tools for the right persona that are integrated

Challenges with traditional data preparation



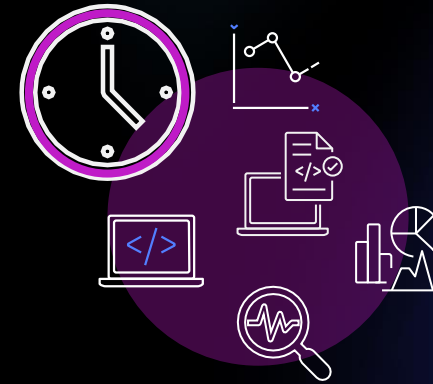
Manual

Needs a lot of code-based heavy-lifting for it to work at scale



Siloed

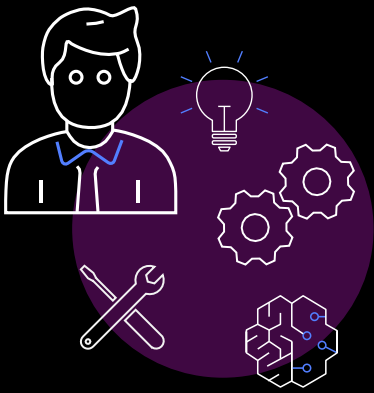
Often requires moving large amounts of data into silos, at times out of VPCs



Time consuming

Needs the right tools for the right persona that are integrated

Challenges with traditional data preparation



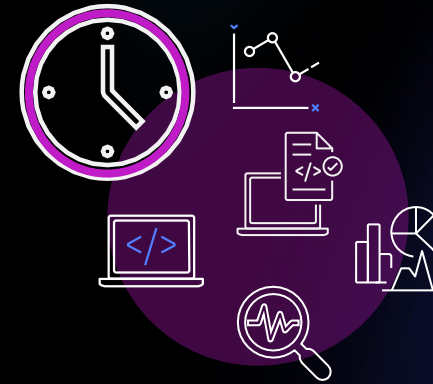
Manual

Needs a lot of code-based heavy-lifting for it to work at scale



Siloed

Often requires moving large amounts of data into silos, at times out of VPCs



Time consuming

Needs the right tools for the right persona that are integrated

Introducing **AWS Glue DataBrew**



Clean and normalize data up to 80% faster

AWS Glue DataBrew is a new visual data preparation tool that makes it easy for data analysts and data scientists to clean and normalize data to prepare it for analytics and machine learning.

The screenshot displays the AWS Glue DataBrew interface for a project named 'nycitibikes'. The interface is divided into several sections:

- Data lineage:** A diagram on the left shows the data flow from an S3 bucket (citibike_sample.csv) through a 'Join' operation to a 'Dataset' (citibike) and then to a 'Recipe' (dataset-met-objects).
- Dataset view:** The main area shows a preview of the 'citibike' dataset with 500 rows. It includes a table with columns: '# start station latitude', '# start station longitude', 'latlong', and '# end station id'. The table shows a list of coordinates and a concatenated string of latitude and longitude values.
- Merge columns dialog:** A modal window on the right titled 'Merge columns' is open. It allows selecting two or more columns to merge into a new column. The 'Source column' section shows 'start station latitude' and 'start station longitude' selected. The 'New column name' section shows 'latlong' entered. The 'Separator - Optional' section is also visible.
- Data insights:** A sidebar on the right provides summary statistics for the dataset, including 'Cardinality' (29% of rows are unique), 'Missing' values (0), and 'Correlations' between variables.

Built for data analysts and data scientists



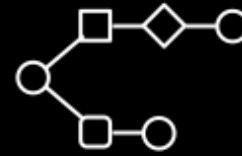
Understand data quality

Understand patterns and detect anomalies using profiles



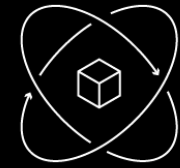
Clean and normalize data

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

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

Built for data analysts and data scientists



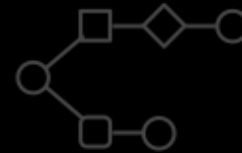
Understand data quality

Understand patterns and detect anomalies using profiles



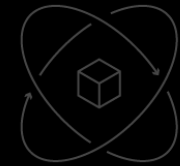
Clean and normalize data

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

Built for data analysts and data scientists



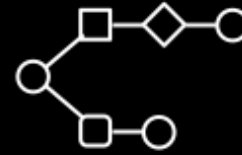
Understand data quality

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

Built for data analysts and data scientists



Understand data quality

Understand patterns and detect anomalies using profiles



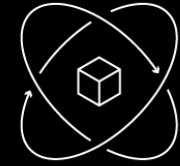
Clean and normalize data

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

Built for data analysts and data scientists



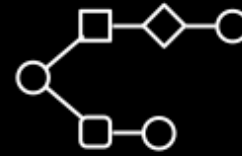
Understand data quality

Understand patterns and detect anomalies using profiles



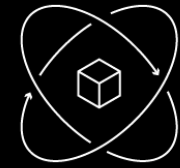
Clean and normalize data

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

Demo



What we will do in the demo



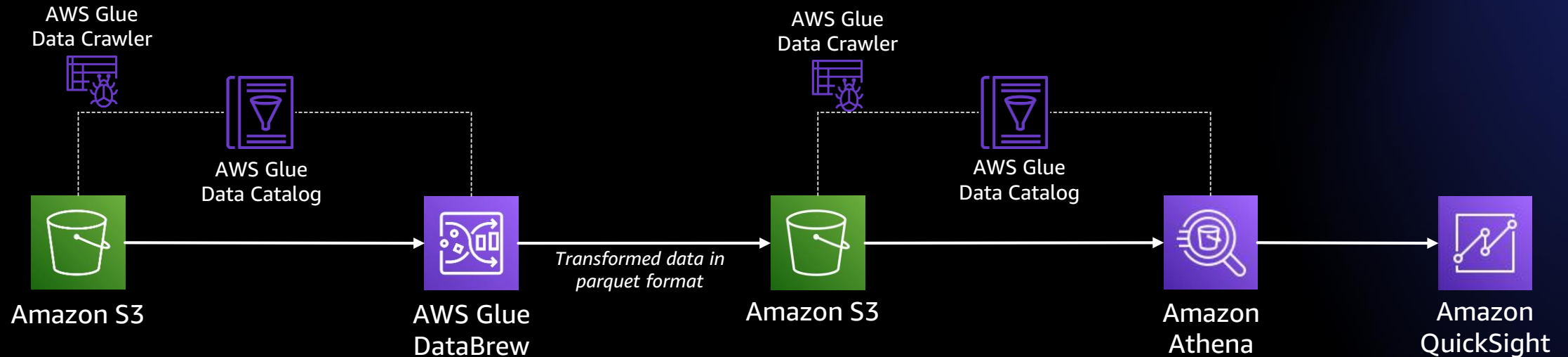
Music App
Startup



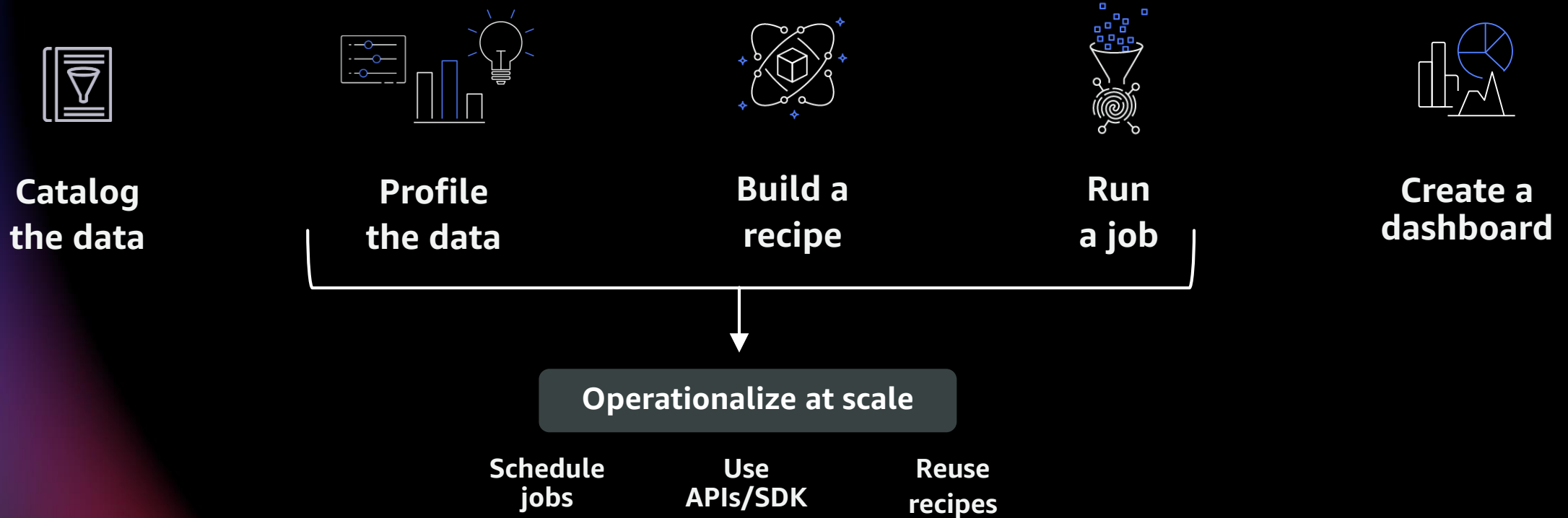
Ana
Data Analyst



Zhang
Data Scientist

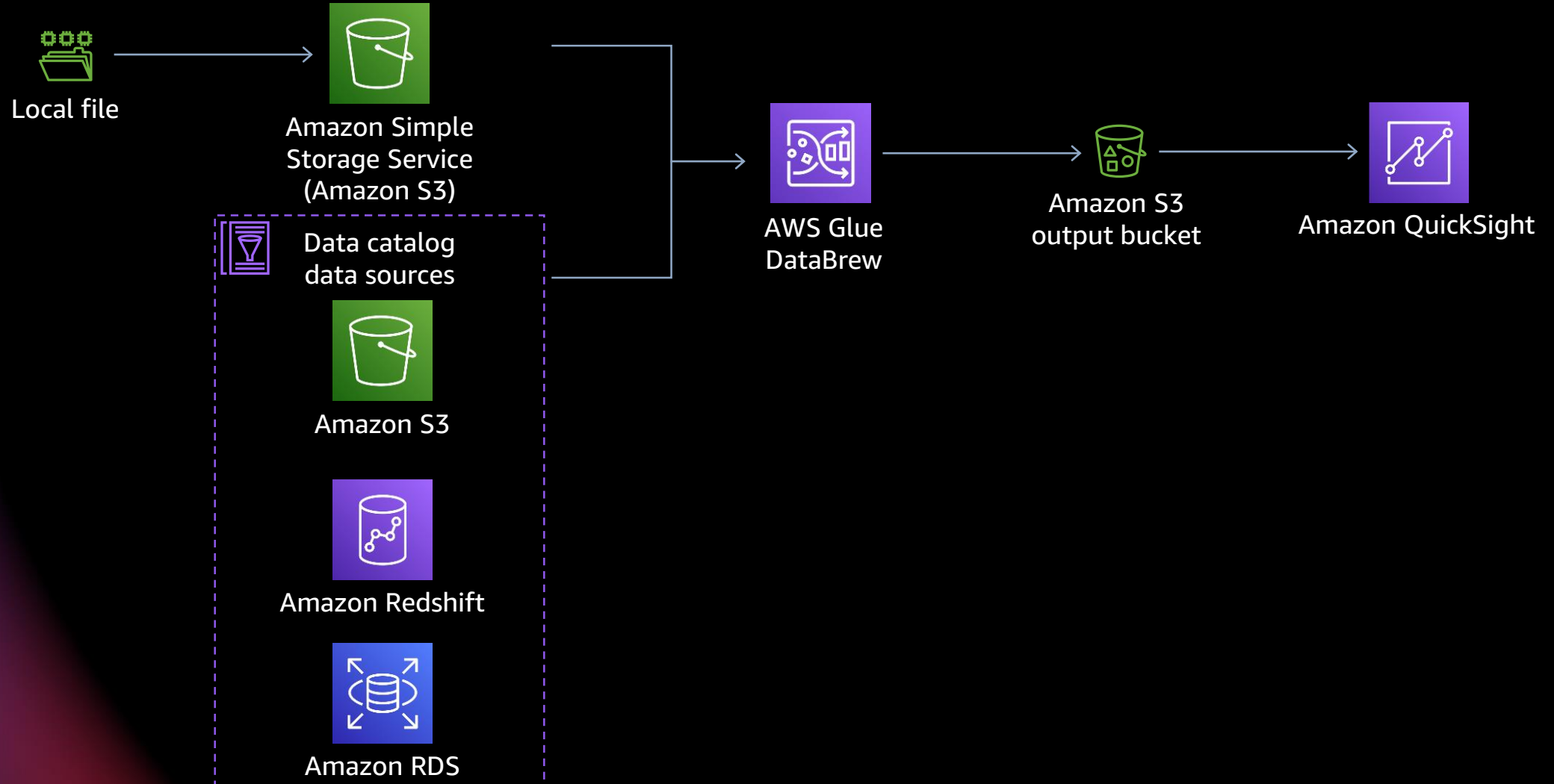


What we saw in the demo



Popular use cases

One-time data analysis for business reporting



Enrich data with unions and joins

Join

DATASETS

PROJECTS

RECIPES


JOBS

Step 1

Select dataset


Step 2

Specify join details




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

Select all rows from Table A and Table B.

Join keys

Table A (this project)

resolution

Table B

states

Add another join key

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Update the schema

DATASETS

PROJECTS

RECIPES

JOBS

nyc-analysis-aug2020

Dataset: citibike-nyc-dataset | Sample: First n sample (500 rows)

1 job in progress [Run job](#)




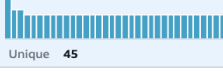



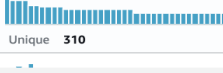
[JOB DETAILS](#) [LINEAGE](#) [ACTIONS](#)

UNDO REDO FILTER COLUMN FORMAT CLEAN EXTRACT MISSING INVALID DUPLICATES SPLIT MERGE CREATE FUNCTIONS UNNEST PIVOT GROUP JOIN UNION TEXT SCALE MAPPING ENCODE

Viewing 29 columns

SAMPLE

GRID SCHEMA PROFILE

	Show/Hide	Column name	Data type	Data quality	Value distribution
<input type="checkbox"/>	<input type="checkbox"/>	tripduration	# number	100% Valid	 Unique 379
<input type="checkbox"/>	<input type="checkbox"/>	tripduration_mean	# number	100% Valid	 Unique 2
<input type="checkbox"/>	<input type="checkbox"/>	day	ABC string	100% Valid	 Unique 8
<input checked="" type="checkbox"/>	<input type="checkbox"/>	day start time	ABC string	100% Valid	 Unique 45
<input type="checkbox"/>	<input type="checkbox"/>	stoptime	ABC string	100% Valid	 Unique 422
<input type="checkbox"/>	<input type="checkbox"/>	start station id	# number	100% Valid	 Unique 310
<input type="checkbox"/>	<input type="checkbox"/>	start station name	ABC string	100% Valid	 Unique 310
<input type="checkbox"/>	<input type="checkbox"/>	latlongmerge	ABC string	100% Valid	 Unique 310

Column details

ABC String starttime_2

The statistics below are only on the sample data.

Column statistics

Recommendations

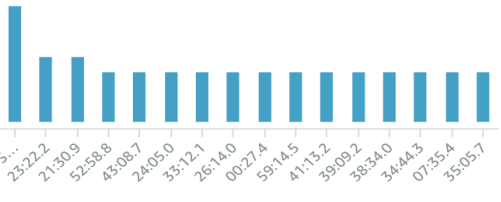
Data quality

VALID VALUES

MISSING VALUES

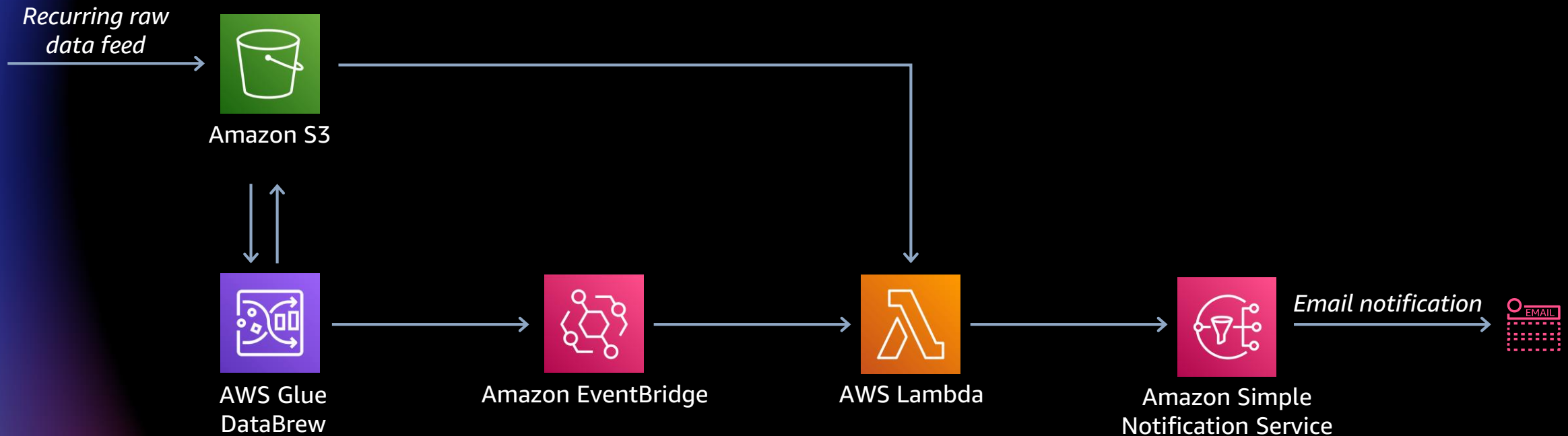
45 100% 0 0%

Value distribution

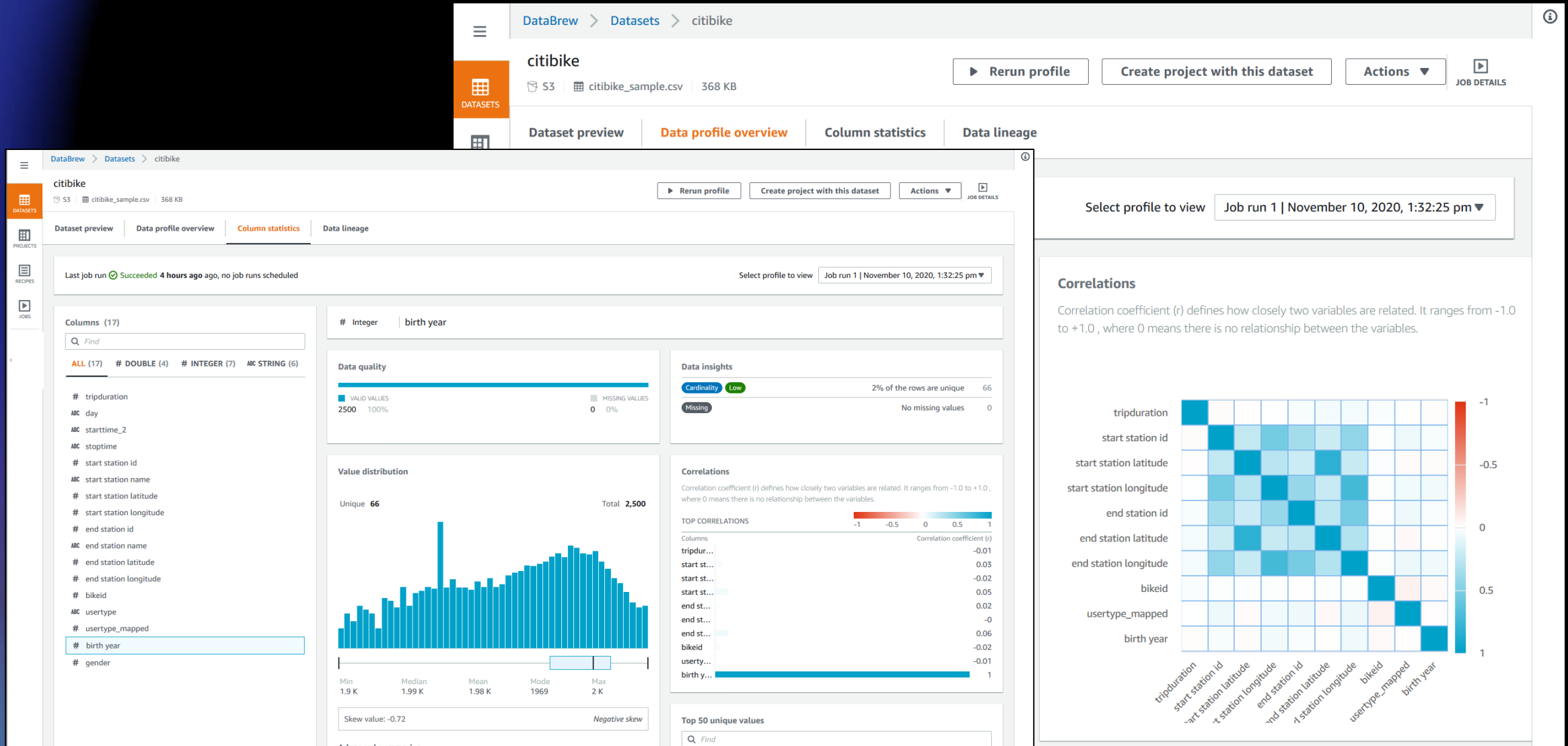


Unique values

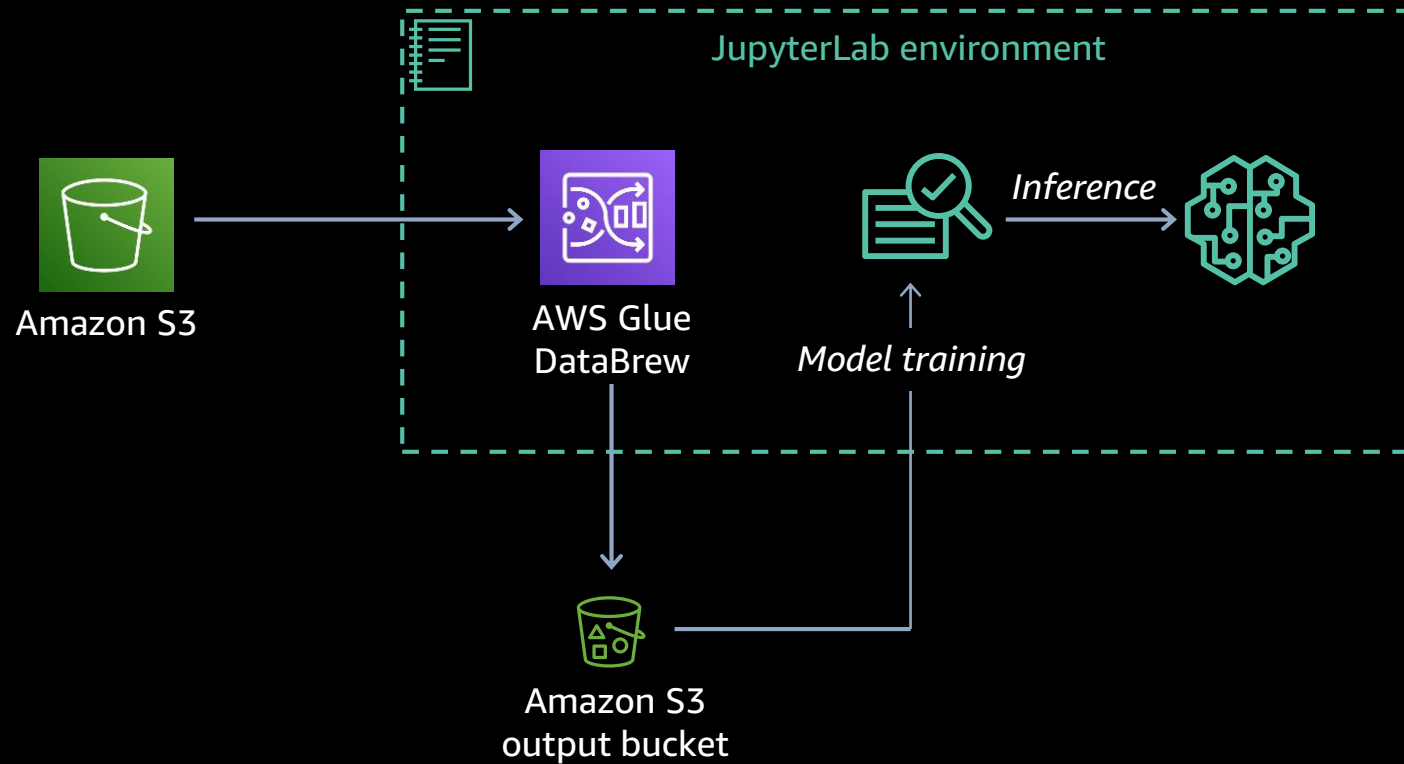
Set up data quality rules with AWS Lambda



Generate data profiles



Data preprocessing for machine learning



Remove outliers

Netflix

Dataset: Netflix titles | Sample: First n sample (500 rows)

Create job

LINEAGE

ACTIONS

DATASETS

PROJECTS

RECIPES

JOBS

UNDO REDO

FILTER COLUMN

FORMAT CLEAN EXTRACT

MISSING INVALID DUPLICATES

SPLIT MERGE CREATE

FUNCTIONS

UNNEST PIVOT GROUP JOIN UNION

TEXT SCALE MAPPING ENCODE

1 RECIPE

Viewing 12 columns 500 rows View highlighted

SAMPLE

GRID

SCHEMA

PROFILE

Column details

duration		listed_in	
Total 30	Unique 13	Total 30	Unique 2
13 43.33%	1 Season	14 46.67%	Children & Family Movies
9 30%	58 min	2 6.67%	Kids' TV
3 10%	65 min	2 6.67%	
5 16.67%	All other values	12 40%	
	90-min		Children & Family Movies, Comedies
	94-min		Stand-Up Comedy
	1 Season		Kids' TV
	1 Season		Kids' TV
	99-min		Comedies
	1 Season		Crime TV Shows, International TV Shows,
	110-min		International Movies, Sci-Fi & Fantasy, Th
	60-min		Stand-Up Comedy
	1 Season		Docuseries, Science & Nature TV
	90-min		Action & Adventure, Thrillers
	78-min		Stand-Up Comedy
	95-min		Action & Adventure, Dramas, internation
	58 min		Children & Family Movies
	62 min		Children & Family Movies
	65 min		Children & Family Movies
	61 min		Children & Family Movies
	65 min		Children & Family Movies
	58 min		Children & Family Movies

Filter values

Filter condition

Is exactly

Enter custom value Enter regex value

Enter a filter value

Children & Family Movies Kids' TV

Find

☒ Children & ...

16 3%

☐ Documenta...

16 3%

☐ Dramas, Ind...

15 3%

☒ Kids' TV

14 2%

Filtered 2/135 values

Results 30 rows

Clear filter

Add to recipe

Filter values

Statistics

Recommendations

ality

values 0 0%

values 0 0%

values 500 100%

10 of 135 unique values

tar... 32 6%

p C... 28 5%

Dramas, In... 27 5%

Children &... 16 3%

Documentar... 16 3%

Dramas, In... 15 3%

Kids' TV 14 2%

Zoom 100%

Use with Jupyter Notebooks

The screenshot displays the AWS Glue DataBrew console interface. The top navigation bar includes options like File, Edit, View, Run, Kernel, Tabs, Settings, and Help. The main workspace is divided into several sections:

- Left Sidebar:** Contains navigation icons and a section titled "AWS GLUE DATABREW" with a link to "Launch AWS Glue DataBrew".
- Top Bar:** Features tabs for Datasets, Projects (active), Recipes, Jobs, and Community. A "Create job" button is also present.
- Main Workspace:**
 - Dataset Information:** Shows the dataset name "test", its source "Dataset: Test", and a sample of "First n sample (500 rows)".
 - Tools Bar:** Includes icons for Undo, Redo, Filter, Column, Format, Clean, Extract, Missing, Invalid, Duplicates, Split, Merge, Create, Functions, Unnest, Pivot, Group, Join, Union, Text, Scale, Mapping, Encode, and Recipe.
 - Data Preview:** Displays a table with 17 columns and 500 rows. The columns are: neighbourhood, latitude, longitude, room_type, and price. The table is filtered to show the first 500 rows.

The data preview table shows the following columns and their corresponding data:

neighbourhood	latitude	longitude	room_type	price
Williamsburg	40.59	-74.09	Entire home/apt	291
East Village	40.72	-73.96	Private room	205
Harlem	40.73	-73.96	Shared room	4
All other values	40.68	-73.96	Private room	149
Kensington	40.68	-73.96	Entire home/apt	225
Midtown	40.70	-73.96	Private room	150
Harlem	40.71	-73.96	Entire home/apt	89
Clinton Hill	40.72	-73.96	Entire home/apt	80
East Harlem	40.73	-73.96	Entire home/apt	200
Murray Hill	40.74	-73.96	Private room	60
Bedford-Stuyvesant	40.75	-73.96	Private room	79
Hell's Kitchen	40.76	-73.96	Private room	79
Upper West Side	40.77	-73.96	Entire home/apt	150
Chinatown	40.78	-73.96	Entire home/apt	135
Upper West Side	40.79	-73.96	Private room	85
Hell's Kitchen	40.80	-73.96	Private room	89
South Slope	40.81	-73.96	Private room	85
Upper West Side	40.82	-73.96	Entire home/apt	120
West Village	40.83	-73.96	Entire home/apt	140
Williamsburg	40.84	-73.96	Entire home/apt	215
Fort Greene	40.85	-73.96	Private room	140
Chelsea	40.86	-73.96		

Feature engineering

Netflix

Dataset: Netflix titles | Sample: First n sample (500 rows)

Create job

LINEAGE

ACTIONS

UNDONE REDO

FILTER COLUMN

FORMAT CLEAN EXTRACT

MISSING INVALID DUPLICATES

SPLIT MERGE CREATE

FUNCTIONS

UNNEST PIVOT GROUP JOIN UNION

TEXT SCALE MAPPING ENCODE

Viewing 13 columns 500 rows View highlighted

SAMPLE

GRID SCHEMA PROFILE

SOURCE			PREVIEW		
ABC cast			ABC Starring Kate Hudson?		
Total	352		Total	445	
Unique	438		Unique	3	
148	29.6%		444	88.8%	
5	1%		55	11%	
4	0.8%		1	0.2%	
343	68.6%				
All other values			All other values		
438			234		
87.6%			46.8%		
Alan Marriott, Andrew Toth, Brian Dobson, Cole Ho...			United States, India, South Korea, China		
Jandino Asporaat			United Kingdom		
Peter Cullen, Sumalee Montano, Frank Welker, Jeff...			United States		
Will Friedle, Darren Criss, Constance Zimmer, Khar...			United States		
Nesta Cooper, Kate Walsh, John Michael Higgins, K...			United States		
Alberto Ammann, Eloy Azorín, Verónica Echegui, L...			Spain		
Antonio Banderas, Dylan McDermott, Melanie Gri...			Bulgaria, United States, Spain, Canada		
Fabrizio Copano			Chile		
null			United States		
James Franco, Kate Hudson, Tom Wilkinson, Omar ...			United States, United Kingdom, Denmark, Sweden		
Joaquín Reyes			null		
Jim Sturgess, Sam Worthington, Ryan Kwanten, An...			Netherlands, Belgium, United Kingdom, United Sta...		
Damandeep Singh Baggan, Smita Malhotra, Baba ...			null		
Damandeep Singh Baggan, Smita Malhotra, Baba ...			null		
Damandeep Singh Baggan, Smita Malhotra, Deepa...			null		
Damandeep Singh Baggan, Smita Malhotra, Baba ...			null		
Rishi Gambhir, Smita Malhotra, Deepak Chachra			null		

Zoom 100%

Create column Info

Create a new column by some transforms

Create column options

Flag values

Source column

Select column to extract values from

cast

Values to flag

☐ Missing values

☒ Custom value

Values to flag

Kate Hudson

Enter string value or regex expression

Flag values as

Yes or no

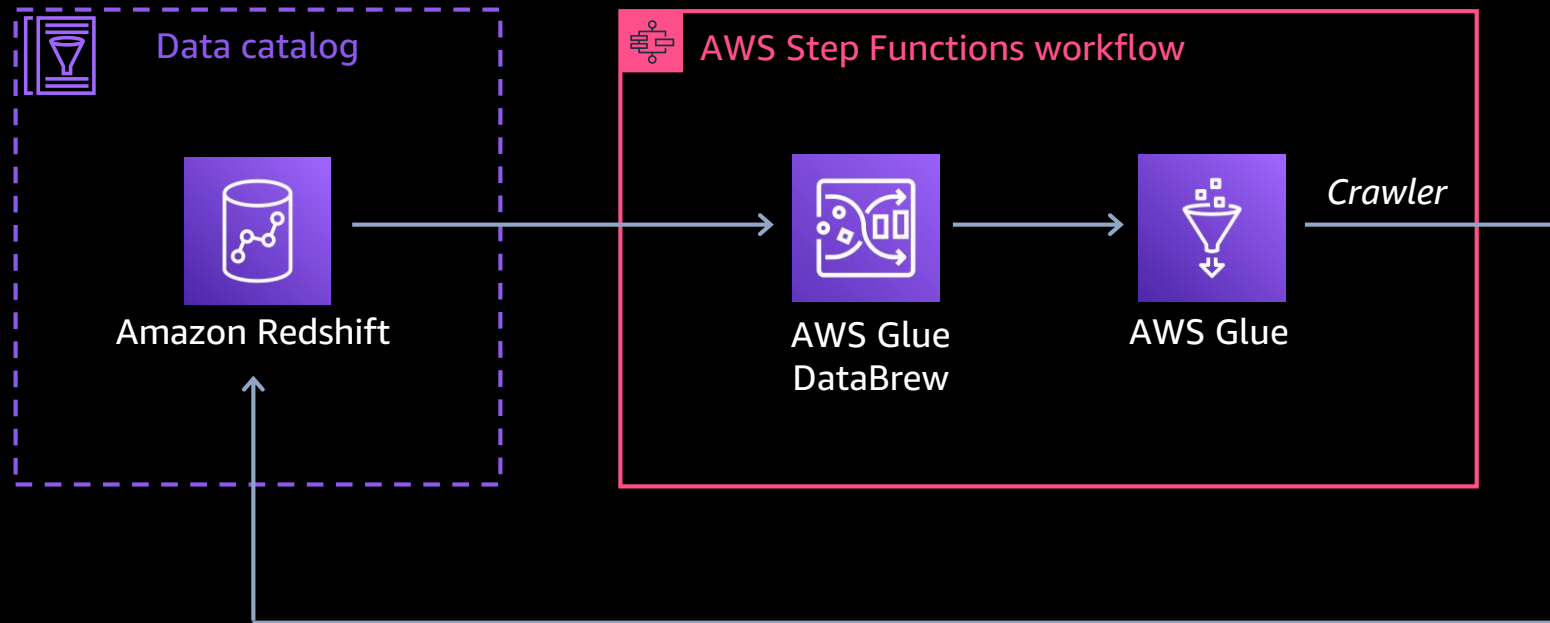
Destination column

Name of the column created with extracted values

Starring Kate Hudson?

Valid characters are alphanumeric, underscore, and space

Orchestrating data preparation in workflows



Recap

- Challenges with traditional data preparation methods
- Saw AWS Glue DataBrew in action with a demo
- Popular use cases of AWS Glue DataBrew
- ~~What is AWS Glue DataBrew?~~ >> AWS Glue DataBrew is so cool!

Additional resources

- [AWS Glue Databrew features](#)
- [AWS Glue DataBrew jupyter extension](#)
- [7 most common data preparation transformations in AWS Glue DataBrew](#)
- [Orchestrating an AWS Glue DataBrew job and Amazon Athena query with AWS Step Functions](#)
- [Setting up automated data quality workflows and alerts using AWS Glue DataBrew and AWS Lambda](#)
- [Build a data quality score card using AWS Glue DataBrew, Amazon Athena, and Amazon QuickSight](#)
- [AWS Step Functions Workflow Studio – A Low-Code Visual Tool for Building State Machines](#)

Visit the AWS Data Resource Hub

Dive deeper with these resources, get inspired and learn how you can use data to make better decisions and innovate faster.

- Building a winning data strategy
- The new leadership mindset for data & analytics
- Harness data to reinvent your organization
- Put your data to work with a modern analytics approach
- Breaking free from on-premises database constraints
- Cloud storage adoption: From cost optimization to agility & innovation
- A strategic playbook for data, analytics, and machine learning
- ... and more!



<https://tinyurl.com/aws-data-resource>

Visit resource hub



AWS Training and Certification

Empower your teams with comprehensive training

By building skills with AWS Training and Certification, businesses and individuals can see the bigger picture understanding the reasoning behind every data point. As training progresses and teams become data-fluent, previously hidden insights come into view.

Build data skills to
unlock any insight

Leverage free digital training

Learn how to harness the world's most valuable resource: data. Access digital and virtual instructor-led courses on data analytics and databases built by the experts at AWS and start your learning journey to become data-driven.

[Take a digital course »](#)



Get certified

Earn industry-recognized credibility and set tangible goals for success with industry-recognized certifications, like *AWS Certified Data Analytics – Specialty*.

[Learn more »](#)



Ramp-up your skills

Deep dive into new topics and focus on knowledge gaps at your own pace with the *AWS Ramp-Up Guide: Database* and *AWS Ramp-Up Guide: Data Analytics*. With a wide range of whitepapers, blog posts, videos, webinars and peer resources available for data professionals to leverage for independent learning.

[Download ramp-up guides »](#)

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey**.
Let us know what you thought of today's event and how we can improve the event experience for you in the future.



aws-apj-marketing@amazon.com



twitter.com/AWSCloud



facebook.com/AmazonWebServices



youtube.com/user/AmazonWebServices



slideshare.net/AmazonWebServices



twitch.tv/aws

Thank you!