SIMPLIFY DATA TRANSFORMATION FOR ANALYTICS / ML TO UNLOCK INSIGHTS

Wendy Wong

AWS Data Community Builder

AGENDA

- Introduction
- Synopsis
- Use Cases
- High Level Architecture
- Getting Started
- What’s New?

Lesson Objectives

In this lesson you will:

- Create IAM permissions for AWS services
- Use AWS Glue and AWS Glue Studio to prepare your data for analytics
- Create an external table using AWS Athena
- Unlock business insights using Amazon QuickSight

INTRODUCTION

About Me

- Data Scientist, Senior Data Analyst, Business Analyst, AWS Community Builder
- Lead instructor data analytics at General Assembly
- Former data analytics instructor at Academy Xi
- Former Senior Consultant and Digital Accelerator at PwC Digital Academy
- Stanford University Women in Data Science Sydney Ambassador in 2018 and 2020
- Director of Women in Big Data Sydney Chapter in 2020
- Editor at Towards Data Science 2017-2019

pwc

GA

Academy Xi

WOMEN IN BIG DATA

WOMEN IN DATA SCIENCE STANFORD UNIVERSITY

AWS community builders

Synopsis

- Real world data is very dirty - from startups, consulting, finance, higher education and government. What's your first use case?
- The journey to cloud, a long transformation project demonstrating the different data analytics life cycle maturity of customers from **descriptive**, **prescriptive** and **predictive**.
- If an organization does not have FTE for data engineers, data scientists and business analysts may complete ETL jobs. E.g. Alteryx, Power Query.

- The AWS services used to build a data solution included:

Amazon Simple Storage Service (Amazon S3)

AWS Identity and Access Management (IAM)

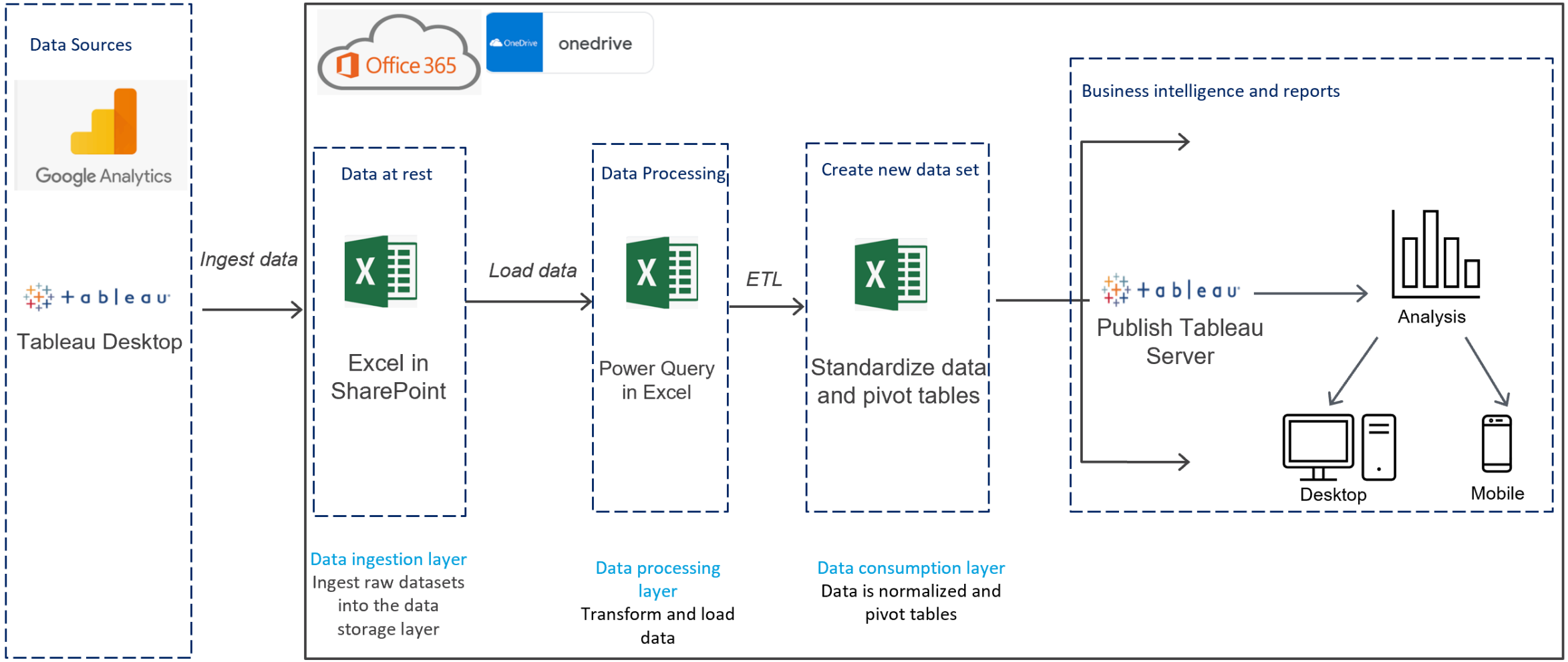
Amazon QuickSight

Amazon Athena

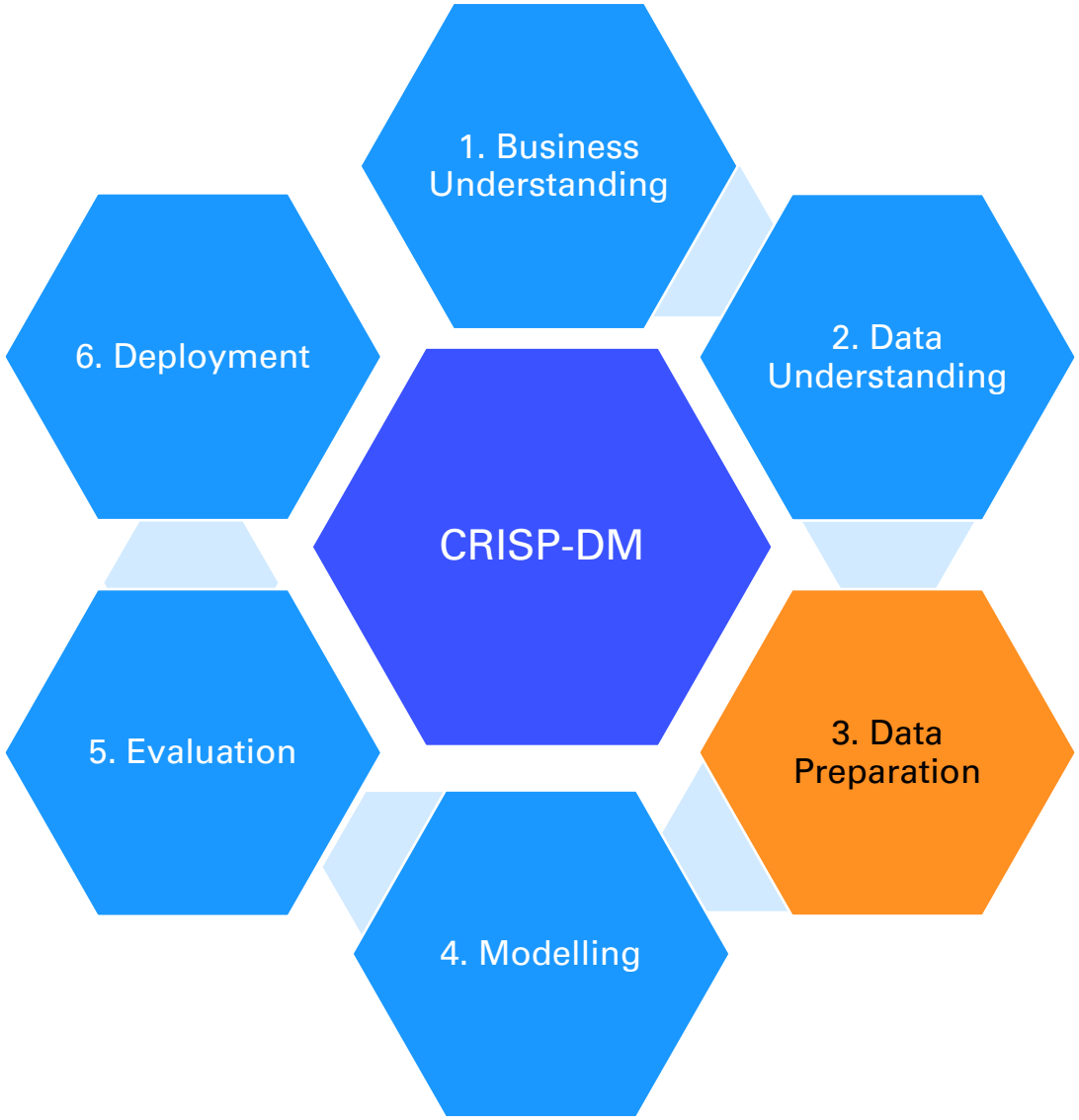
AWS Glue

So you want to be a data scientist

What is ETL?


Data Sources
Google Analytics
Tableau Desktop
Ingest data
Office 365
OneDrive
onedrive
Data at rest
Excel in SharePoint
Load data
Data Processing
Power Query in Excel
ETL
Create new data set
Standardize data and pivot tables
Business intelligence and reports
Publish Tableau Server
Analysis
Desktop
Mobile
Data ingestion layer
Ingest raw datasets into the data storage layer
Data processing layer
Transform and load data
Data consumption layer
Data is normalized and pivot tables

Data Analytics Workflow (CRISP - DM)

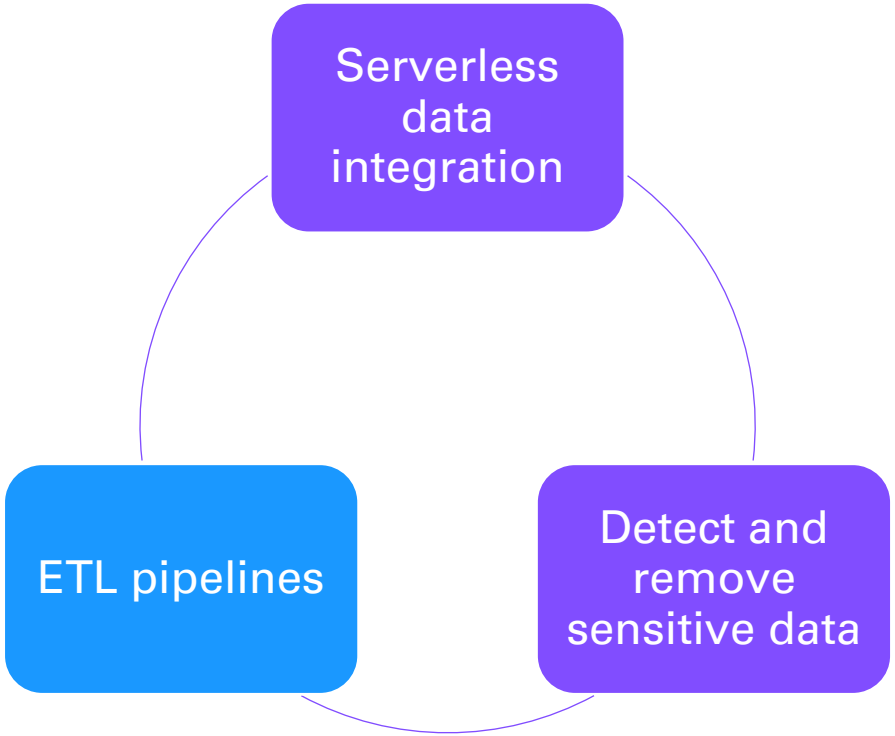


Data Preparation:

- ETL considerations
- Clean and normalize the data

AWS GLUE

What is Amazon Glue?



Languages

- Python, Scala

Benefits

- Catalog data for use in data lake and data warehouse
- Connect to multiple data sources
- You do not need to provision hardware
- **Schedule job** based on trigger event
- Data integration for analytics and ML
- Edit job scripts in AWS Glue Studio

Use Cases

AWS Glue

- Schedule jobs (on **demand** / trigger event)
- Glue Data Catalog
- Discover data properties
- S3, Redshift, AWS databases

Data Engineer

AWS Glue Studio

- Visual editor to create ETL workflows
- **No-code ETL jobs**
- Job monitoring
- Retries
- Generate scripts

Business Analyst,
Data Engineer

AWS Glue DataBrew

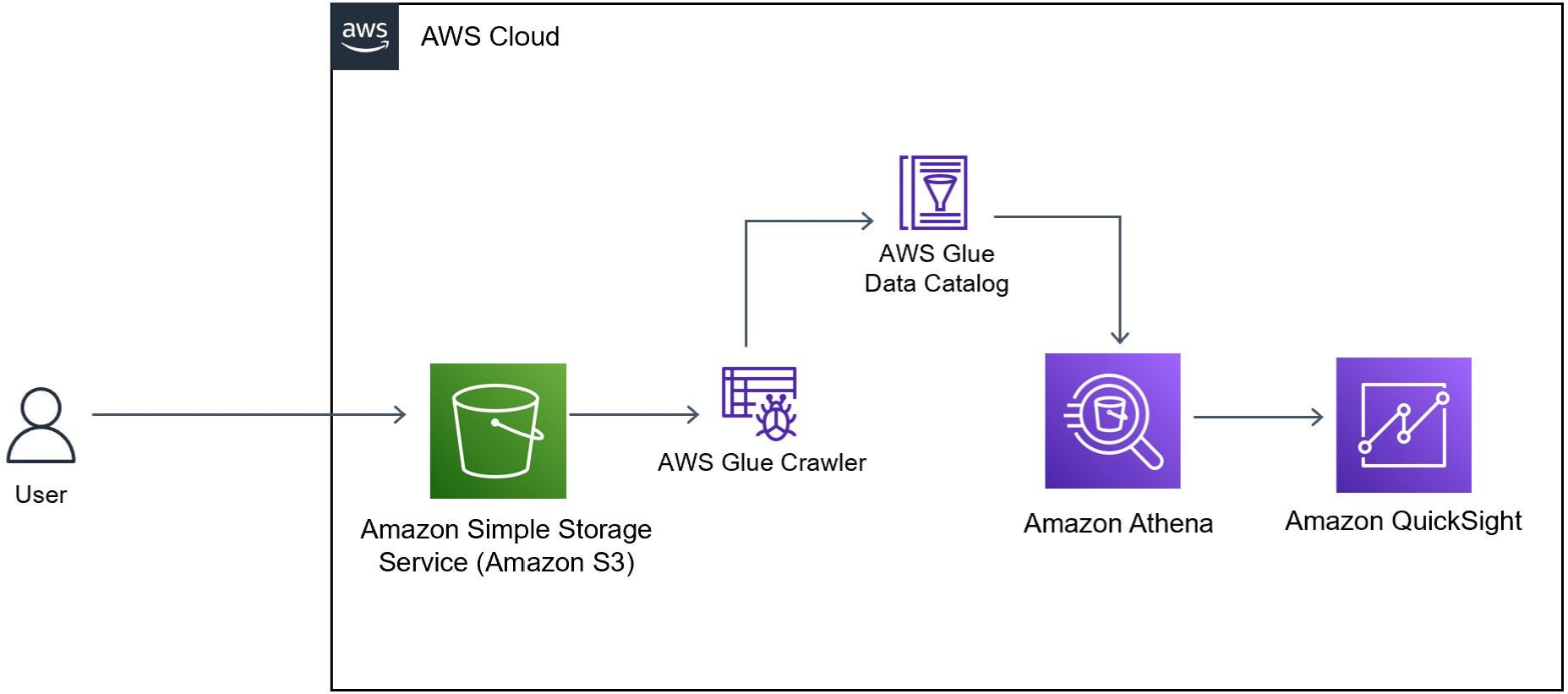
- Normalize data
- Inspect outliers
- Format data
- Create new variables

Data Scientist

HIGH LEVEL ARCHITECTURE

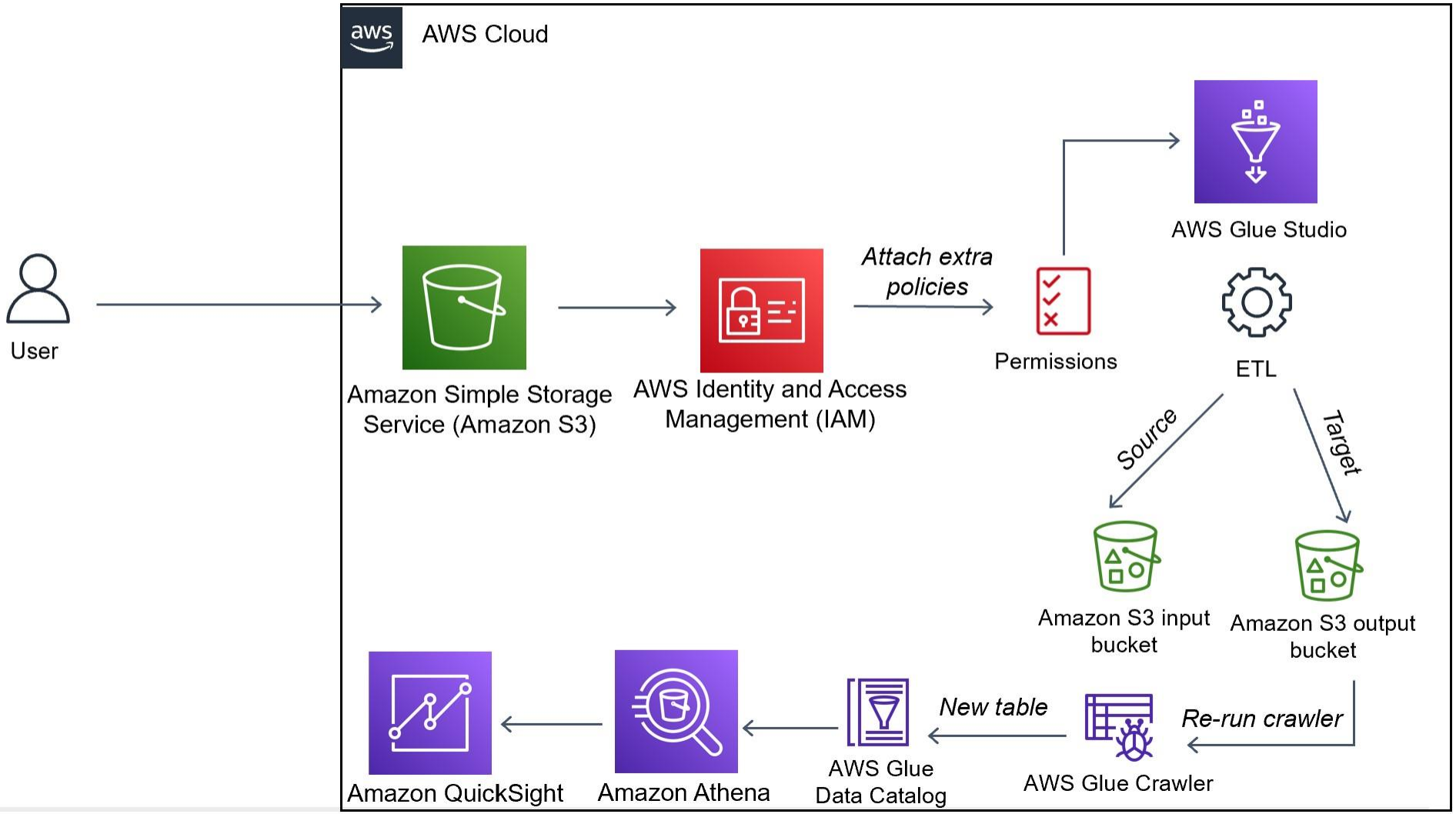
SCHEDULE AWS GLUE JOB

High Level Architecture – Use AWS Glue to crawl and catalog data



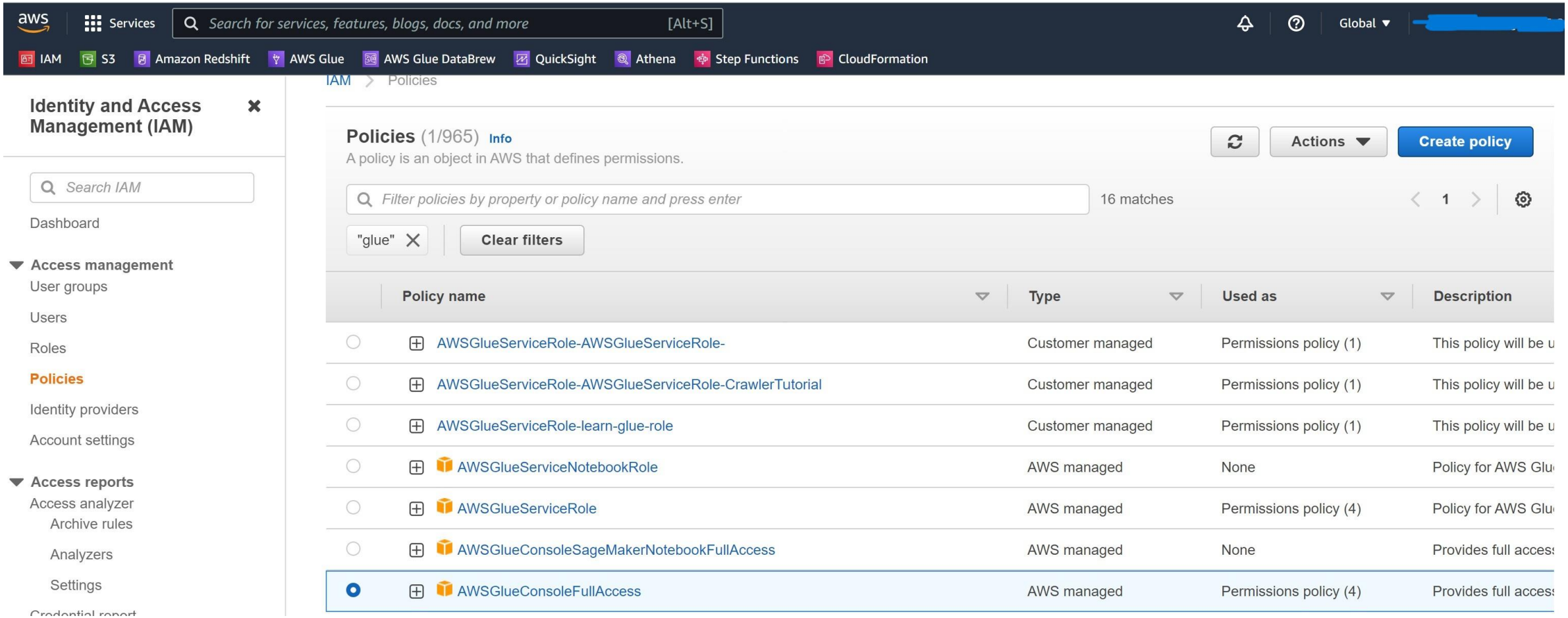
AWS GLUE INTERACTIVE

High Level Architecture – Use AWS Glue Studio to perform data ETL



1.ETL – AWS GLUE

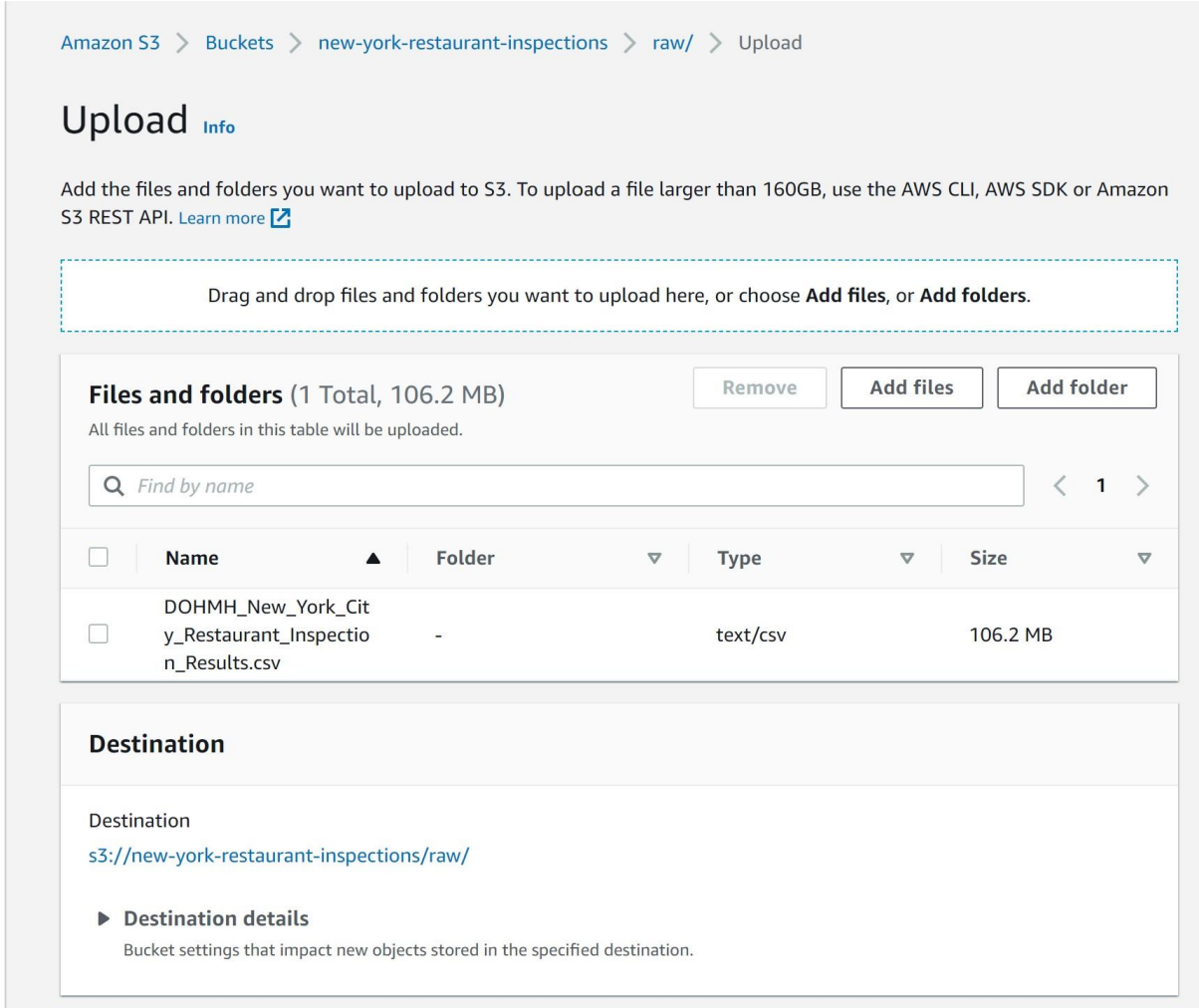
Getting started - IAM Permissions



- AWSGlueConsoleFullAccess and AmazonS3FullAccess

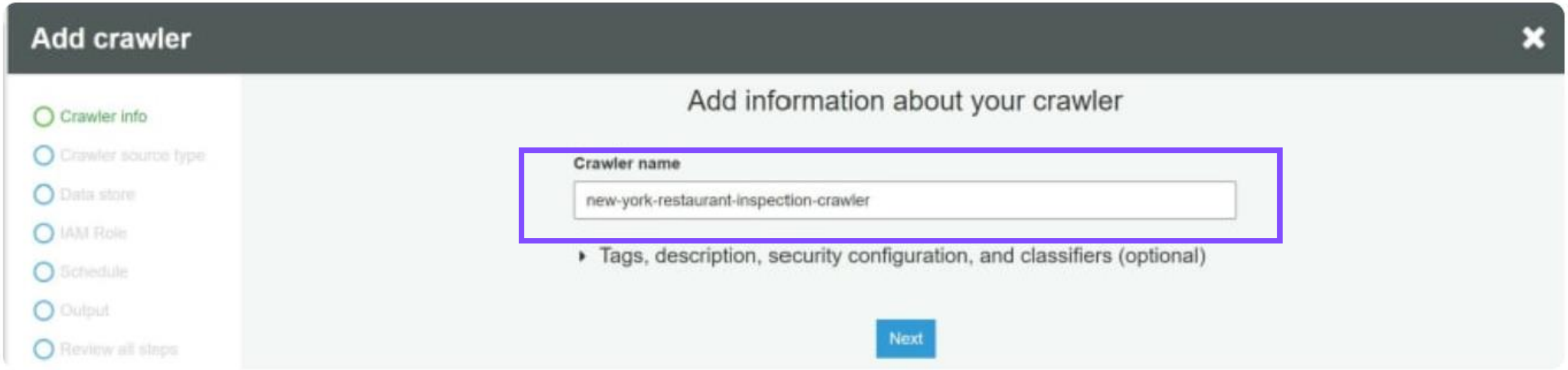
Upload data into Amazon S3

	A	B	C	D	E	F	G	H	I	J	K	L
1	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DE	INSPECTION	ACTION	VIOLATION	VIOLAT
2	41011076	YAKITORI S	Manhattan	2707	BROADWA	10025	2.13E+09	Japanese	02/14/201	Violations	04K	Eviden
3	41316819	HIDEAWAY	Manhattan	185	DUANE STF	10013	2.12E+09	American	03/15/202	Violations	04K	Eviden
4	50050712	PIG BEACH	Brooklyn	480	UNION STR	11231	9.17E+09	Barbecue	10/26/201	Violations	06D	Food c
5	41420319	OCEANA	Manhattan	120	WEST 49	10019	2.13E+09	Seafood	09/13/201	Violations	02G	Cold fo
6	50040939	WASAN BR	Brooklyn	440	BERGEN ST	11217	3.48E+09	Japanese	03/20/201	Violations	10B	Plumbi
7	50118108		Brooklyn	398	KNICKERBC	11237	9.18E+09		1/01/1900			
8	41227925	GOOD EAT	Queens	69-32	GRAND AV	11378	7.18E+09	American	5/09/2019	Violations	10F	Non-fo
9	50077587	WYCKOFF F	Brooklyn	250	WYCKOFF A	11237	3.47E+09	Mexican	05/30/201	Violations	10B	Plumbi
10	41706711	NOT GUILT	Staten Islar	19	HYATT STRI	10301	7.18E+09	Sandwiches	10/17/201	Establishm	06C	Food n
11	41210703	LOS POTRIL	Staten Islar	150	PORT RICHI	10302	7.19E+09	Mexican	07/19/202	Violations	08A	Facility
12	50003047	ELI ZABAR	Manhattan	922	MADISON A	10021	6.47E+09	American	03/17/202	Violations	02G	Cold fo
13	50076675	ELIM BISTR	Manhattan	11	PARK PLAC	10007	2.13E+09	Sandwiches	07/23/201	Violations	10B	Plumbi
14	50101465	NEW CHAN	Bronx	2570	BRONXWO	10469	7.18E+09	Pakistani	8/03/2022	Violations	10B	Anti-si
15	50101465	NEW CHAN	Bronx	2570	BRONXWO	10469	7.18E+09	Pakistani	8/03/2022	Violations	10B	Anti-si
16	41579144	SALT AND F	Manhattan	139	WEST 33	10001	2.12E+09	Soups	08/16/202	Violations	04N	Filth fli
17	50059509	BARN JOO	Manhattan	35	UNION SQU	10003	6.46E+09	Korean	2/12/2018	Violations	10B	Plumbi
18	50104728	El Gallo Tac	Manhattan	369	BROOME S	10013	9.17E+09	Mexican	01/26/202	Violations	02B	Hot foo
19	50080649	TURKO'S GI	Brooklyn	110	MOORE STI	11206	9.18E+09	Mediterran	6/05/2019	No violatio	22C	Bulb no
20	50004456	JUSTINO'S I	Manhattan	77	PEARL STRE	10004	2.13E+09	Pizza	1/04/2019	Violations	10F	Non-fo
21	41168300	ORIGINAL N	Brooklyn	594	CRESCENT	11208	7.19E+09	Pizza	08/17/202	Violations	02G	Cold fo
22	50044246	FLEET BAKE	Manhattan	24	BOWERY	10013	6.46E+09	Bakery Pro	05/14/201	Violations	06B	Tobacc
23	50048697	DYCKMAN	Manhattan	100	DYCKMAN	10040	2.13E+09	American	08/29/201	Violations	02B	Hot foo

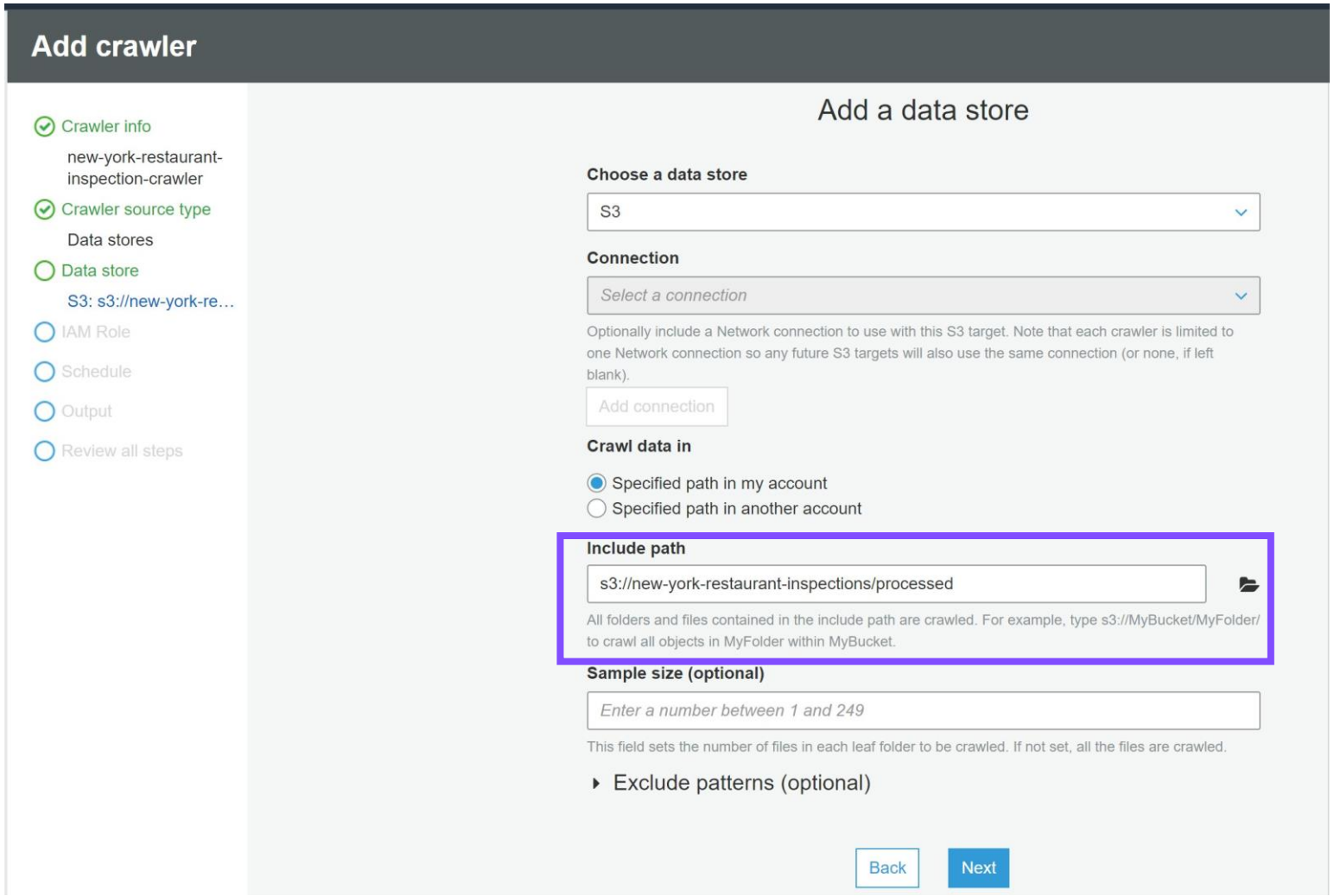


DOHMH New York City Restaurant Inspection Results | NYC Open Data (cityofnewyork.us)

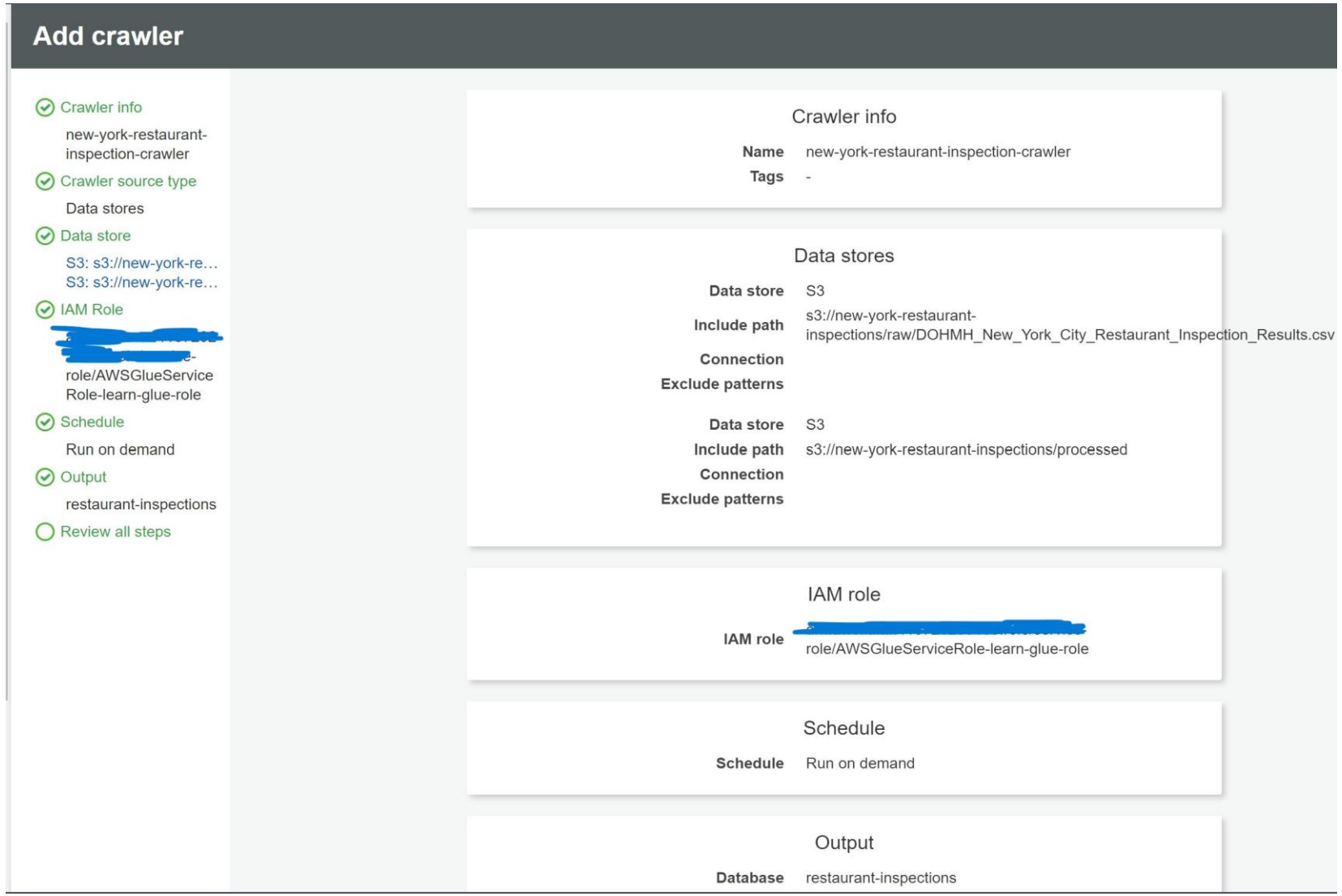
AWS Glue Crawler



AWS Glue Crawler

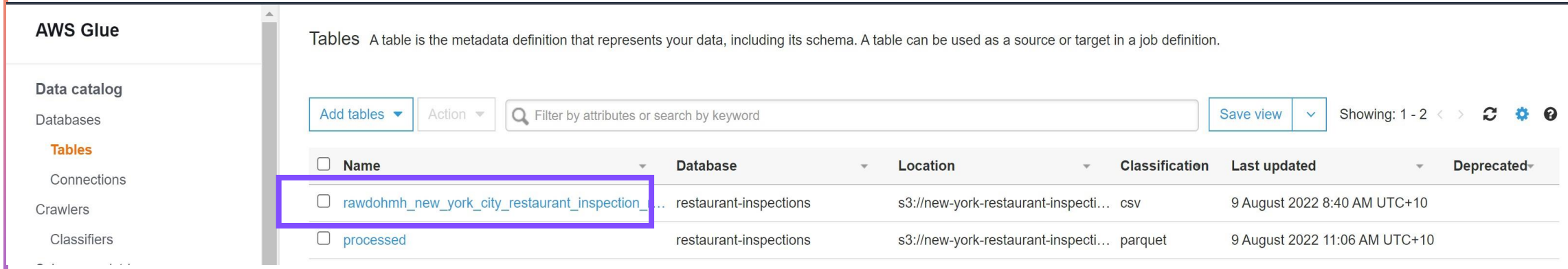


AWS Glue Crawler



Check crawler details

AWS Glue Data Catalog



AWS Glue Data Catalog

Schema

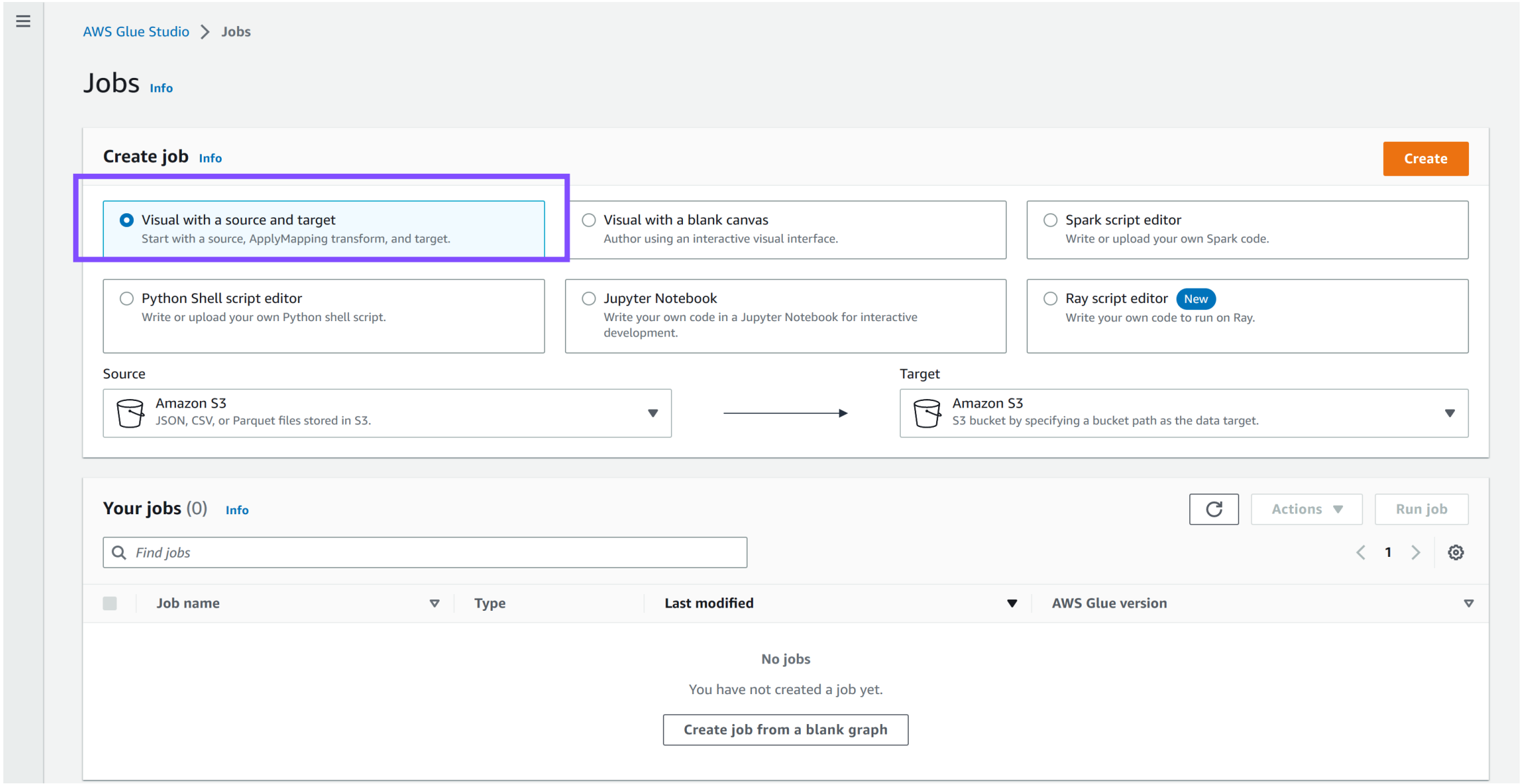
Showing: 1 - 26 of 26

	Column name	Data type	Partition key	Comment
1	camis	bigint		
2	dba	string		
3	boro	string		
4	building	string		
5	street	string		
6	zipcode	bigint		
7	phone	string		
8	cuisine description	string		
9	inspection date	string		
10	action	string		
11	violation code	string		
12	violation description	string		
13	critical flag	string		
14	score	bigint		
15	grade	string		
16	grade date	string		
17	record date	string		
18	inspection type	string		

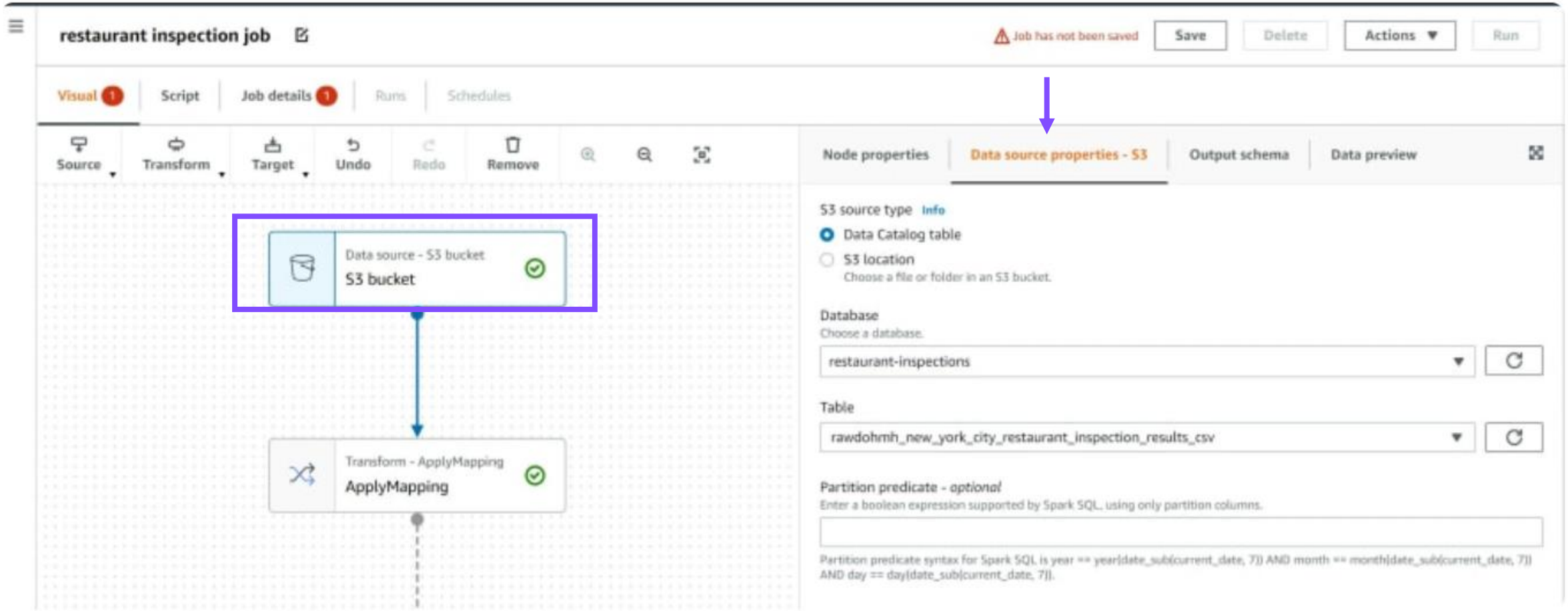
Edit Schema

2.ETL – AWS GLUE STUDIO

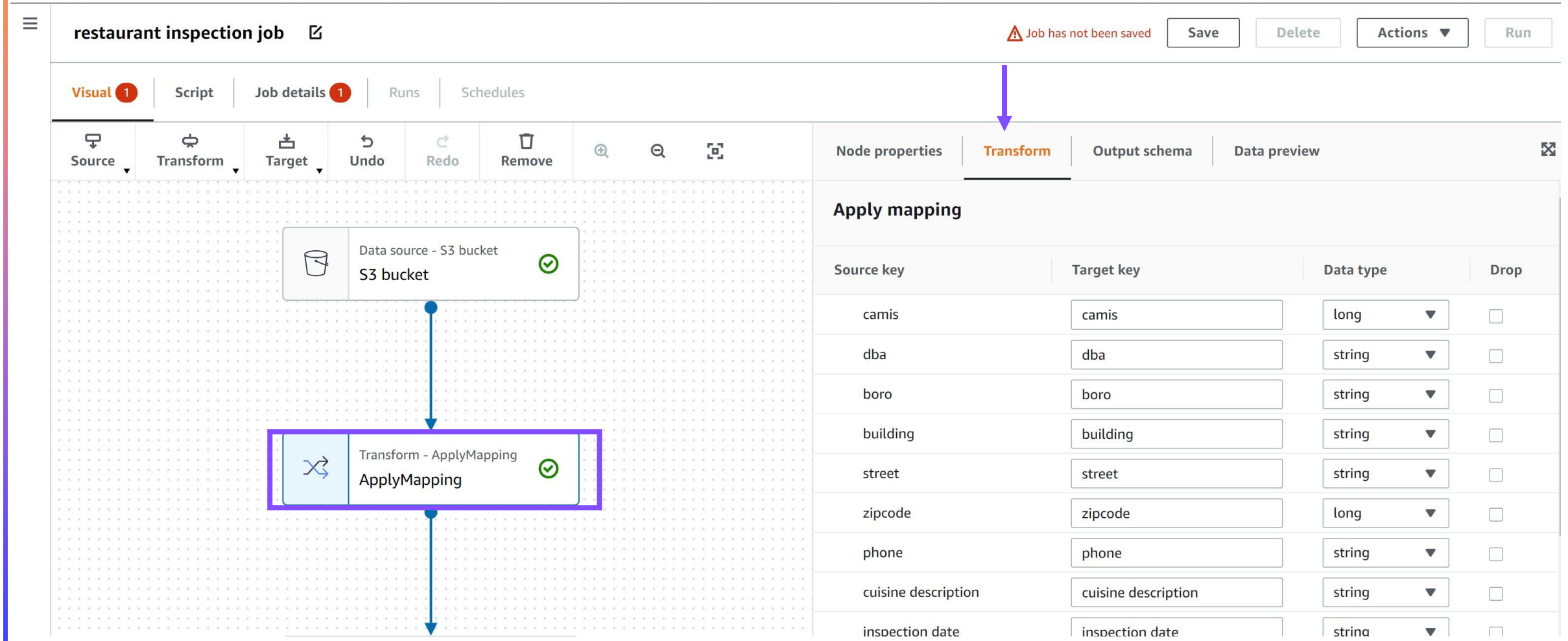
AWS Glue Studio



AWS Glue Studio

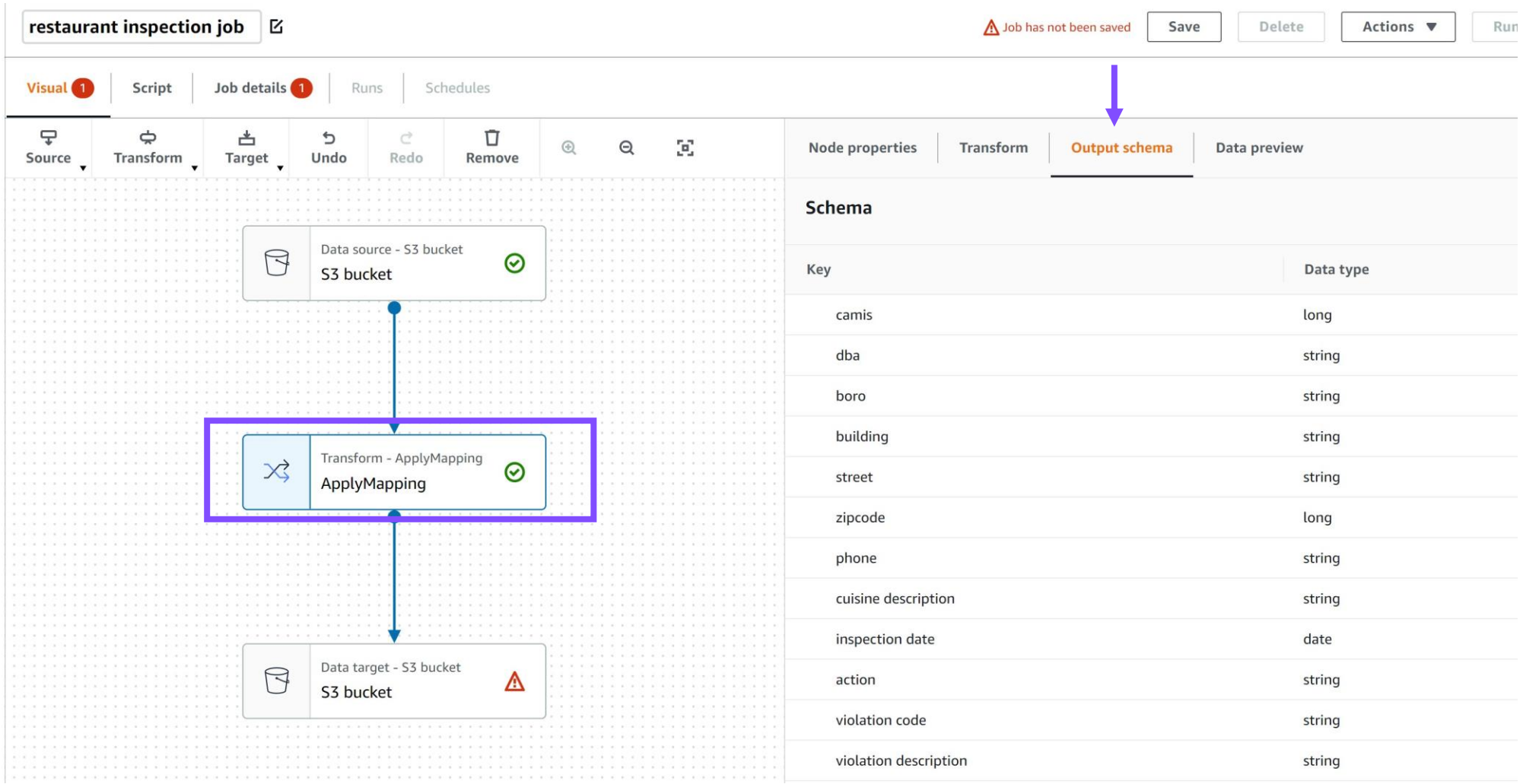


AWS Glue Studio

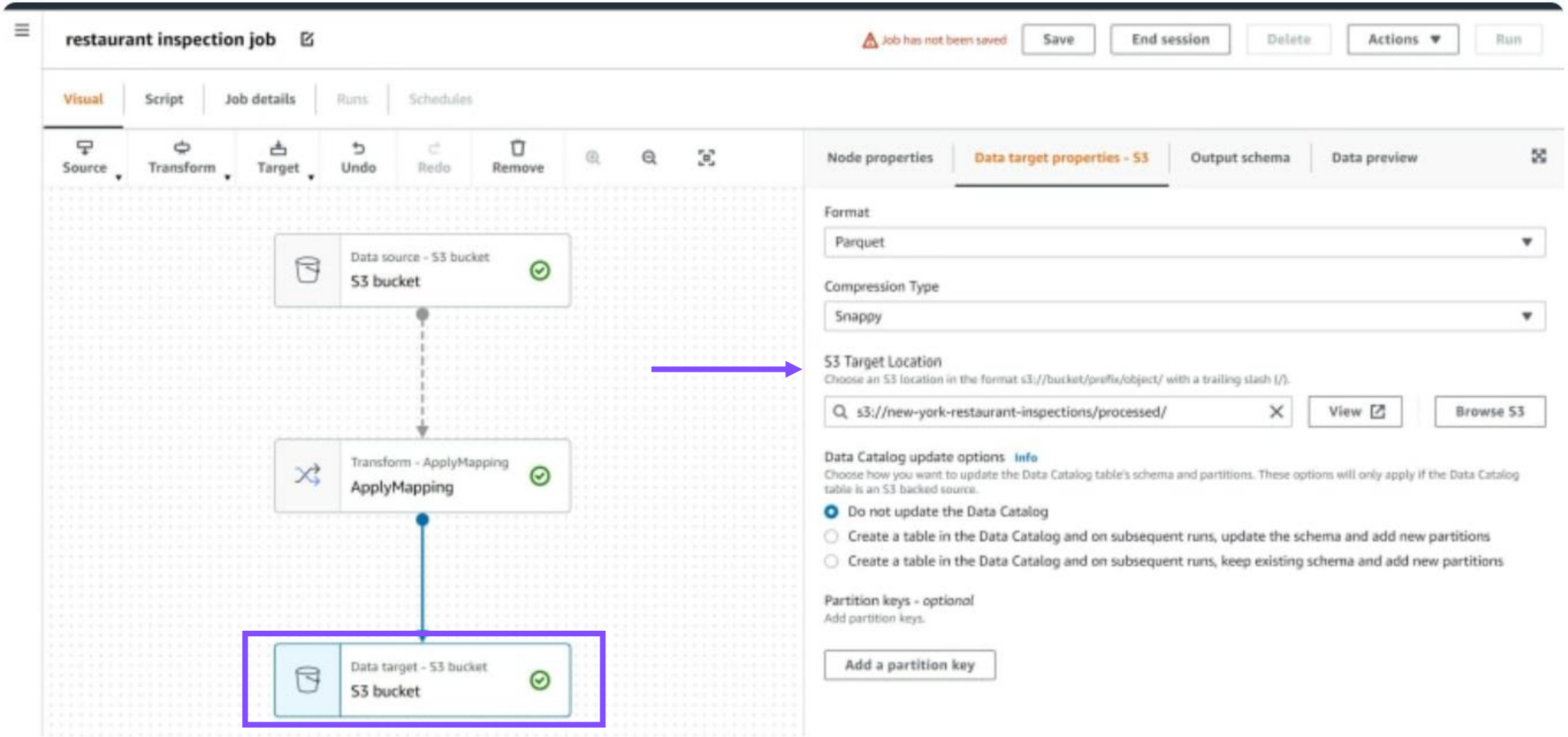


Edit Schema

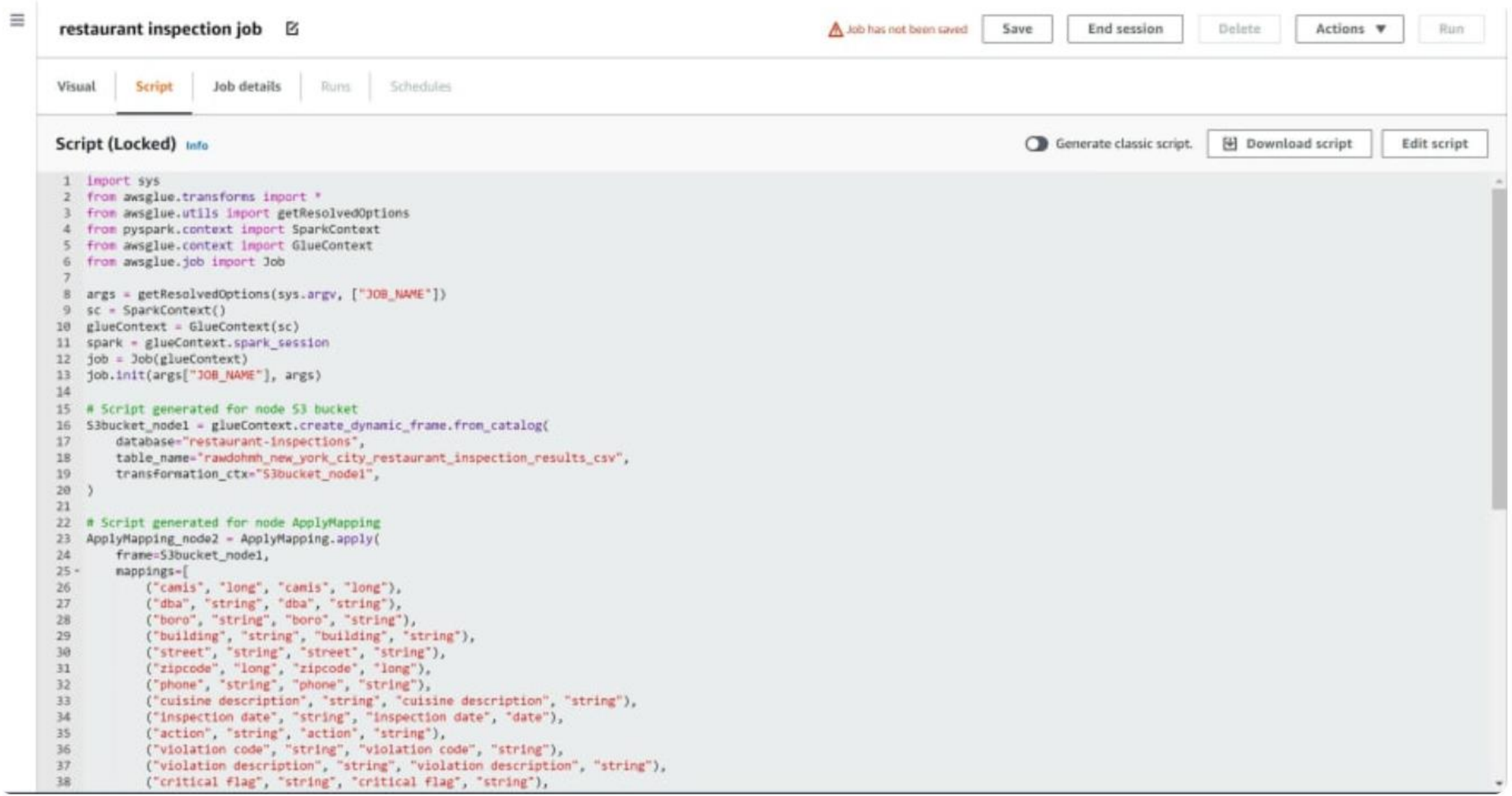
AWS Glue Studio



AWS Glue Studio

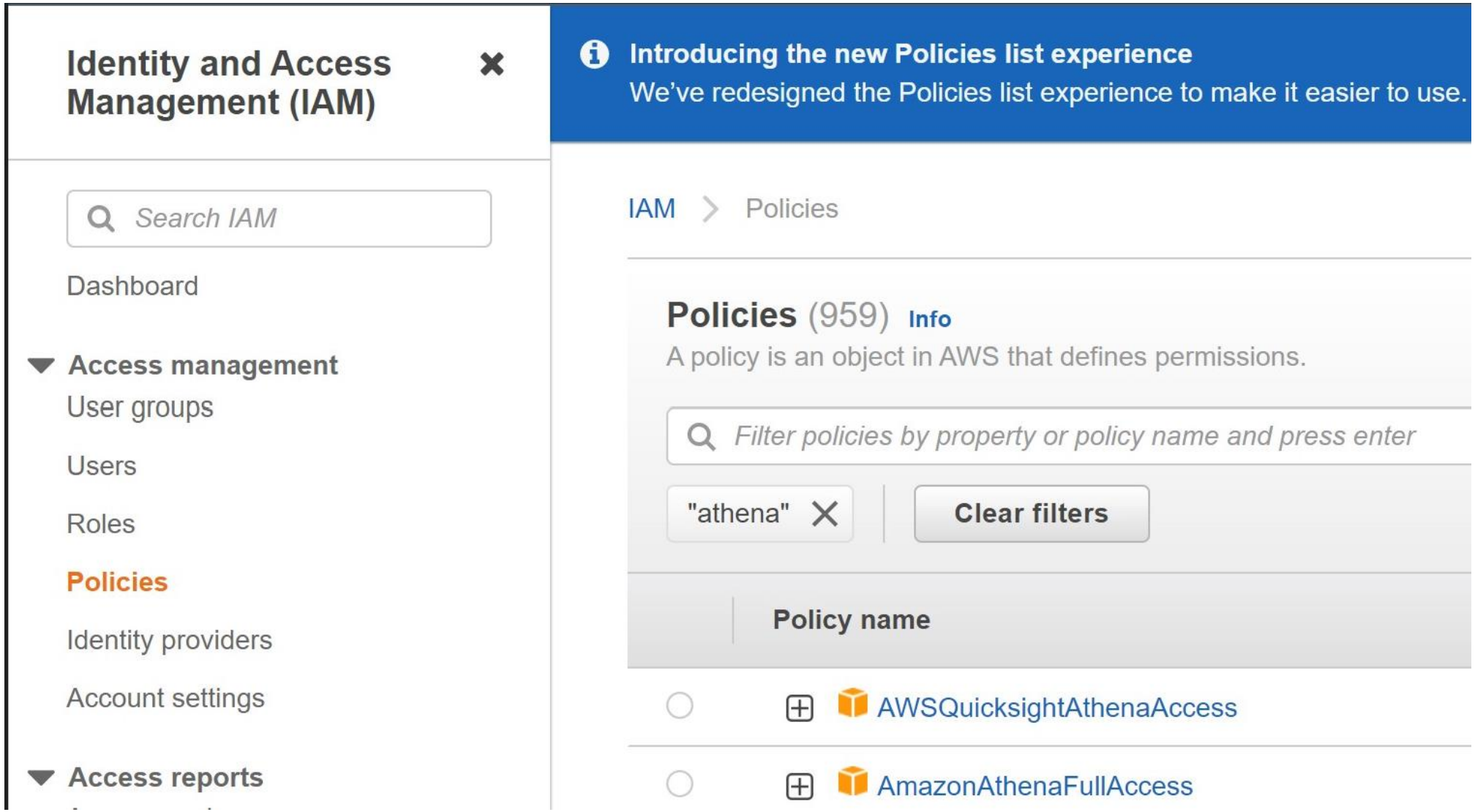


AWS Glue Studio

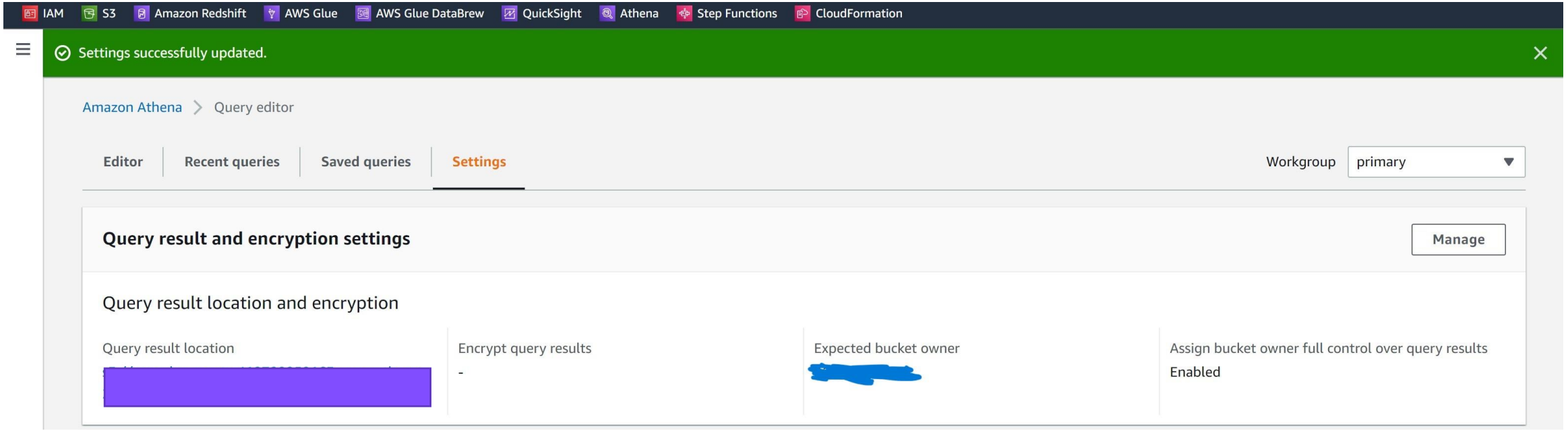


3.CREATE AN EXTERNAL TABLE WITH AMAZON ATHENA

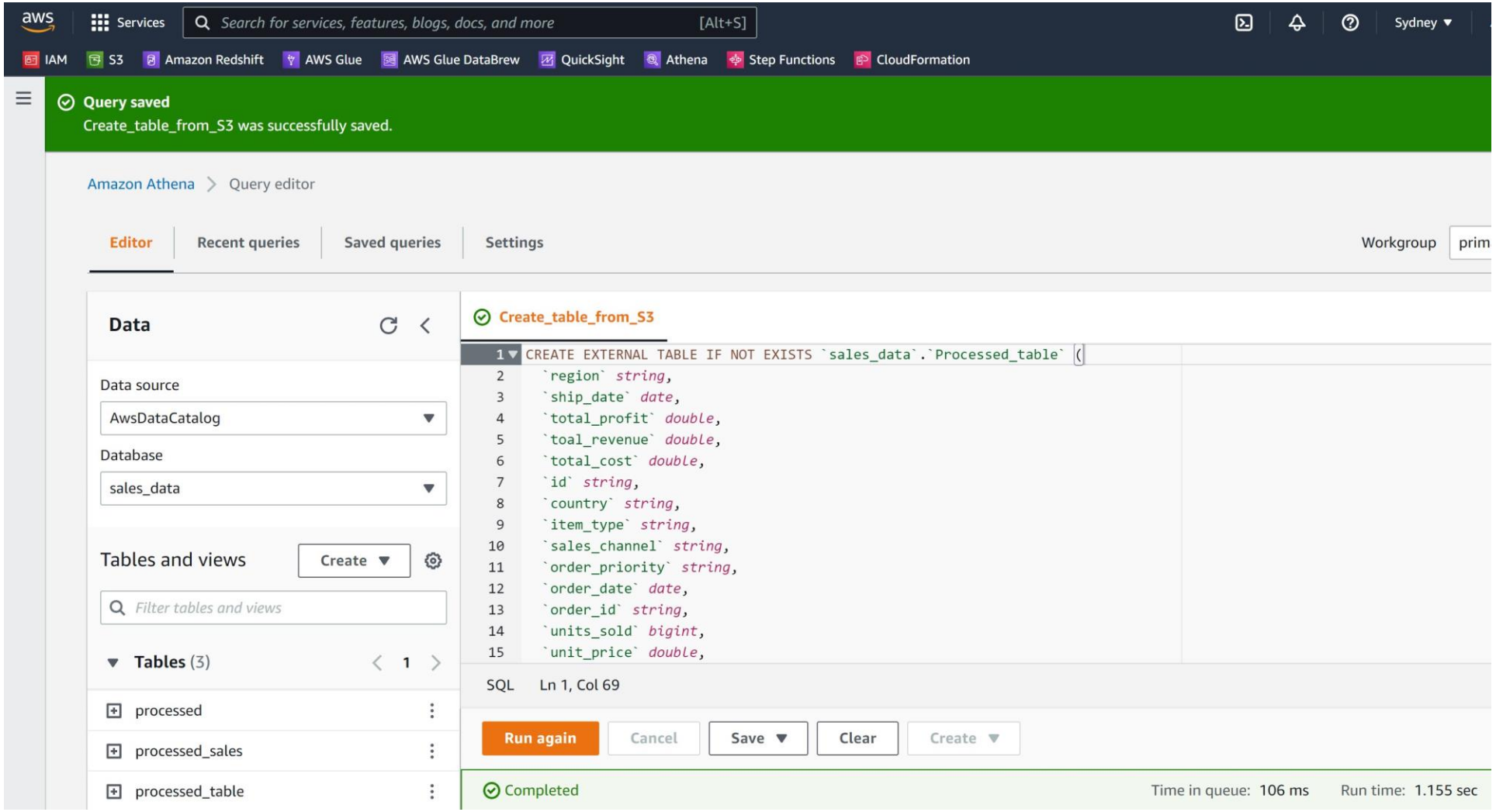
Getting started - IAM Permissions



Check S3 can access Amazon Athena

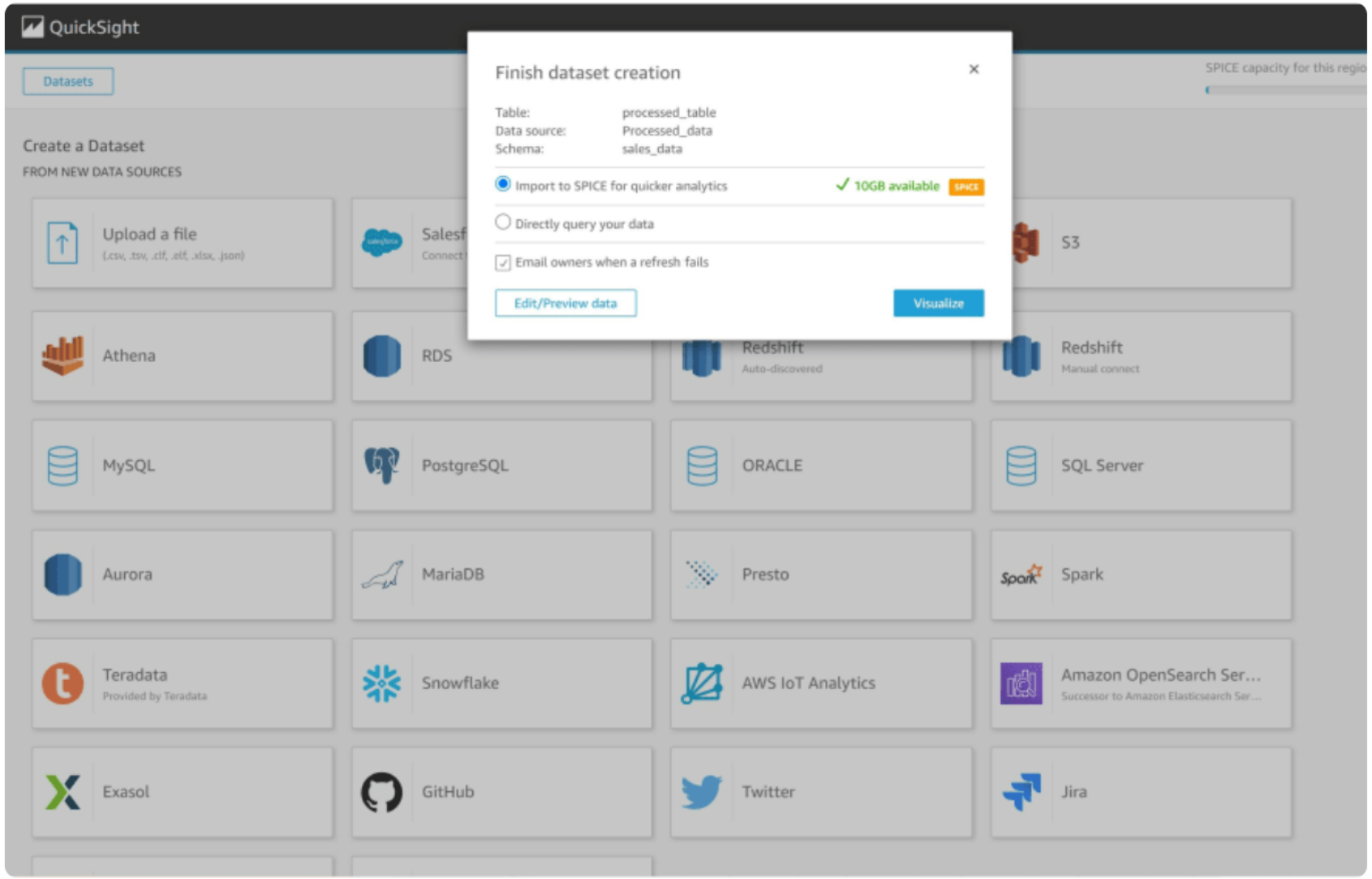


Create external table

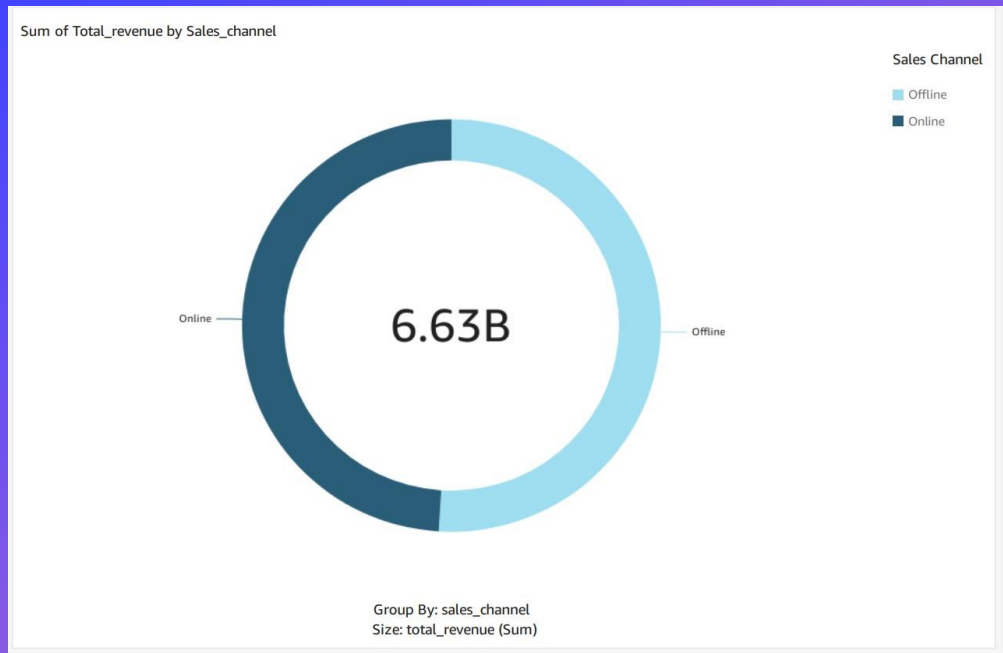


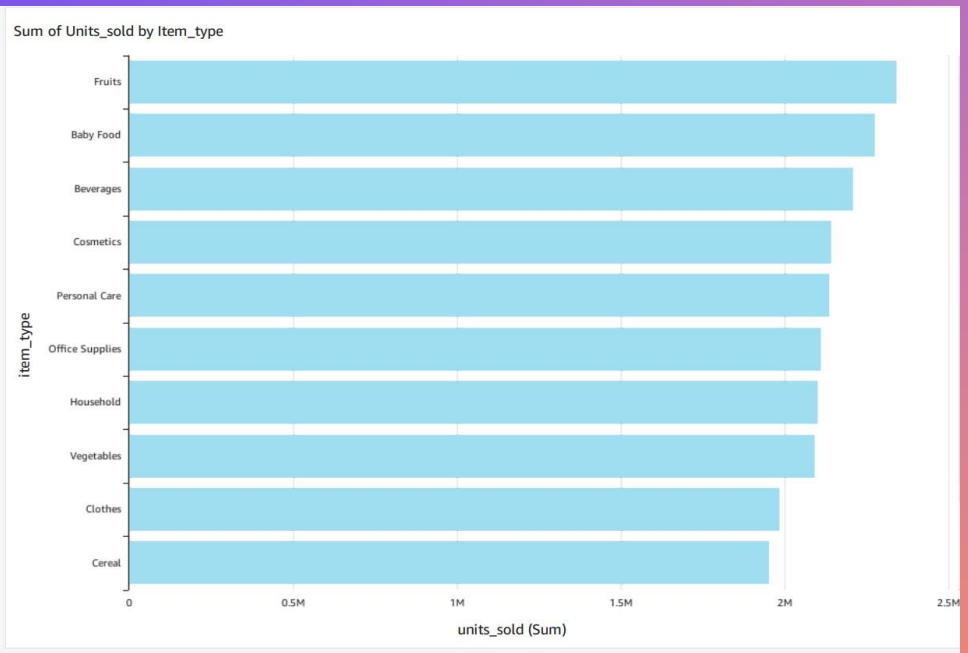
4. CREATE VISUALIZATION WITH AMAZON QUICKSIGHT

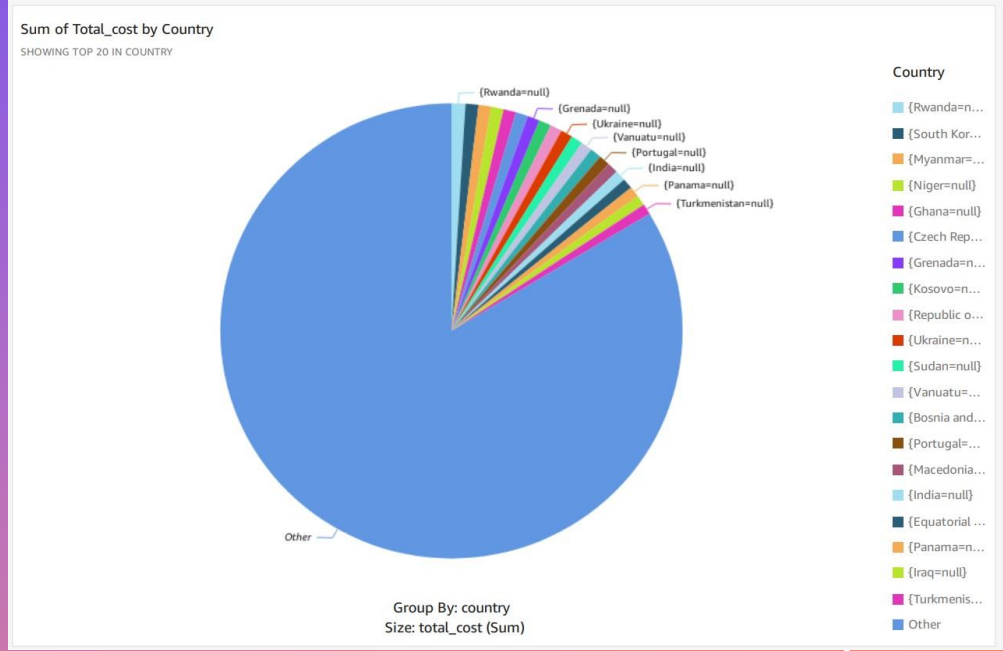
Upload data from Amazon Athena

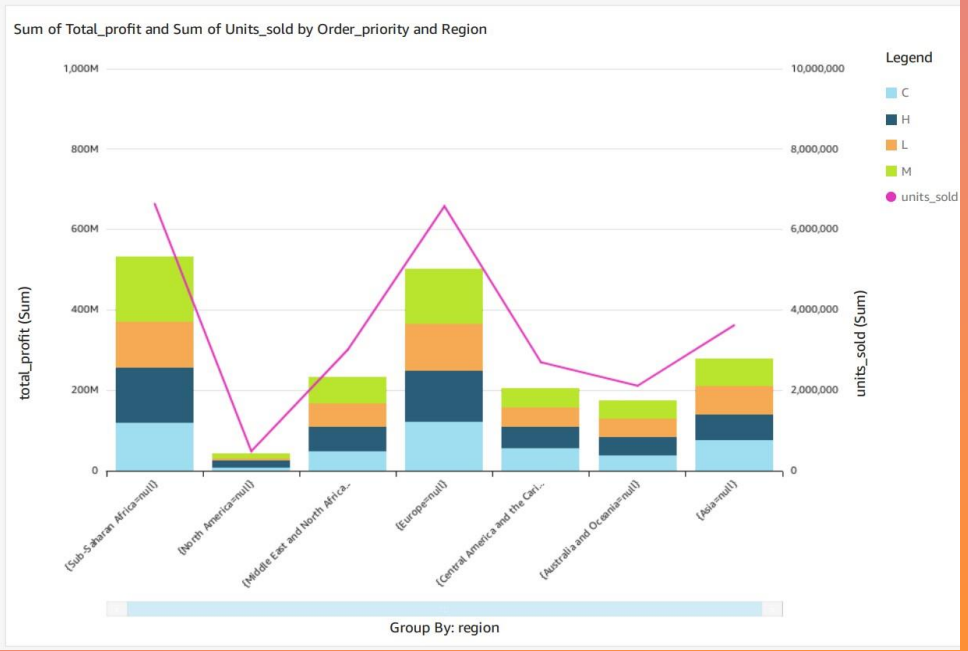


AMAZON QUICKSIGHT









WHAT’S NEW?

Amazon re:Invent 2022

AWS Re:invent 2022

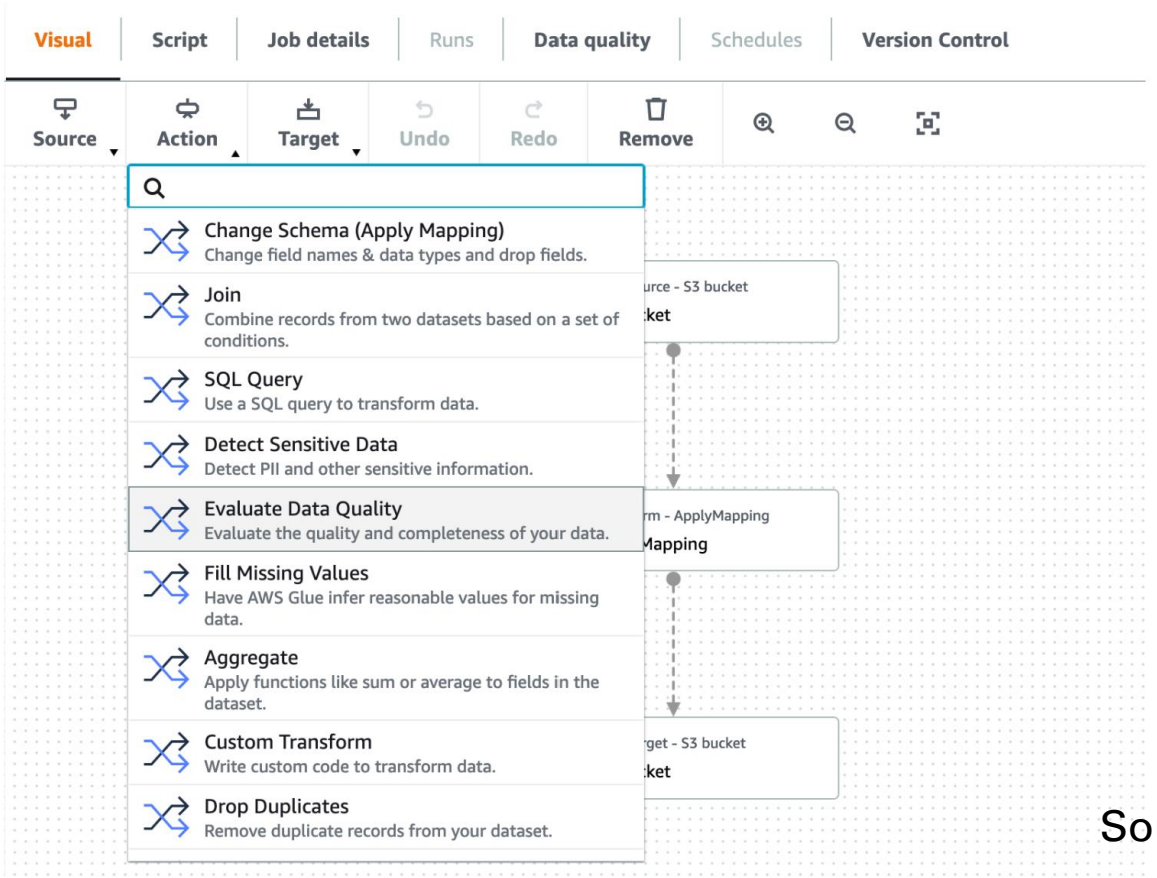
- Keynotes, Leadership sessions
- Data, Analytics
- AI/Machine Learning

https://www.youtube.com/user/amazonwebservices

Preview: AWS Glue Data Quality

Preview

- Automatic data quality recommendations on your data, based on rules



Source: AWS Glue Documentation

THANK YOU

https://dev.to/abc_wendsss