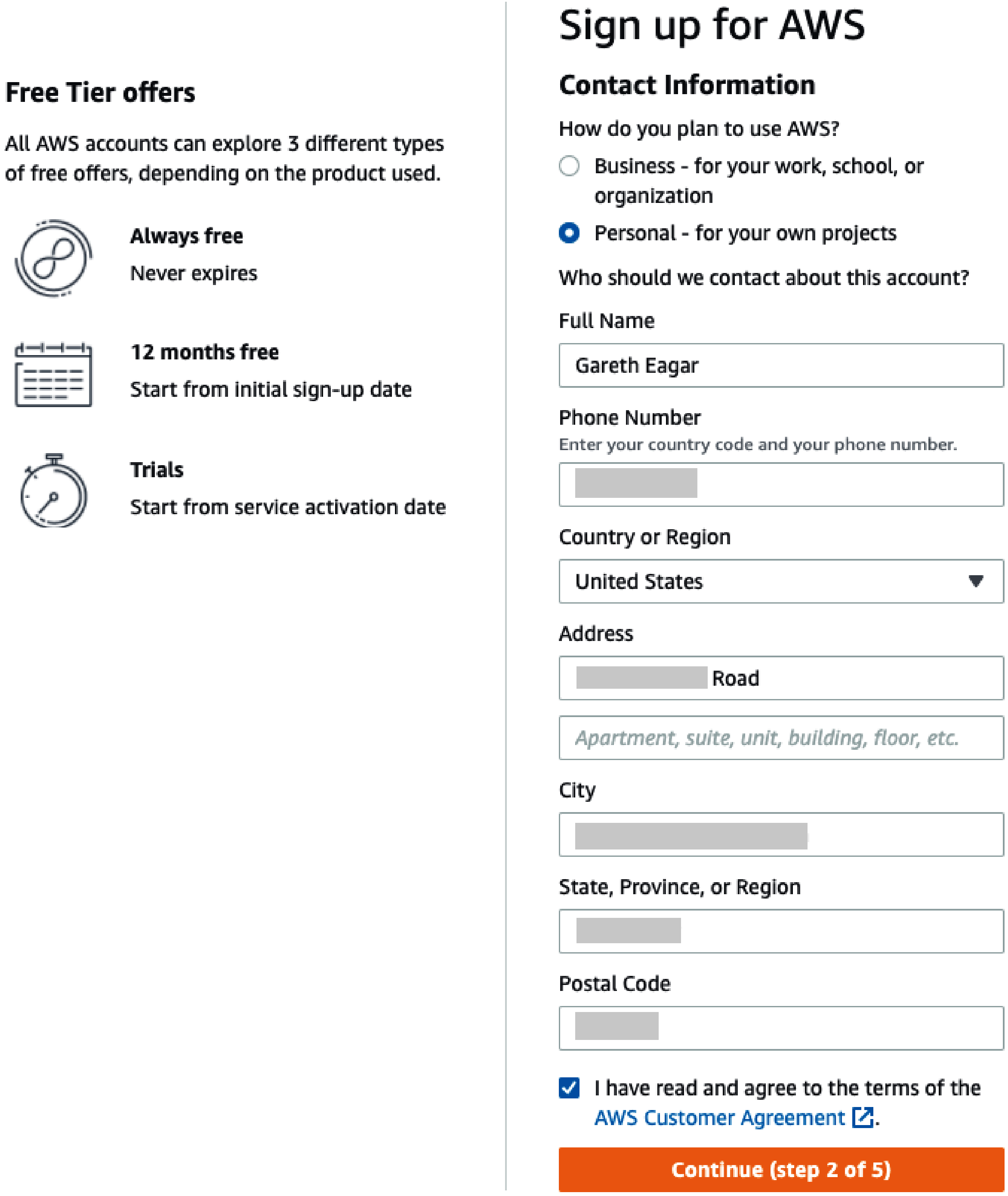


Free Tier offers

All AWS accounts can explore 3 different types of free offers, depending on the product used.

Always free
Never expires

12 months free
Start from initial sign-up date

Trials
Start from service activation date

Sign up for AWS

Contact Information

How do you plan to use AWS?

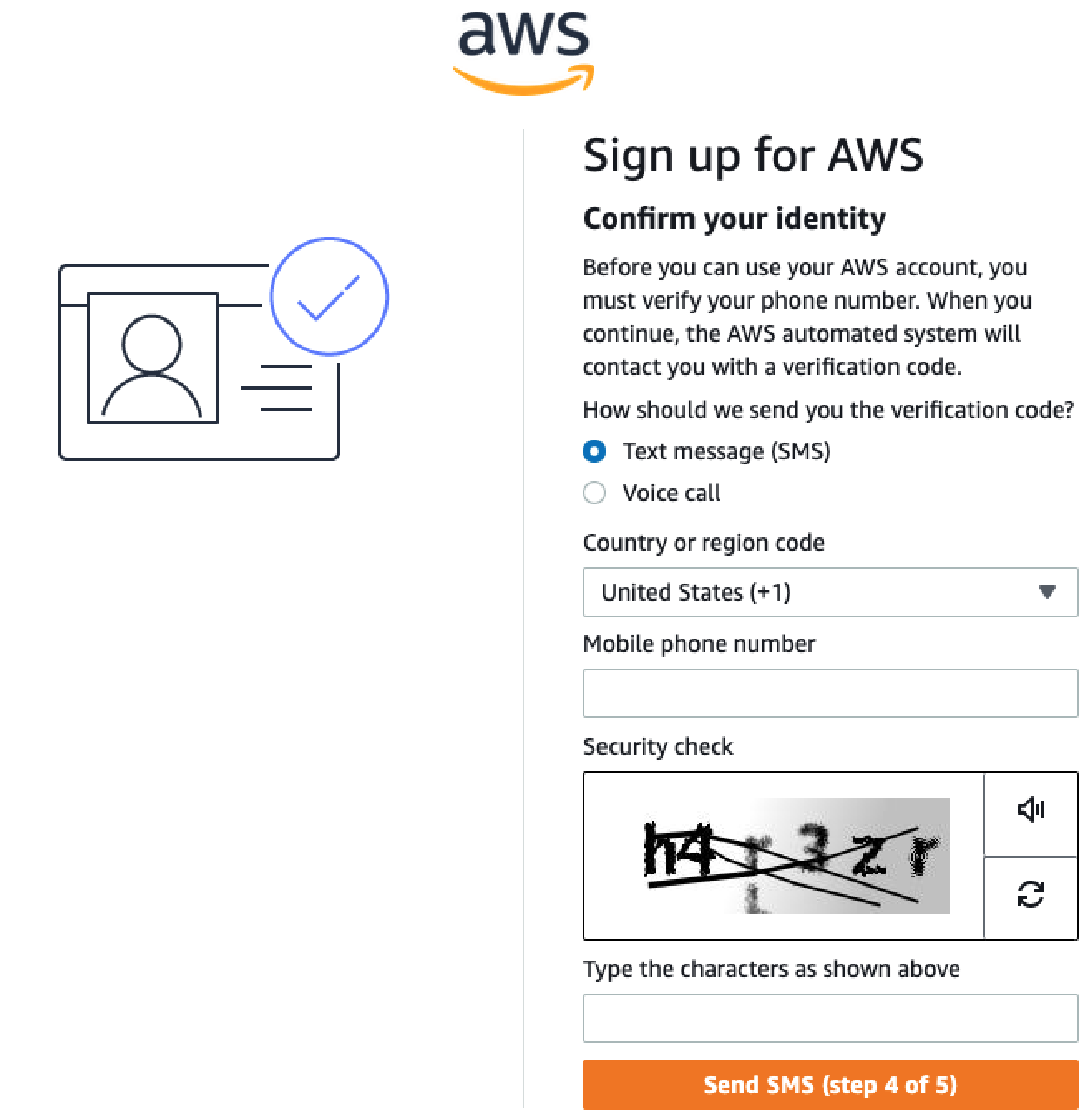
○ Business - for your work, school, or organization

● Personal - for your own projects

Who should we contact about this account?

Full Name	Gareth Eagar
Phone Number Enter your country code and your phone number.	[redacted]
Country or Region	United States ▼
Address	[redacted] Road
	Apartment, suite, unit, building, floor, etc.
City	[redacted]
State, Province, or Region	[redacted]
Postal Code	[redacted]

☑ I have read and agree to the terms of the AWS Customer Agreement.



Sign up for AWS

Confirm your identity

Before you can use your AWS account, you must verify your phone number. When you continue, the AWS automated system will contact you with a verification code.

How should we send you the verification code?

● Text message (SMS)

○ Voice call

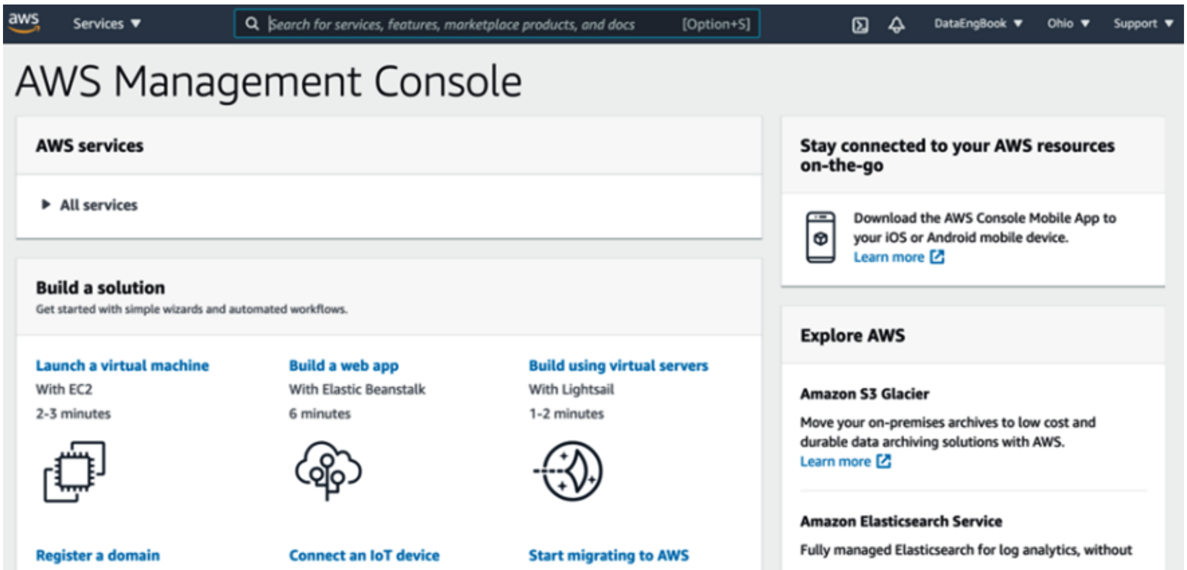
Country or region code

United States (+1)

Mobile phone number

Security check

Type the characters as shown above



aws
Services ▼
Search for services, features, marketplace products, and docs [Option+S]
DataEngBook ▼
Ohio ▼
Support ▼

AWS Management Console

AWS services

Build a solution

Get started with simple wizards and automated workflows.

Launch a virtual machine
With EC2
2-3 minutes

Build a web app
With Elastic Beanstalk
6 minutes

Build using virtual servers
With Lightsail
1-2 minutes

Stay connected to your AWS resources on-the-go

Download the AWS Console Mobile App to your iOS or Android mobile device.
Learn more

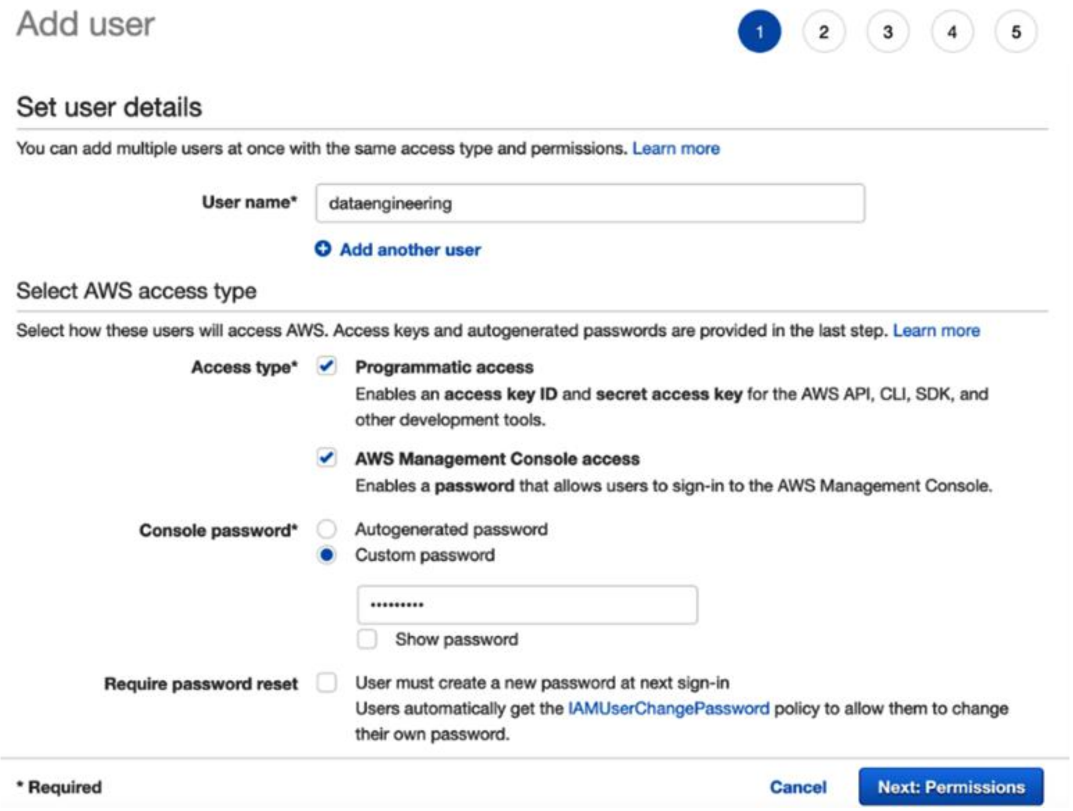
Explore AWS

Amazon S3 Glacier

Move your on-premises archives to low cost and durable data archiving solutions with AWS.
Learn more

Amazon Elasticsearch Service

Fully managed Elasticsearch for log analytics, without



Set user details

You can add multiple users at once with the same access type and permissions. Learn more

User name* dataengineering

Add another user

Select AWS access type

Select how these users will access AWS. Access keys and autogenerated passwords are provided in the last step. Learn more

Access type* ☑ **Programmatic access**
Enables an **access key ID** and **secret access key** for the AWS API, CLI, SDK, and other development tools.

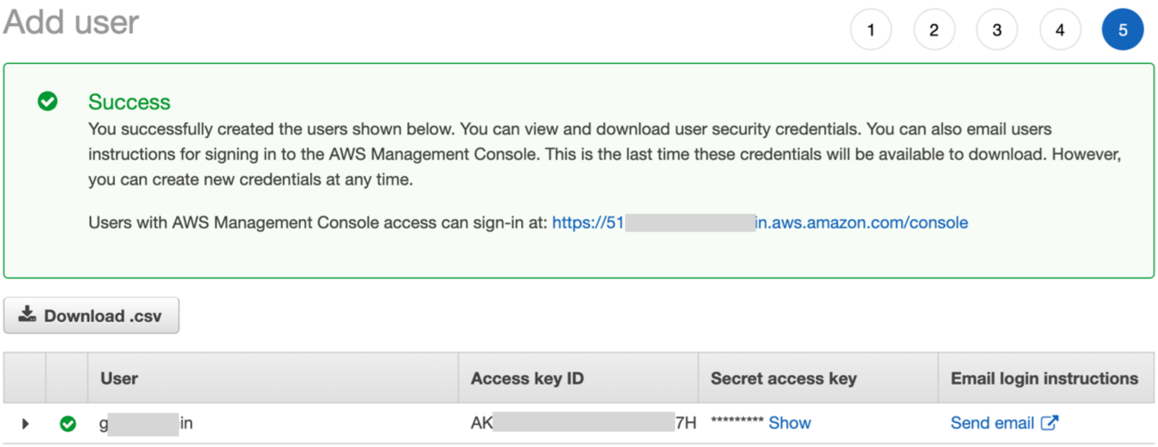
☑ **AWS Management Console access**
Enables a **password** that allows users to sign-in to the AWS Management Console.

Console password* ○ Autogenerated password
● Custom password

•••••••••

□ Show password

Require password reset □ User must create a new password at next sign-in
Users automatically get the IAMUserChangePassword policy to allow them to change their own password.



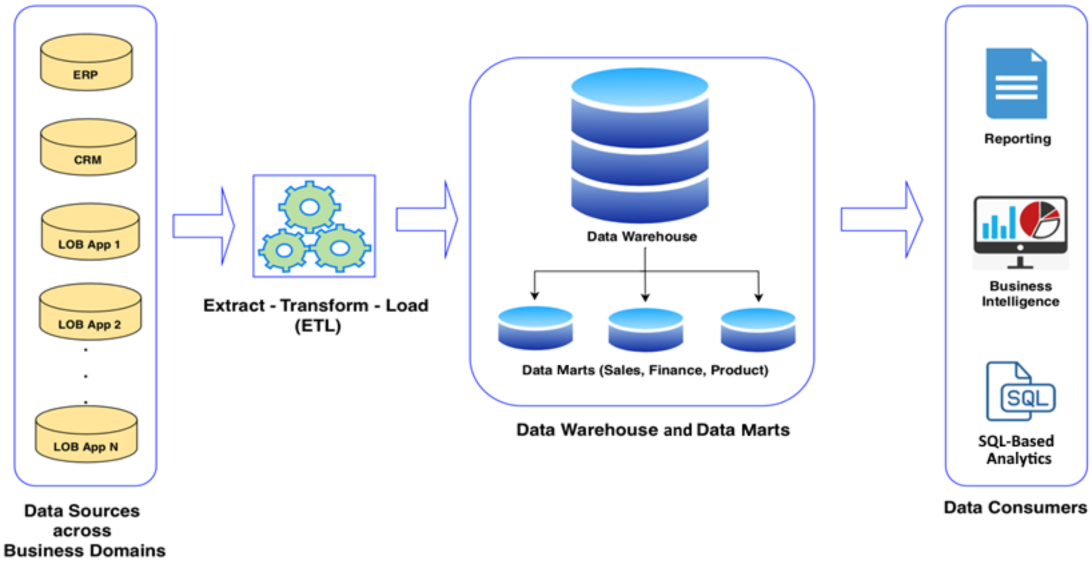
Success

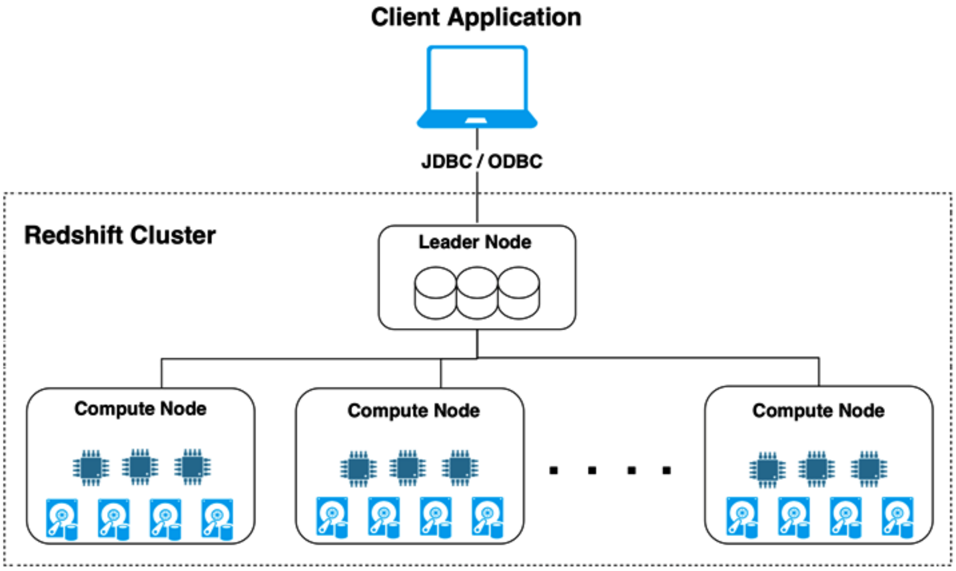
You successfully created the users shown below. You can view and download user security credentials. You can also email users instructions for signing in to the AWS Management Console. This is the last time these credentials will be available to download. However, you can create new credentials at any time.

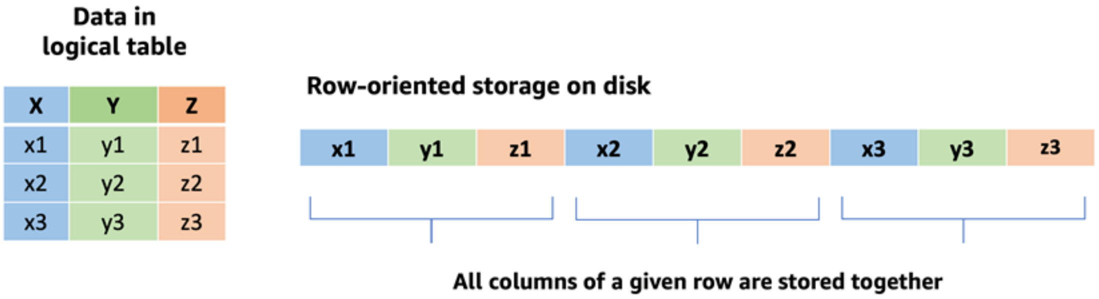
Users with AWS Management Console access can sign-in at: https://51[redacted]in.aws.amazon.com/console

		User	Access key ID	Secret access key	Email login instructions
▸	✓	g[redacted]in	AK[redacted]7H	********* Show	Send email

Chapter 2: Data Management Architectures for Analytics





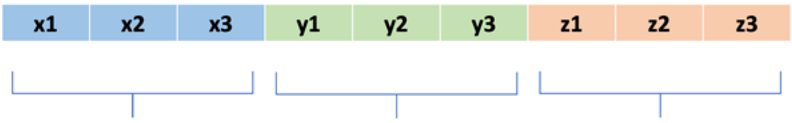

Data in
logical table
Row-oriented storage on disk
x1
y1
z1
x2
y2
z2
x3
y3
z3
All columns of a given row are stored together

X	Y	Z
x1	y1	z1
x2	y2	z2
x3	y3	z3

Data in logical table

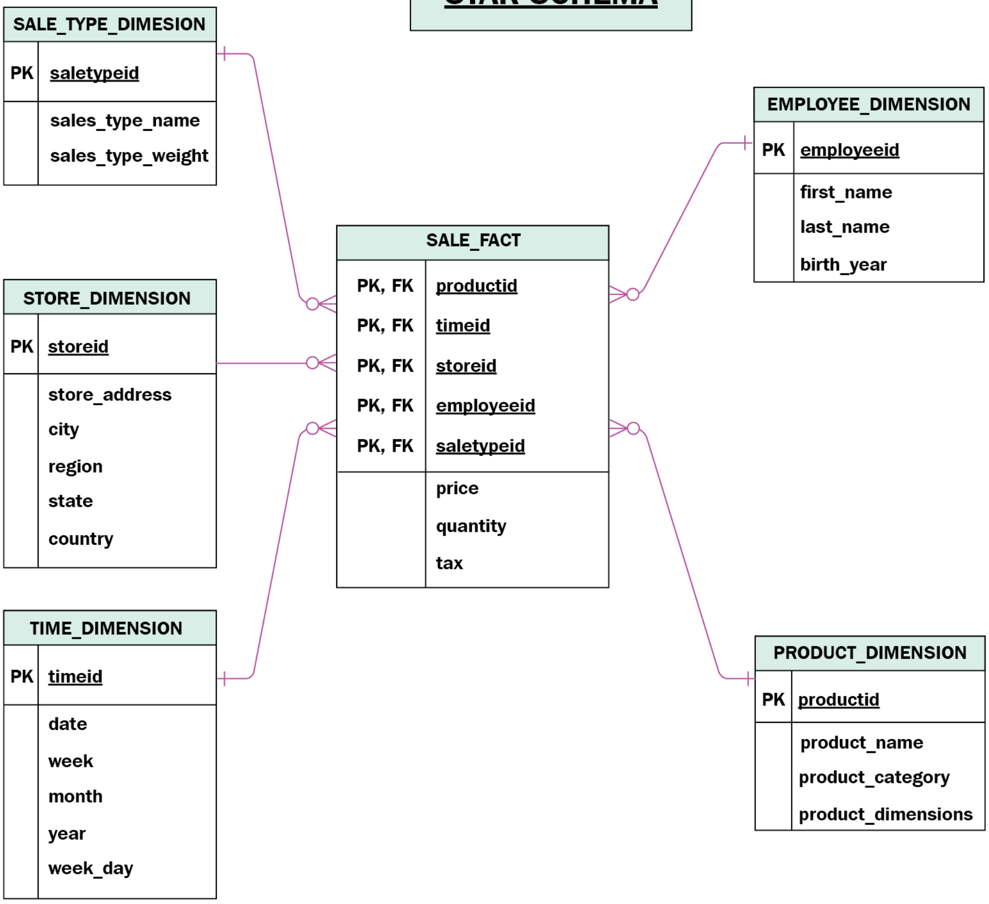
X	Y	Z
x1	y1	z1
x2	y2	z2
x3	y3	z3

Column-oriented storage on disk

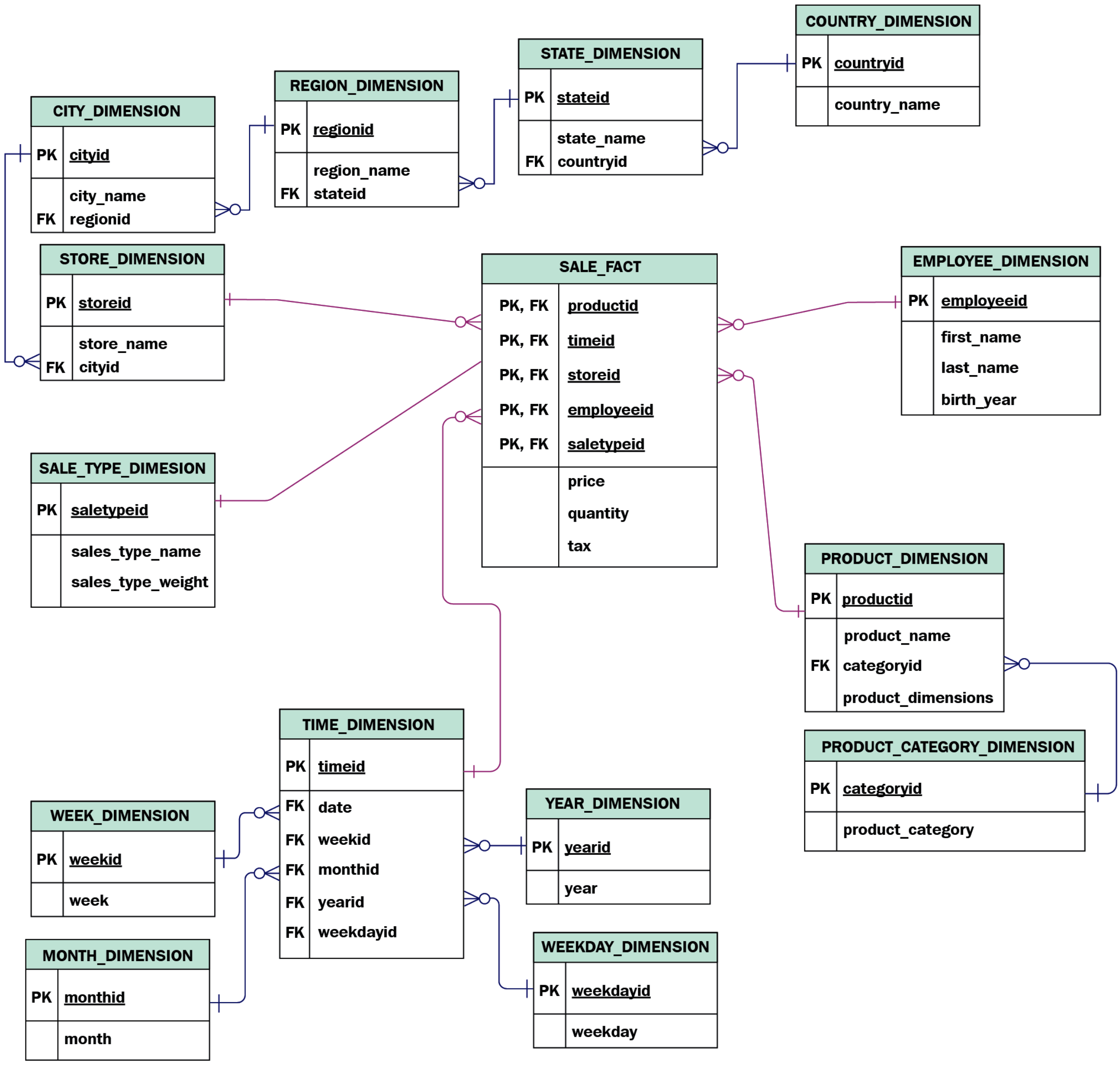


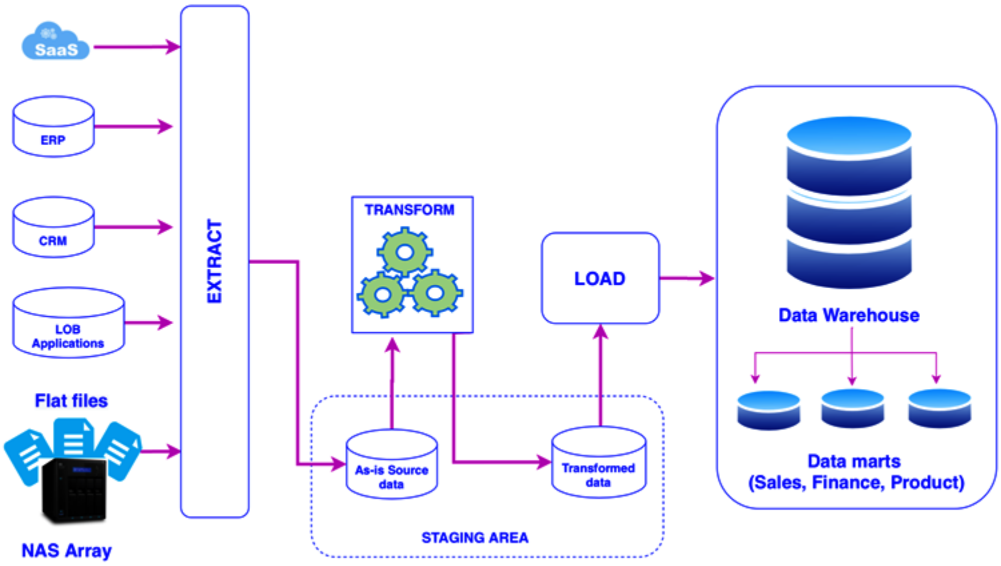
For a set of rows (a.k.a. "chunk"), all values of per column are stored together

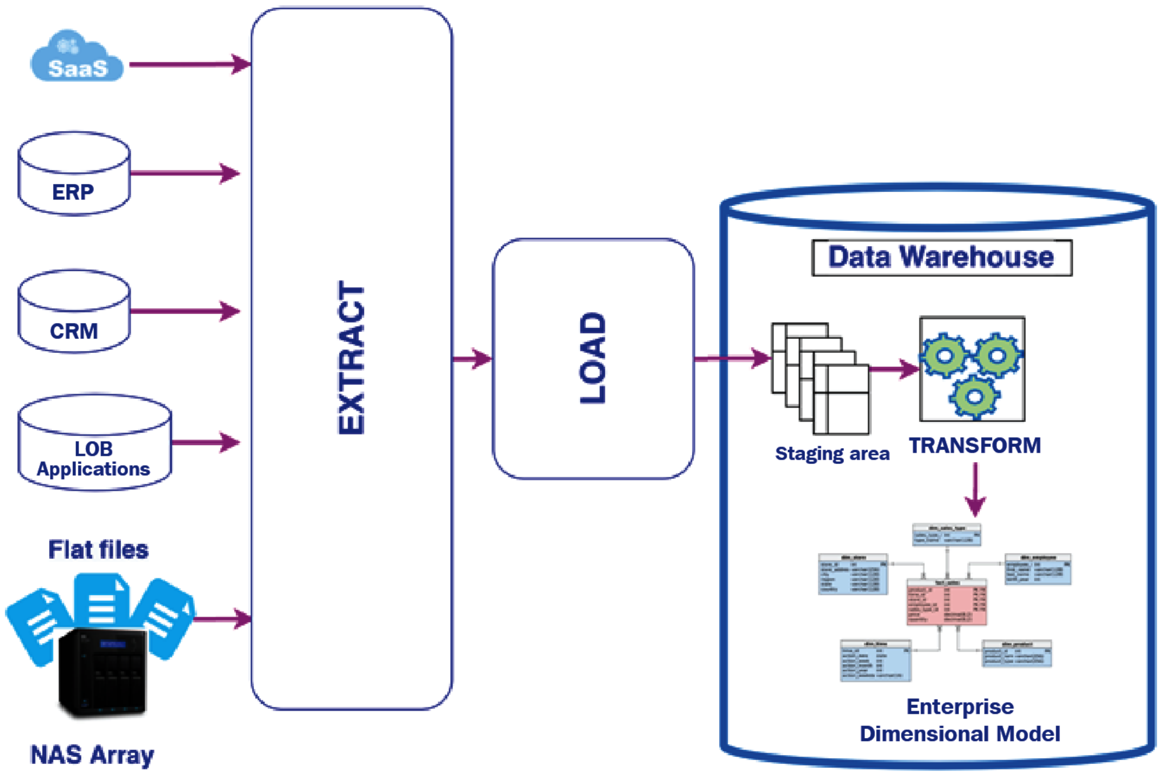
STAR SCHEMA

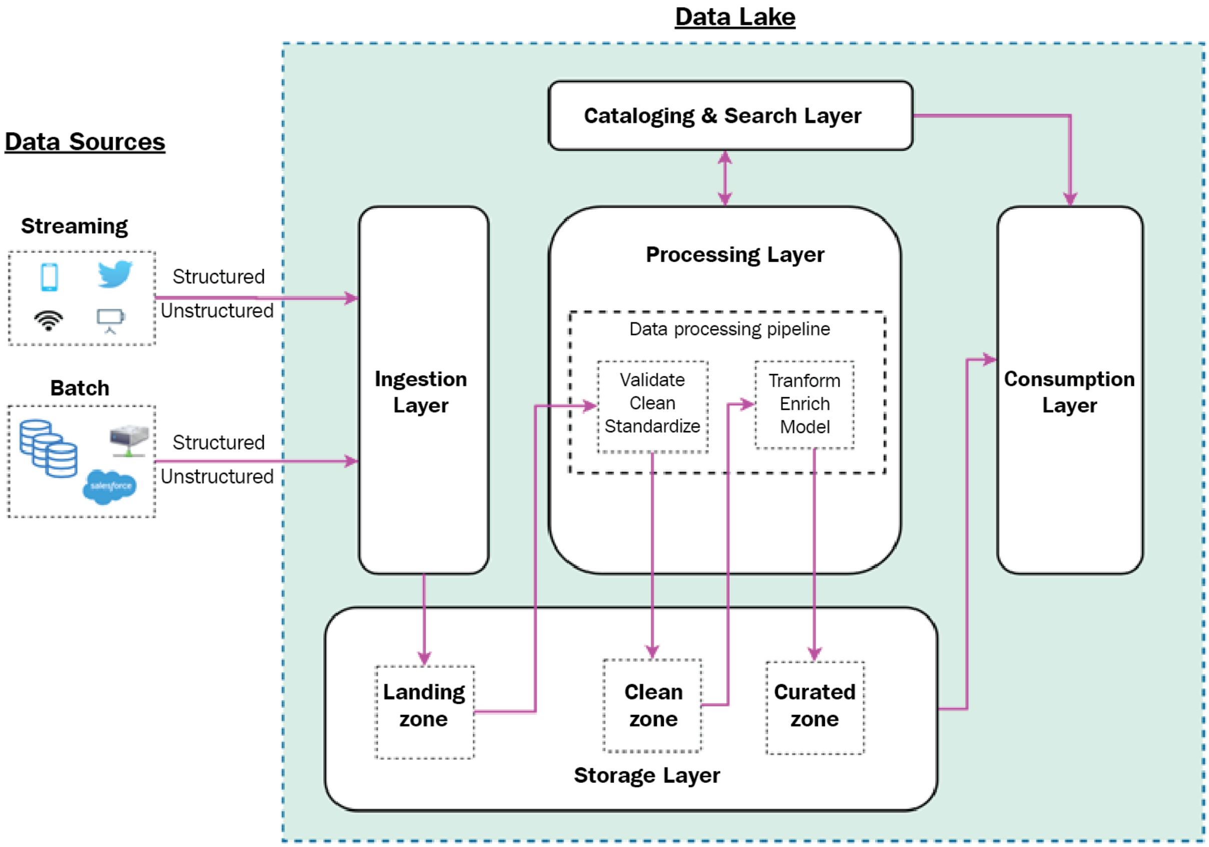


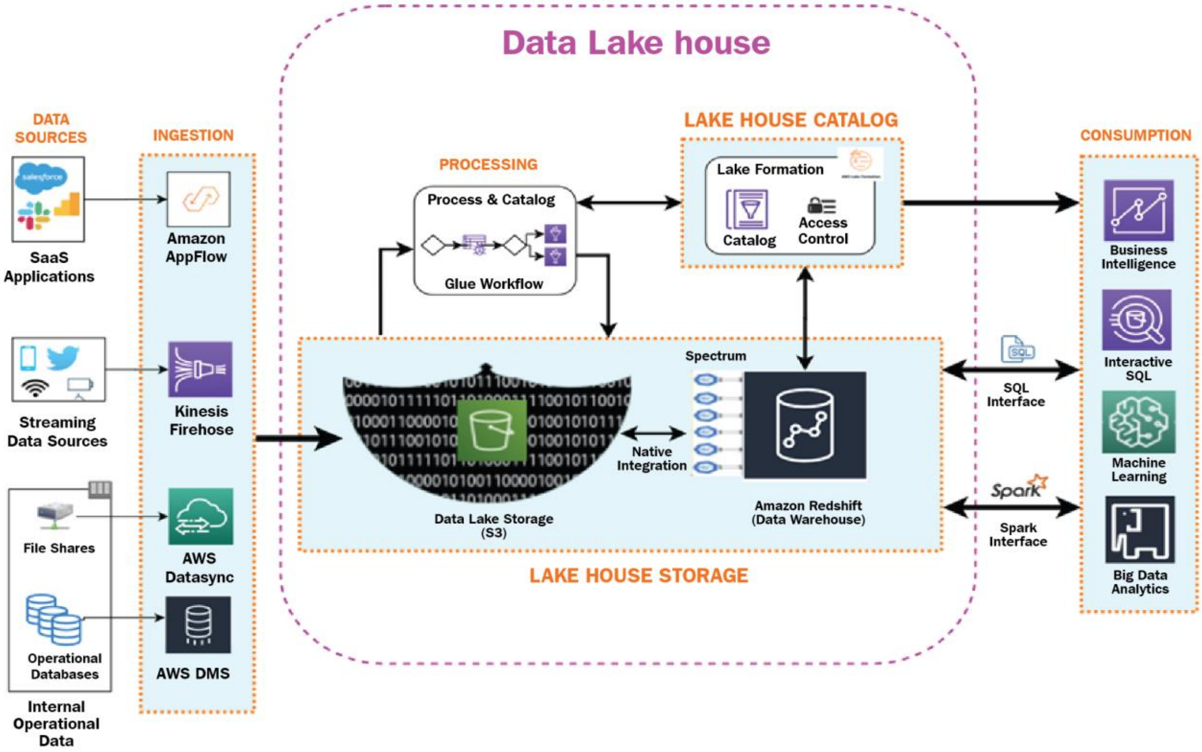
SNOWFLAKE SCHEMA



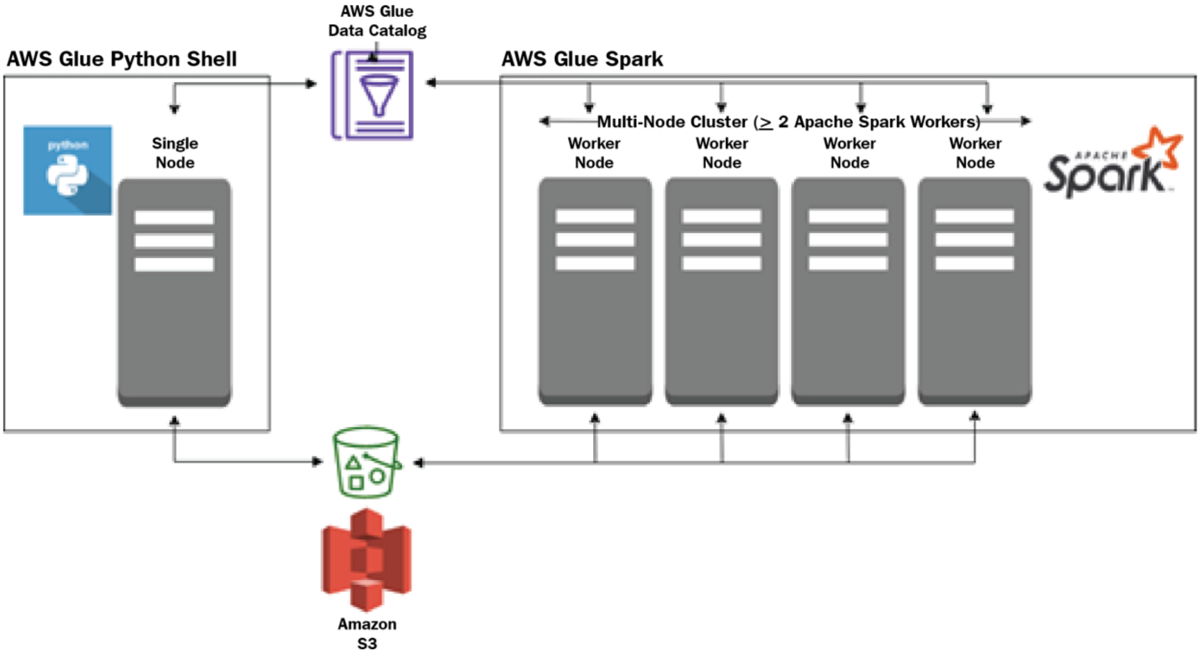


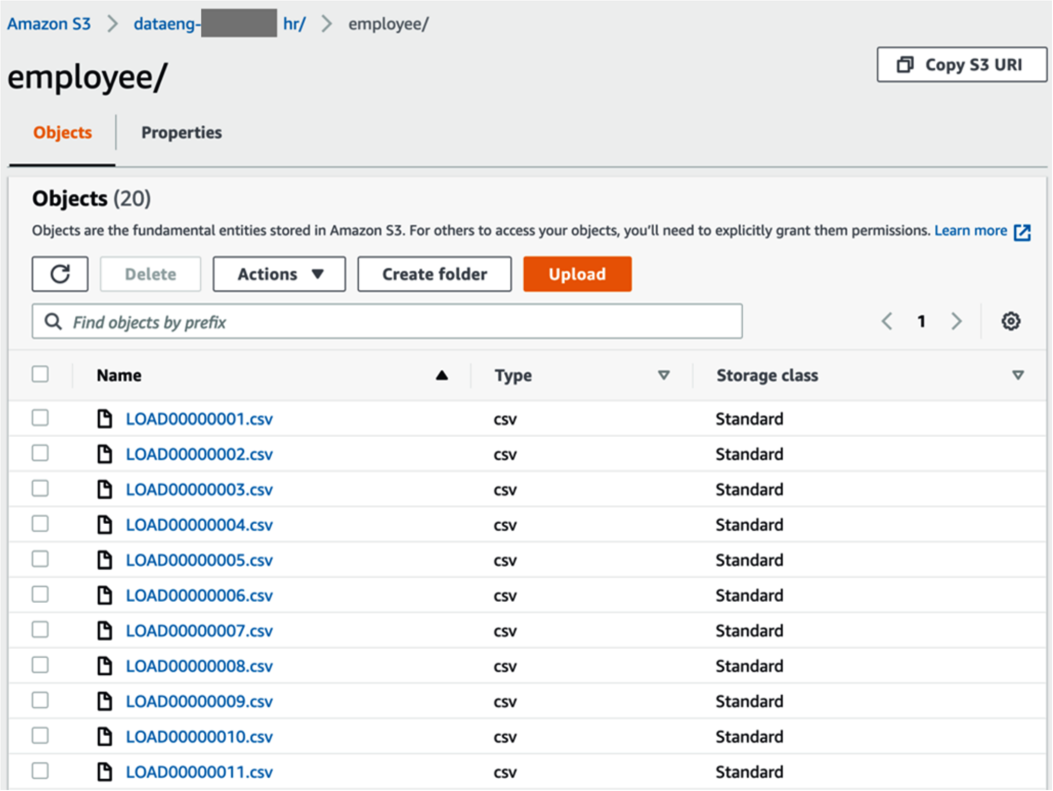

SaaS
ERP
CRM
LOB Applications
Flat files
NAS Array
EXTRACT
LOAD
Data Warehouse
Staging area
TRANSFORM
Enterprise Dimensional Model


Data Lake
Data Sources
Streaming
Structured
Unstructured
Batch
Structured
Unstructured
Cataloging & Search Layer
Ingestion Layer
Processing Layer
Data processing pipeline
Validate
Clean
Standardize
Tranform
Enrich
Model
Consumption Layer
Landing zone
Clean zone
Curated zone
Storage Layer


Data Lake house
DATA SOURCES
SaaS Applications
Streaming Data Sources
File Shares
Operational Databases
Internal Operational Data
INGESTION
Amazon AppFlow
Kinesis Firehose
AWS Datasync
AWS DMS
PROCESSING
Process & Catalog
Glue Workflow
LAKE HOUSE CATALOG
Lake Formation
Catalog
Access Control
Spectrum
Native Integration
Data Lake Storage (S3)
Amazon Redshift (Data Warehouse)
LAKE HOUSE STORAGE
SQL Interface
Spark Interface
CONSUMPTION
Business Intelligence
Interactive SQL
Machine Learning
Big Data Analytics

Chapter 3: The AWS Data Engineer's Toolkit



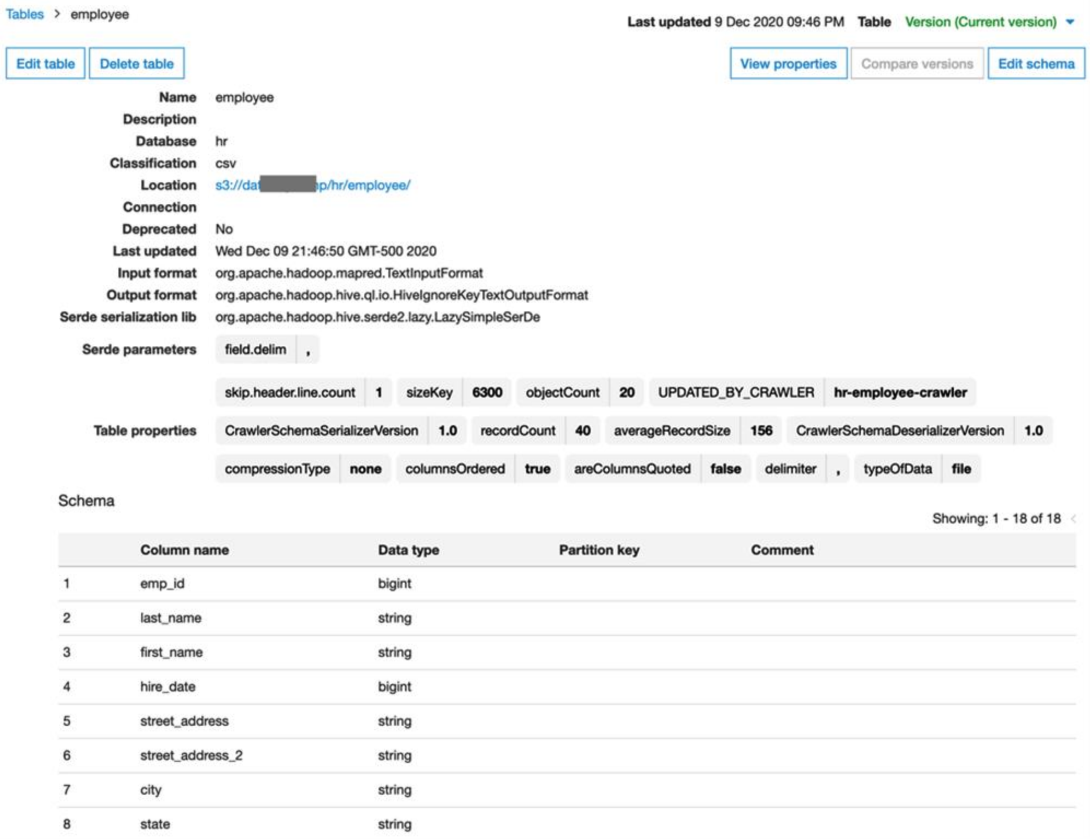


employee/

Objects (20)

Objects are the fundamental entities stored in Amazon S3. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Storage class
□	LOAD00000001.csv	csv	Standard
□	LOAD00000002.csv	csv	Standard
□	LOAD00000003.csv	csv	Standard
□	LOAD00000004.csv	csv	Standard
□	LOAD00000005.csv	csv	Standard
□	LOAD00000006.csv	csv	Standard
□	LOAD00000007.csv	csv	Standard
□	LOAD00000008.csv	csv	Standard
□	LOAD00000009.csv	csv	Standard
□	LOAD00000010.csv	csv	Standard
□	LOAD00000011.csv	csv	Standard



Tables > employee

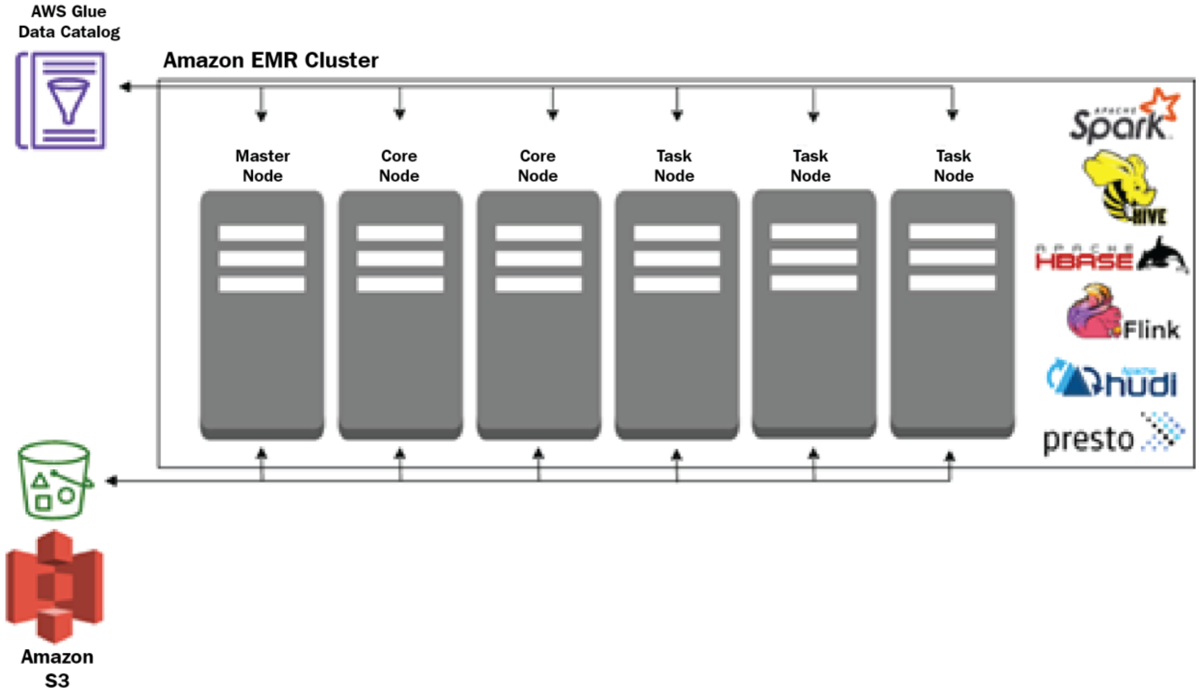
Last updated 9 Dec 2020 09:46 PM **Table** Version (Current version)

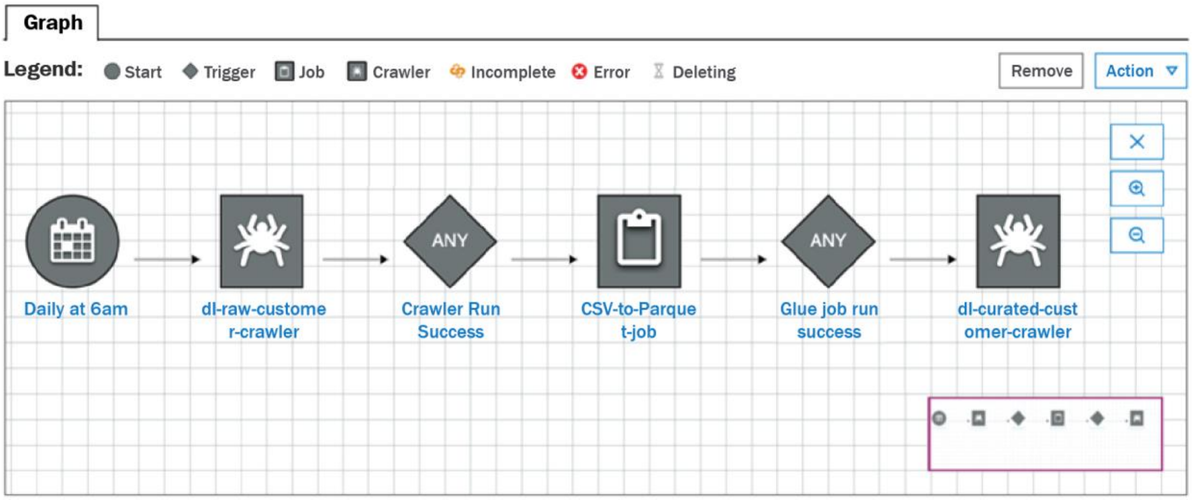
Name	employee
Description	
Database	hr
Classification	csv
Location	s3://dat[redacted]p/hr/employee/
Connection	
Deprecated	No
Last updated	Wed Dec 09 21:46:50 GMT-500 2020
Input format	org.apache.hadoop.mapred.TextInputFormat
Output format	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters	field.delim **,**
Table properties	skip.header.line.count **1** sizeKey **6300** objectCount **20** UPDATED_BY_CRAWLER **hr-employee-crawler** CrawlerSchemaSerializerVersion **1.0** recordCount **40** averageRecordSize **156** CrawlerSchemaDeserializerVersion **1.0** compressionType **none** columnsOrdered **true** areColumnsQuoted **false** delimiter **,** typeOfData **file**

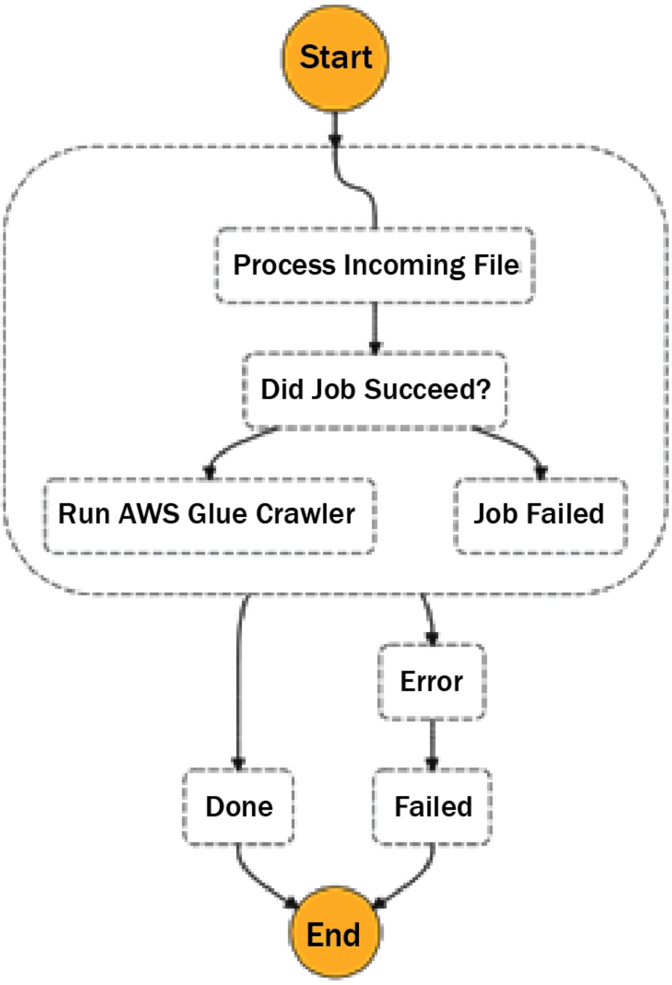
Schema

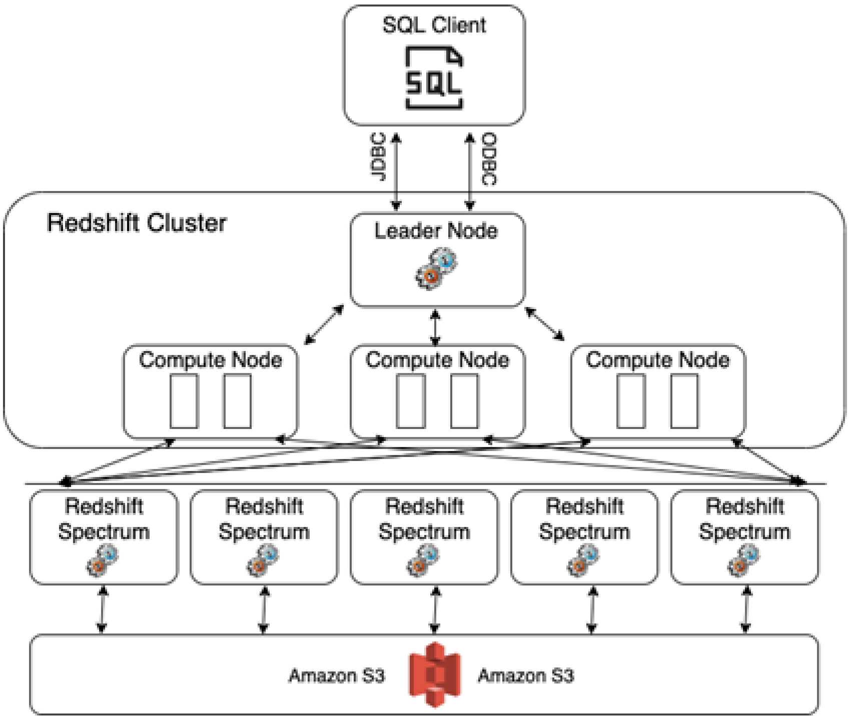
Showing: 1 - 18 of 18

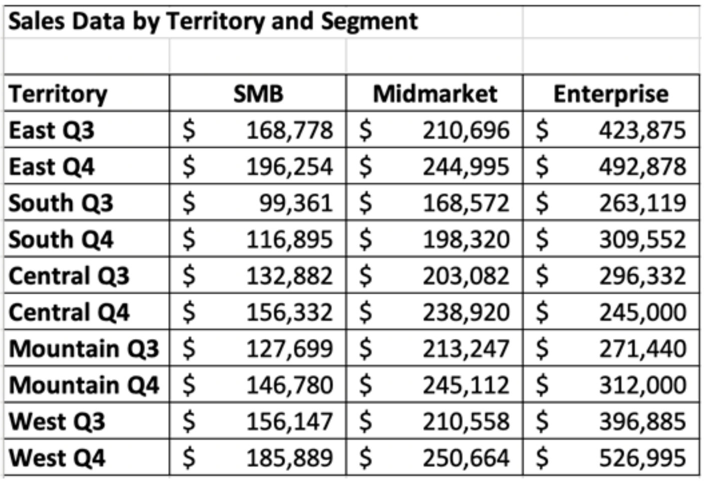
	Column name	Data type	Partition key	Comment
1	emp_id	bigint		
2	last_name	string		
3	first_name	string		
4	hire_date	bigint		
5	street_address	string		
6	street_address_2	string		
7	city	string		
8	state	string		




Graph
Legend:
● Start
◆ Trigger
Job
Crawler
Incomplete
Error
Deleting
Remove
Action ▽
Daily at 6am
dl-raw-customer-crawler
ANY
Crawler Run Success
CSV-to-Parquet-job
ANY
Glue job run success
dl-curated-customer-crawler

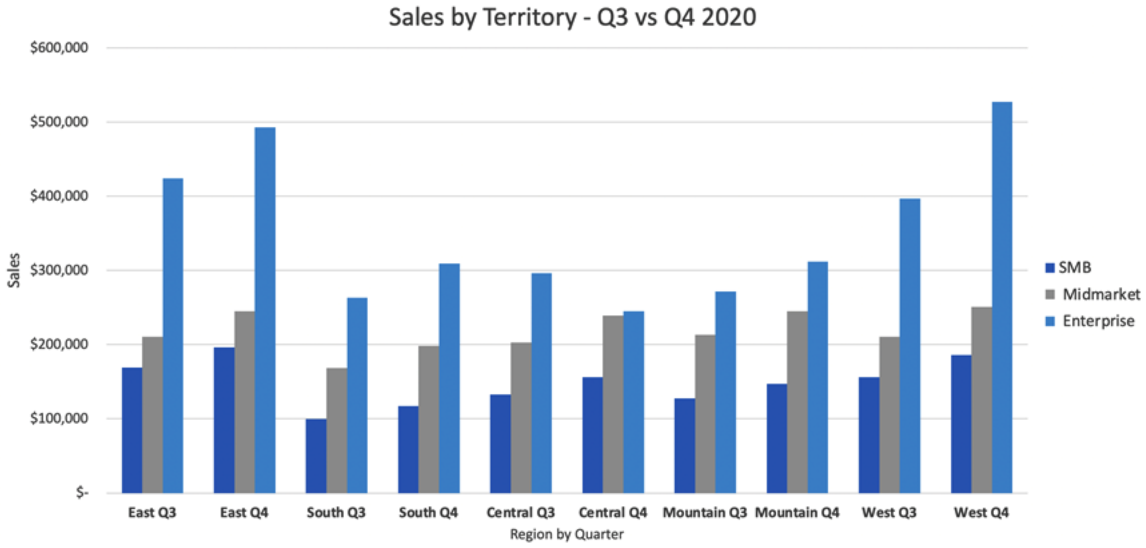

Start
Process Incoming File
Did Job Succeed?
Run AWS Glue Crawler
Job Failed
Error
Done
Failed
End

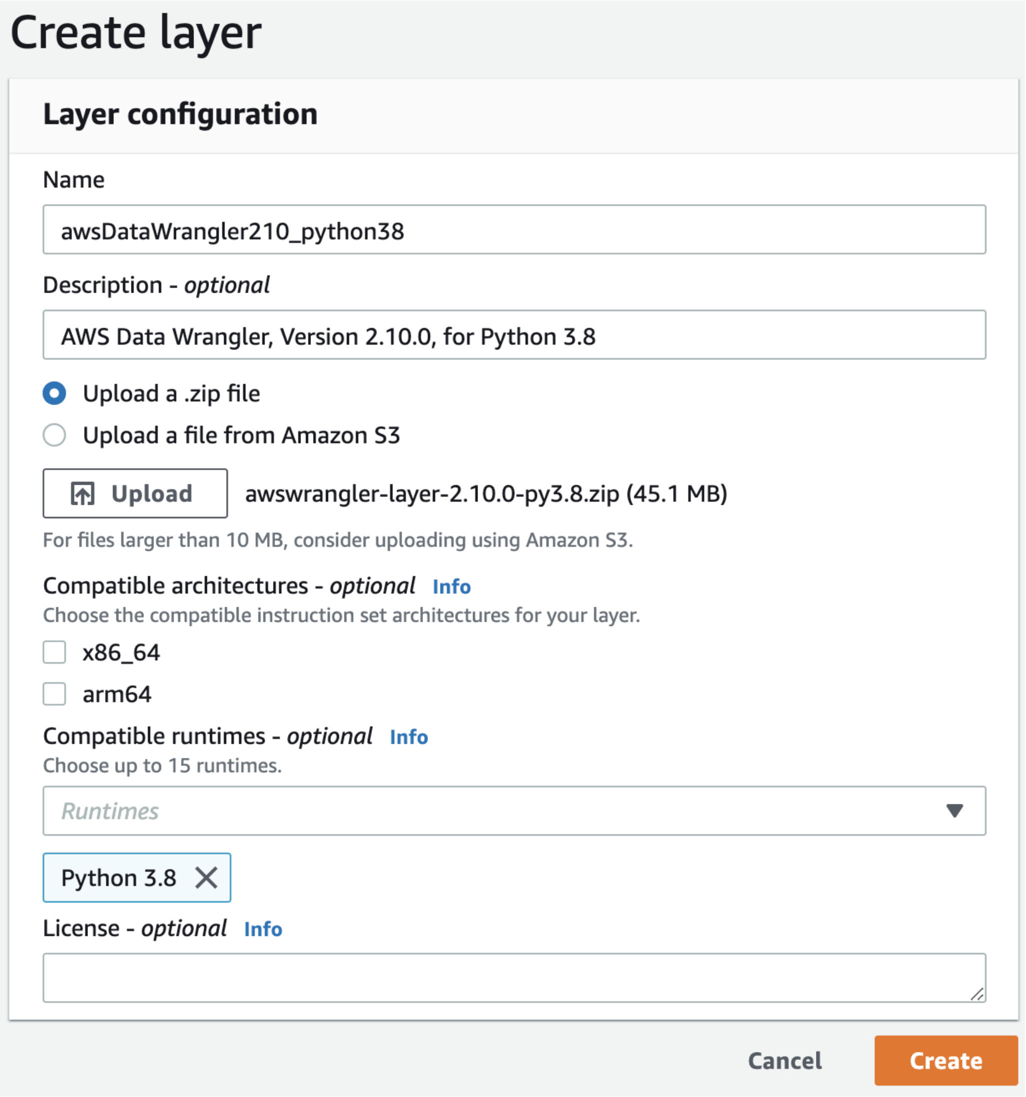




Sales Data by Territory and Segment

Territory	SMB	Midmarket	Enterprise
East Q3	$ 168,778	$ 210,696	$ 423,875
East Q4	$ 196,254	$ 244,995	$ 492,878
South Q3	$ 99,361	$ 168,572	$ 263,119
South Q4	$ 116,895	$ 198,320	$ 309,552
Central Q3	$ 132,882	$ 203,082	$ 296,332
Central Q4	$ 156,332	$ 238,920	$ 245,000
Mountain Q3	$ 127,699	$ 213,247	$ 271,440
Mountain Q4	$ 146,780	$ 245,112	$ 312,000
West Q3	$ 156,147	$ 210,558	$ 396,885
West Q4	$ 185,889	$ 250,664	$ 526,995





Create layer

Layer configuration

Name

awsDataWrangler210_python38

Description - *optional*

AWS Data Wrangler, Version 2.10.0, for Python 3.8

● Upload a .zip file

○ Upload a file from Amazon S3

awswrangler-layer-2.10.0-py3.8.zip (45.1 MB)

For files larger than 10 MB, consider uploading using Amazon S3.

Compatible architectures - *optional* Info

Choose the compatible instruction set architectures for your layer.

□ x86_64

□ arm64

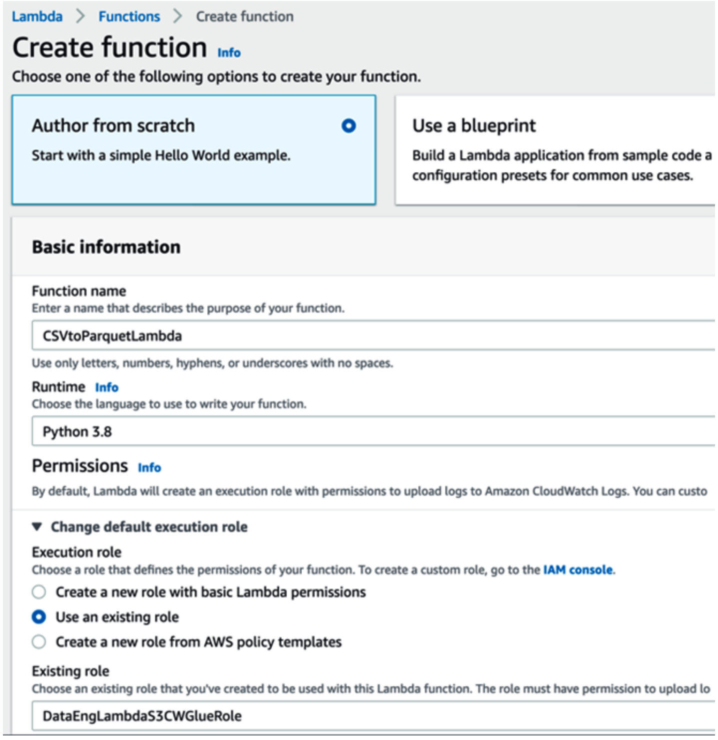
Compatible runtimes - *optional* Info

Choose up to 15 runtimes.

Runtimes

Python 3.8 ✕

License - *optional* Info



Create function Info

Choose one of the following options to create your function.

Author from scratch ●

Start with a simple Hello World example.

Use a blueprint

Build a Lambda application from sample code a
configuration presets for common use cases.

Basic information

Function name

Enter a name that describes the purpose of your function.

CSVtoParquetLambda

Use only letters, numbers, hyphens, or underscores with no spaces.

Runtime Info

Choose the language to use to write your function.

Python 3.8

Permissions Info

By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can custo

▼ Change default execution role

Execution role

Choose a role that defines the permissions of your function. To create a custom role, go to the IAM console.

○ Create a new role with basic Lambda permissions

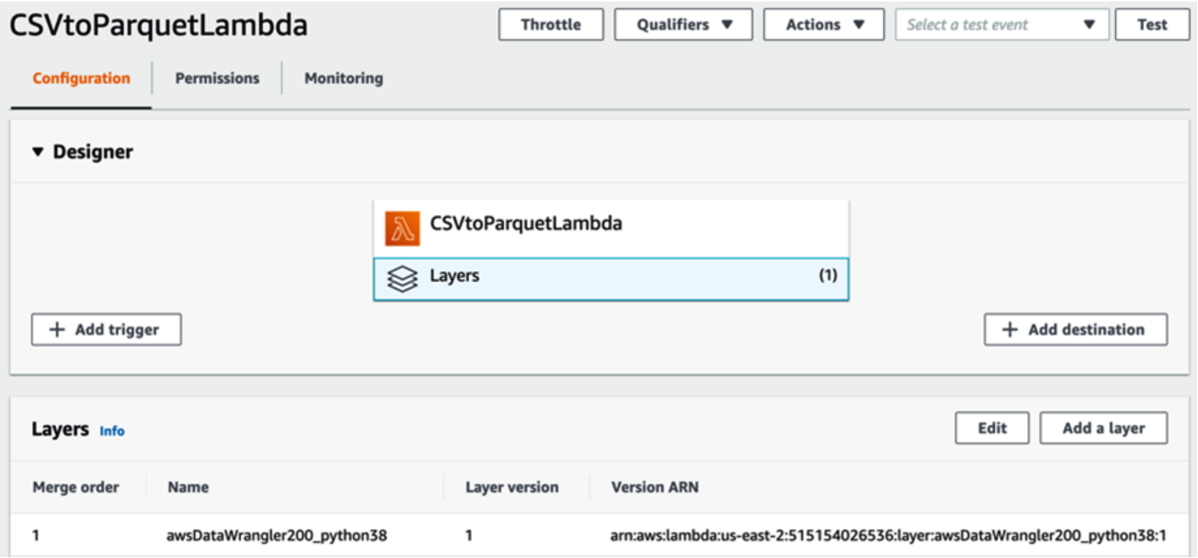
● Use an existing role

○ Create a new role from AWS policy templates

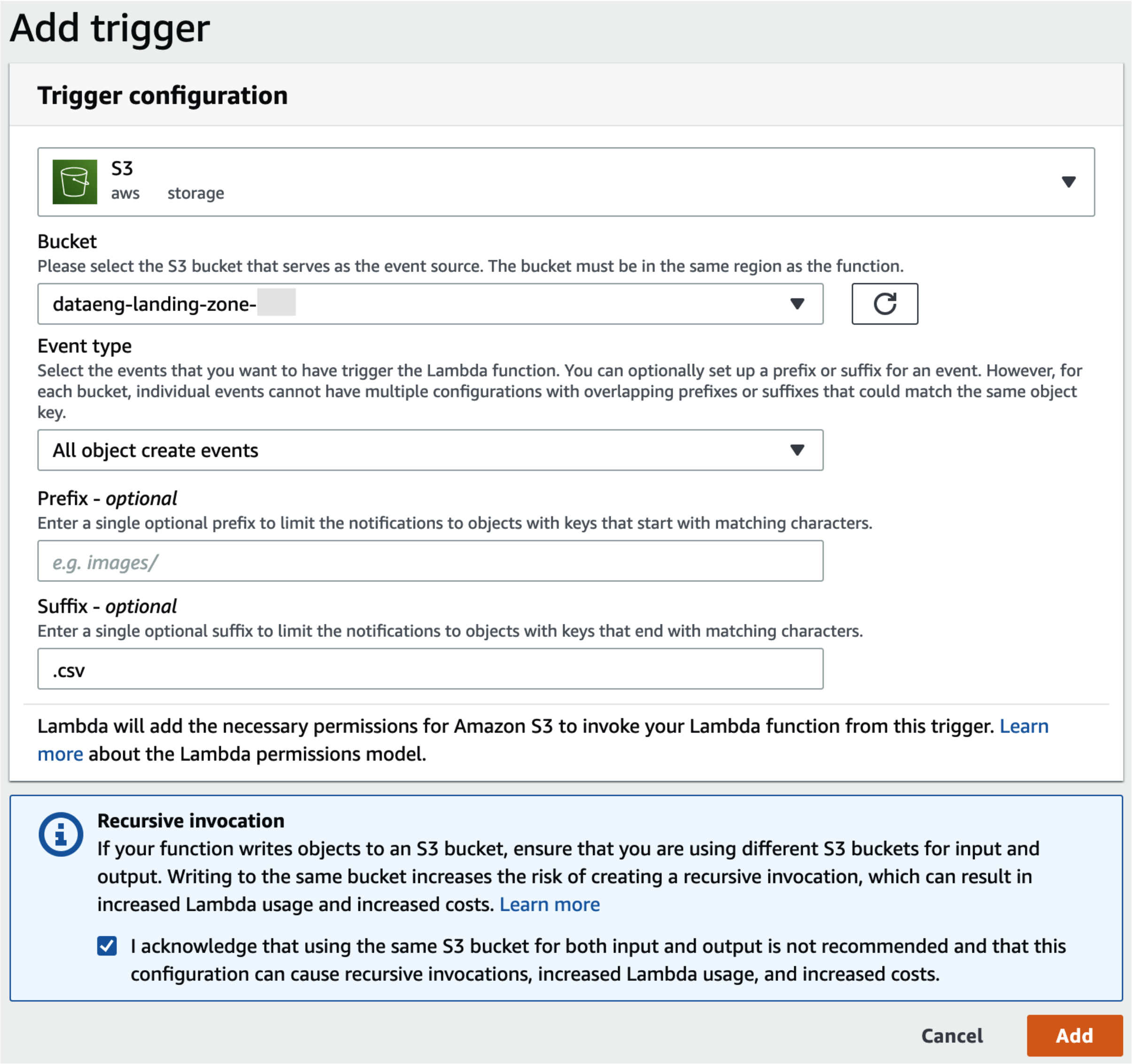
Existing role

Choose an existing role that you've created to be used with this Lambda function. The role must have permission to upload lo

DataEngLambdaS3CWGlueRole



Merge order	Name	Layer version	Version ARN
1	awsDataWrangler200_python38	1	arn:aws:lambda:us-east-2:515154026536:layer:awsDataWrangler200_python38:1



Bucket

Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.

dataeng-landing-zone-[redacted]

Event type

Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.

All object create events

Prefix - *optional*

Enter a single optional prefix to limit the notifications to objects with keys that start with matching characters.

e.g. images/

Suffix - *optional*

Enter a single optional suffix to limit the notifications to objects with keys that end with matching characters.

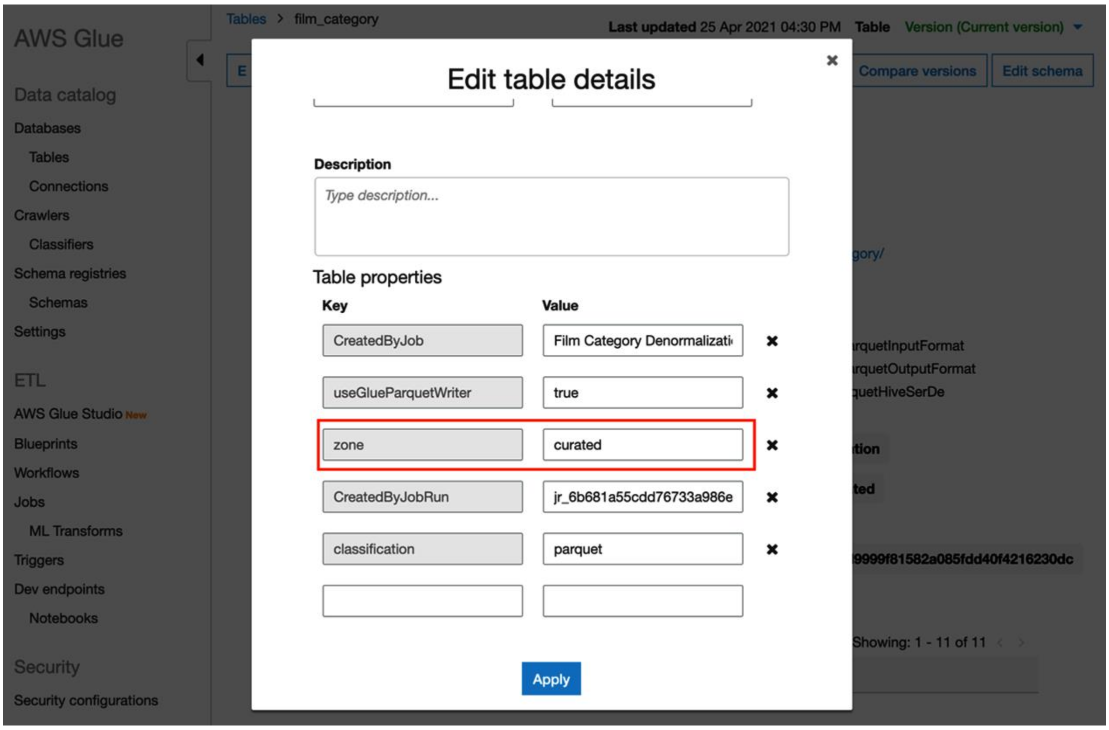
.csv

Lambda will add the necessary permissions for Amazon S3 to invoke your Lambda function from this trigger. Learn more about the Lambda permissions model.

Recursive invocation

If your function writes objects to an S3 bucket, ensure that you are using different S3 buckets for input and output. Writing to the same bucket increases the risk of creating a recursive invocation, which can result in increased Lambda usage and increased costs. Learn more

☑ I acknowledge that using the same S3 bucket for both input and output is not recommended and that this configuration can cause recursive invocations, increased Lambda usage, and increased costs.



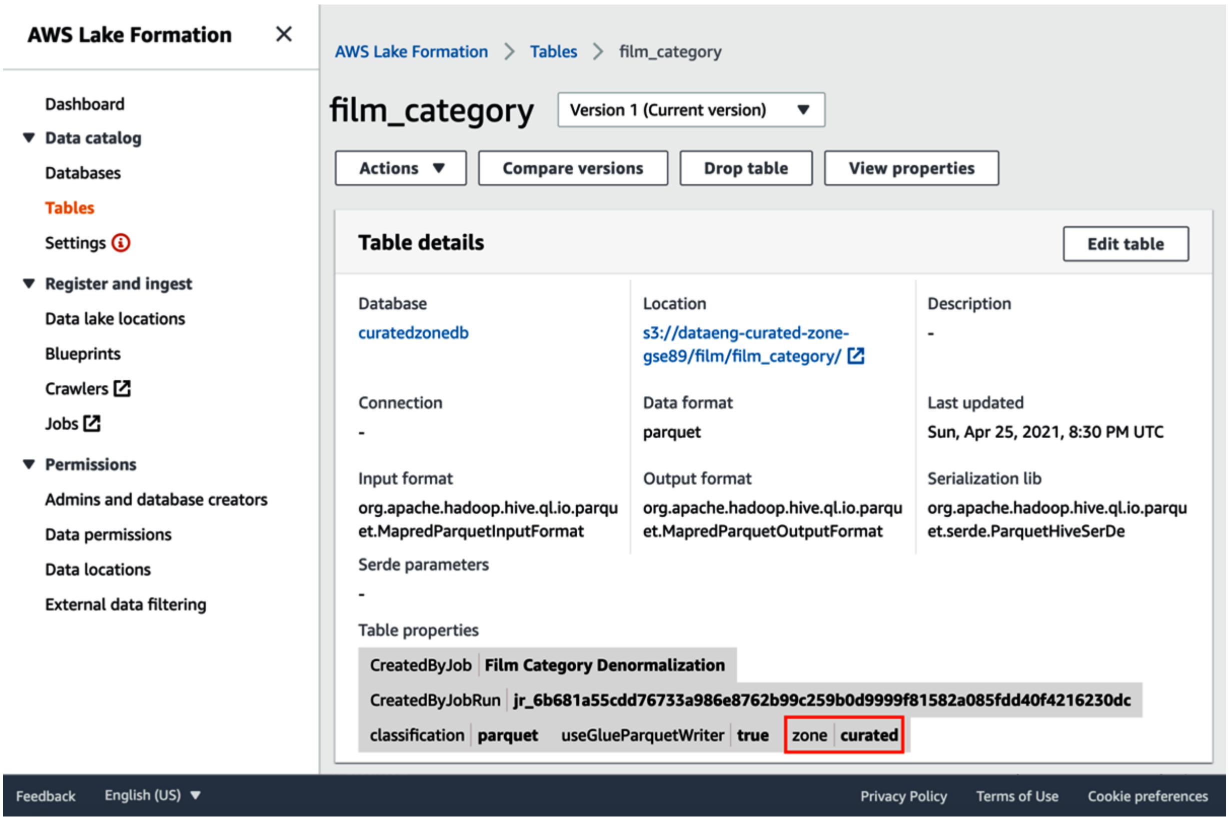
Edit table details

Description

Type description...

Table properties

Key	Value
CreatedByJob	Film Category Denormalizati
useGlueParquetWriter	true
zone	curated
CreatedByJobRun	jr_6b681a55cdd76733a986e
classification	parquet



film_category

Version 1 (Current version)

Table details

Database
curatedzonedb

Location
s3://dataeng-curated-zone-gse89/film/film_category/

Description
-

Connection
-

Data format
parquet

Last updated
Sun, Apr 25, 2021, 8:30 PM UTC

Input format
org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat

Output format
org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat

Serialization lib
org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe

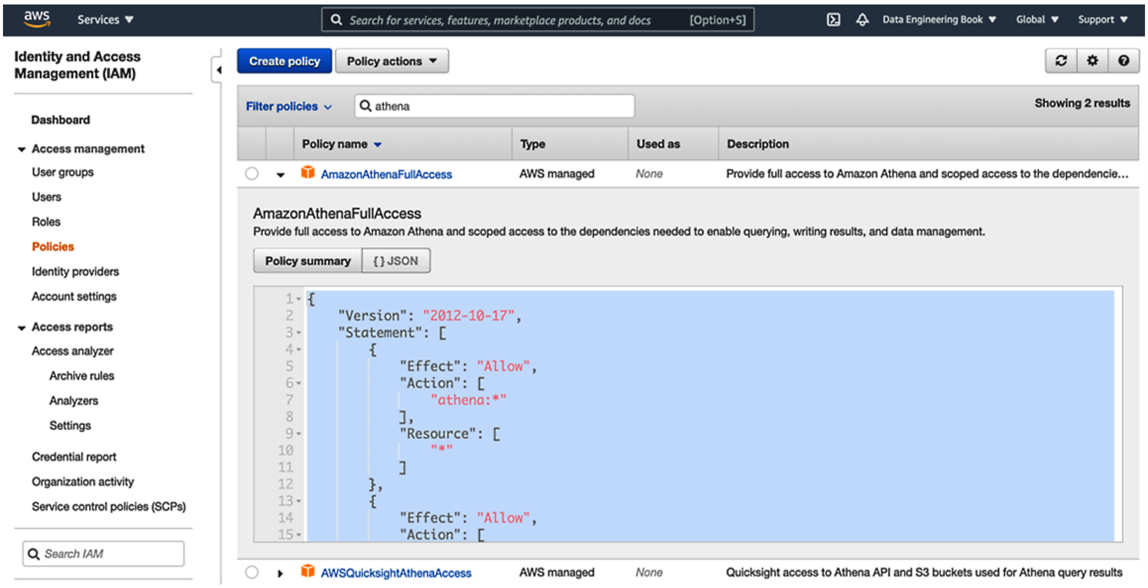
Serde parameters
-

Table properties

CreatedByJob | Film Category Denormalization

CreatedByJobRun | jr_6b681a55cdd76733a986e8762b99c259b0d9999f81582a085fdd40f4216230dc

classification | parquet useGlueParquetWriter | true zone | curated



athena

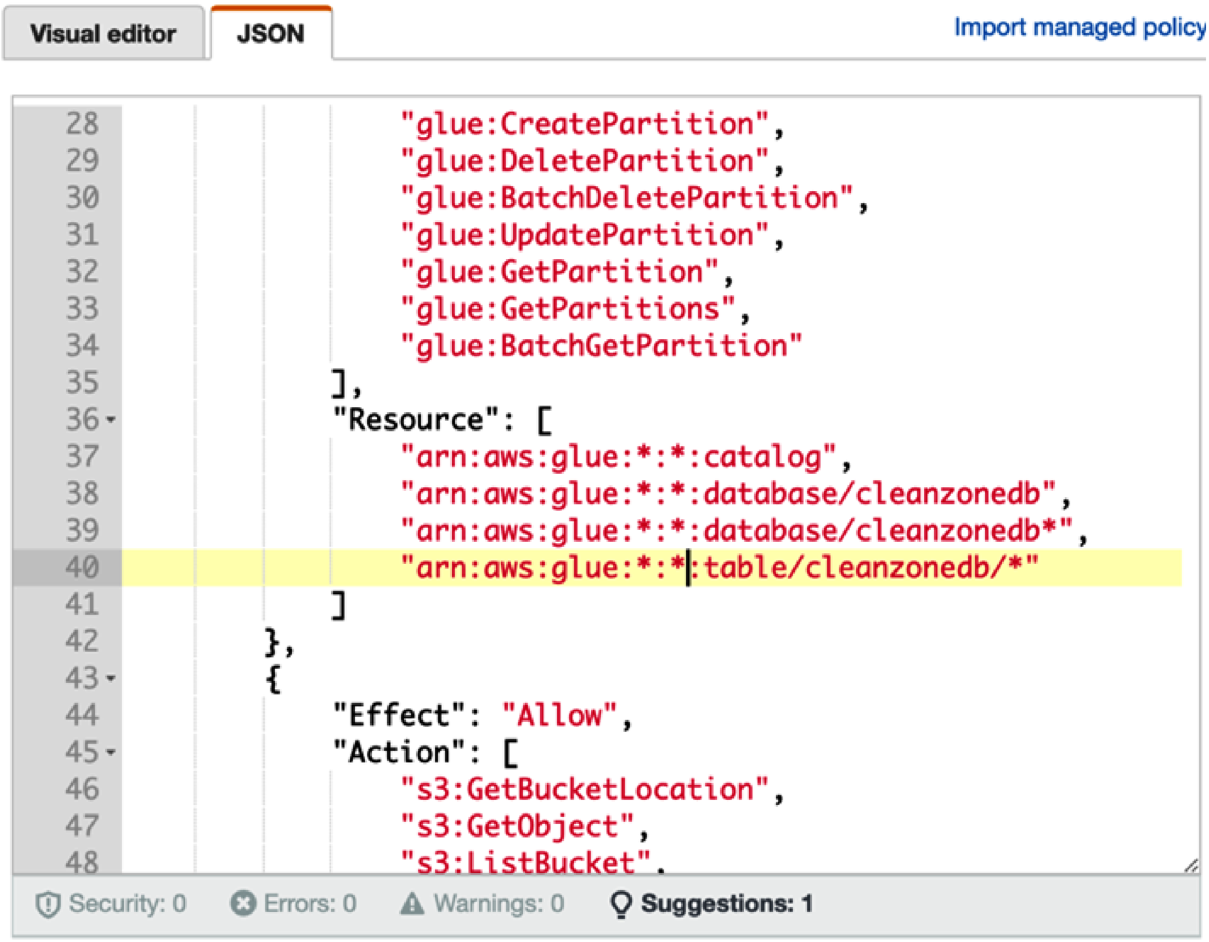
Showing 2 results

	Policy name ▼	Type	Used as	Description
○ ▼	AmazonAthenaFullAccess	AWS managed	None	Provide full access to Amazon Athena and scoped access to the dependencie...
○ ▶	AWSQuicksightAthenaAccess	AWS managed	None	Quicksight access to Athena API and S3 buckets used for Athena query results

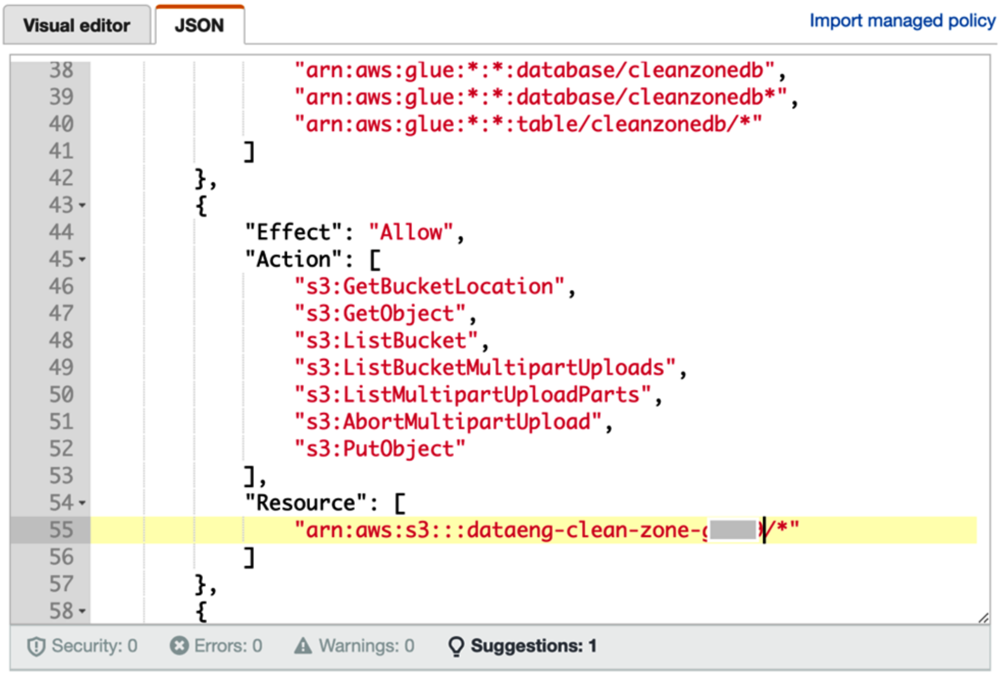
AmazonAthenaFullAccess

Provide full access to Amazon Athena and scoped access to the dependencies needed to enable querying, writing results, and data management.

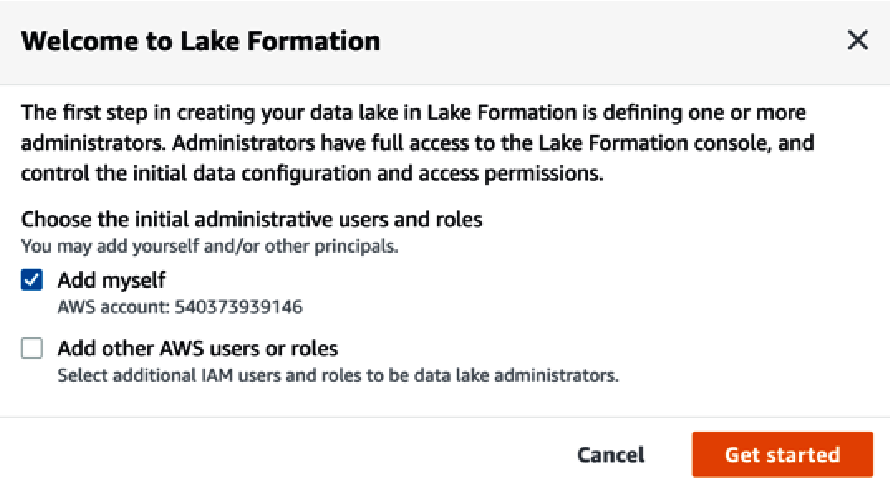
```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "athena:*"
            ],
            "Resource": [
                "*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
```



```
                "glue:CreatePartition",
                "glue:DeletePartition",
                "glue:BatchDeletePartition",
                "glue:UpdatePartition",
                "glue:GetPartition",
                "glue:GetPartitions",
                "glue:BatchGetPartition"
            ],
            "Resource": [
                "arn:aws:glue:*:*:catalog",
                "arn:aws:glue:*:*:database/cleanzonedb",
                "arn:aws:glue:*:*:database/cleanzonedb*",
                "arn:aws:glue:*:*:table/cleanzonedb/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetBucketLocation",
                "s3:GetObject",
                "s3:ListBucket",
```



```
                "arn:aws:glue:*:*:database/cleanzonedb",
                "arn:aws:glue:*:*:database/cleanzonedb*",
                "arn:aws:glue:*:*:table/cleanzonedb/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetBucketLocation",
                "s3:GetObject",
                "s3:ListBucket",
                "s3:ListBucketMultipartUploads",
                "s3:ListMultipartUploadParts",
                "s3:AbortMultipartUpload",
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::dataeng-clean-zone-[redacted]/*"
            ]
        },
        {
```



Welcome to Lake Formation

The first step in creating your data lake in Lake Formation is defining one or more administrators. Administrators have full access to the Lake Formation console, and control the initial data configuration and access permissions.

Choose the initial administrative users and roles

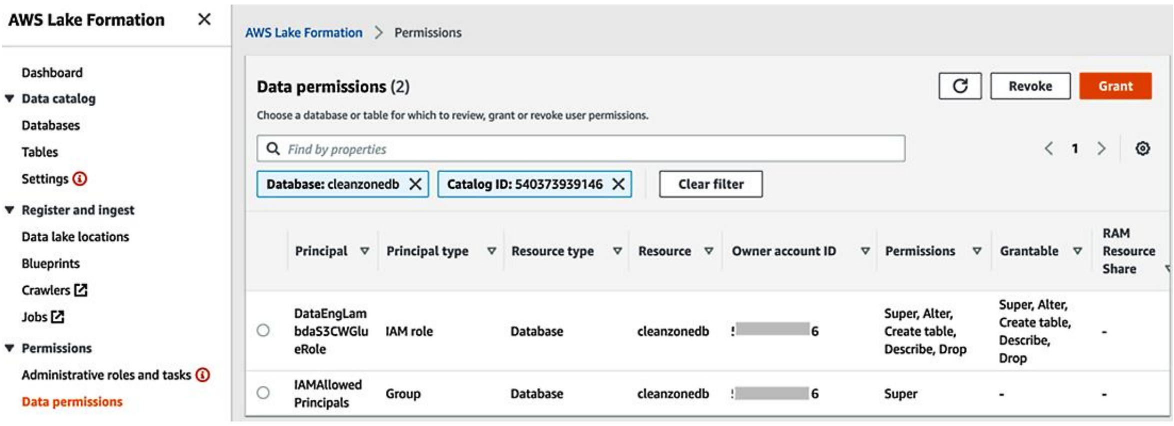
You may add yourself and/or other principals.

☑ Add myself

AWS account: 540373939146

□ Add other AWS users or roles

Select additional IAM users and roles to be data lake administrators.

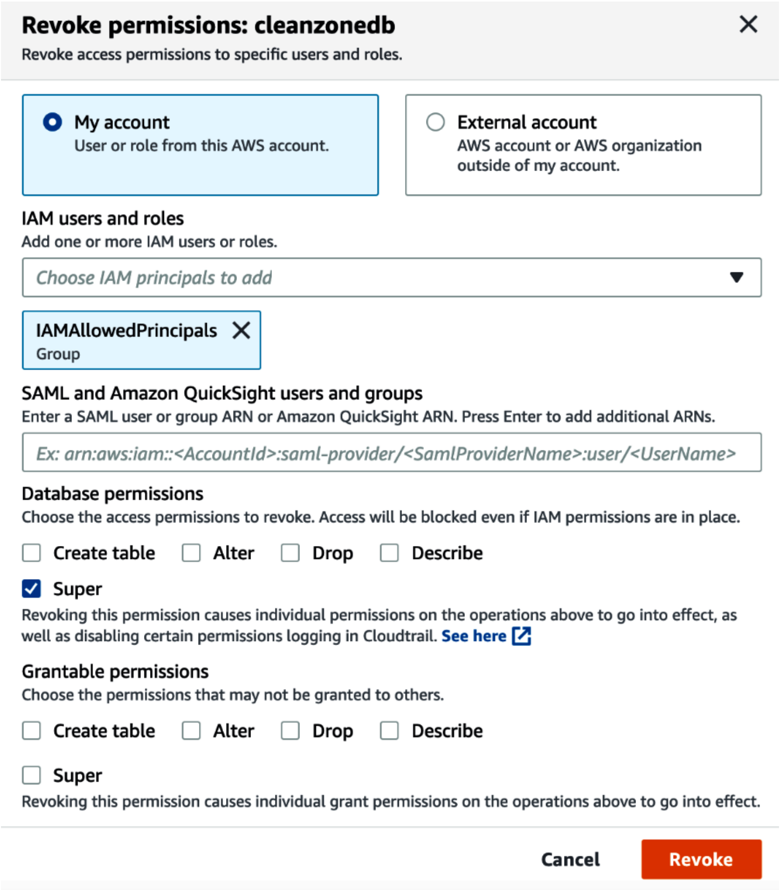


Data permissions (2)

Choose a database or table for which to review, grant or revoke user permissions.

Database: cleanzonedb ✕ Catalog ID: 540373939146 ✕

	Principal	Principal type	Resource type	Resource	Owner account ID	Permissions	Grantable	RAM Resource Share
○	DataEngLambdaS3CWGlueRole	IAM role	Database	cleanzonedb	[redacted]	Super, Alter, Create table, Describe, Drop	Super, Alter, Create table, Describe, Drop	-
○	IAMAllowedPrincipals	Group	Database	cleanzonedb	[redacted]	Super	-	-



Revoke permissions: cleanzonedb

Revoke access permissions to specific users and roles.

● **My account**
User or role from this AWS account.

○ **External account**
AWS account or AWS organization outside of my account.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add

IAMAllowedPrincipals
Group

SAML and Amazon QuickSight users and groups
Enter a SAML user or group ARN or Amazon QuickSight ARN. Press Enter to add additional ARNs.

Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Database permissions
Choose the access permissions to revoke. Access will be blocked even if IAM permissions are in place.

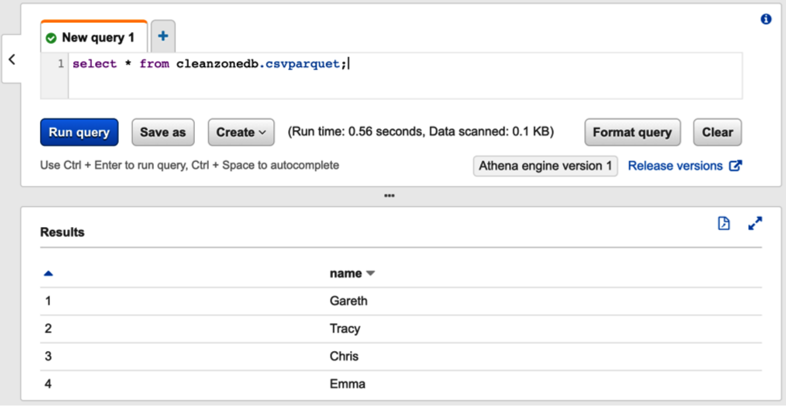
□ Create table □ Alter □ Drop □ Describe

☑ Super
Revoking this permission causes individual permissions on the operations above to go into effect, as well as disabling certain permissions logging in Cloudtrail. **See here**

Grantable permissions
Choose the permissions that may not be granted to others.

□ Create table □ Alter □ Drop □ Describe

□ Super
Revoking this permission causes individual grant permissions on the operations above to go into effect.



```
select * from cleanzonedb.csvparquet;
```

(Run time: 0.56 seconds, Data scanned: 0.1 KB)

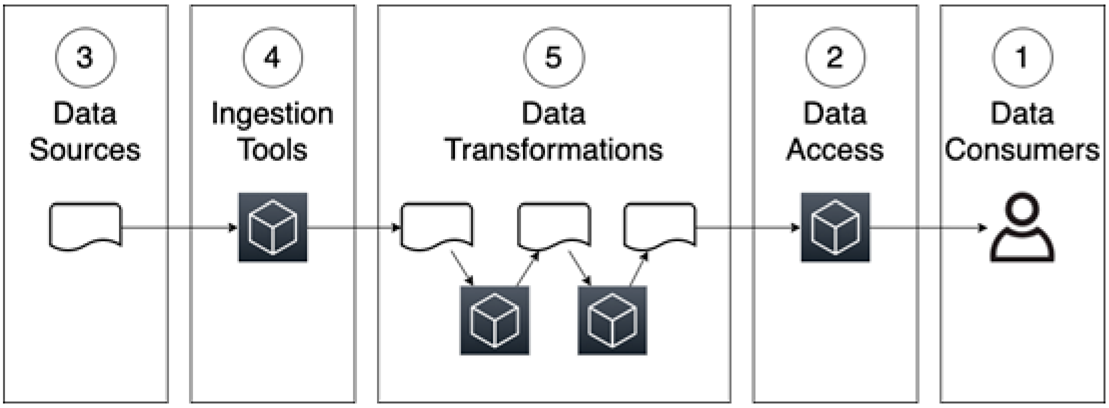
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

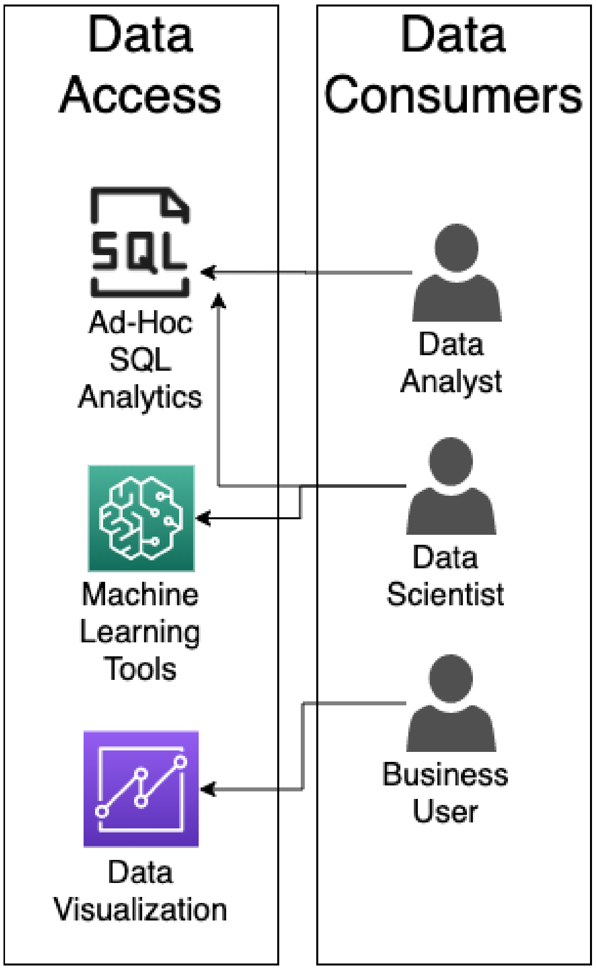
Athena engine version 1 Release versions

Results

	name
1	Gareth
2	Tracy
3	Chris
4	Emma

Chapter 5: Architecting Data Engineering Pipelines



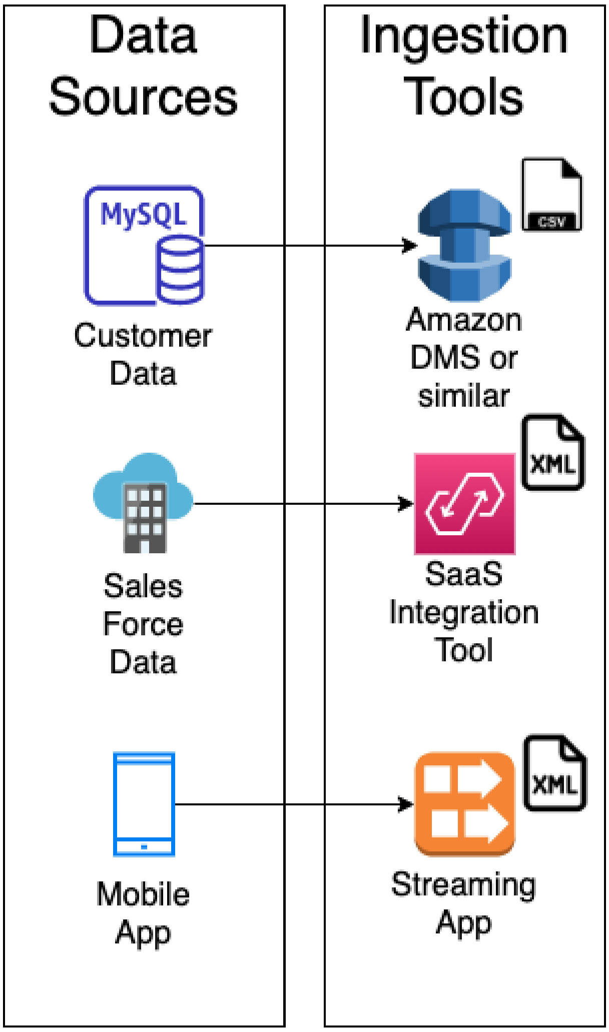


Notes:

- **Data Analysts:** A team of 4 - 6 data analysts will be responsible for creating reports and drawing insights from the data that will be delivered to senior Sales Management. This team has experience in using SQL.

- **Data Scientists:** A team of 2 - 3 data scientists will be tasked with creating Machine Learning models based on historical data that is part of this project. This team wants SQL access for exploring the data, as well as access to specialized machine learning tools

- **Business Users:** This project will enable sales operations teams across the country. Total users approx 25 - 30. They want easy access to summarized data via a data visualization tool that lets them filter, drill-down, work with different graphs, etc. Some of this team have experience with Tableau for visualization, but we do not currently have enough licenses for all business users. Open to exploring alternate tools.

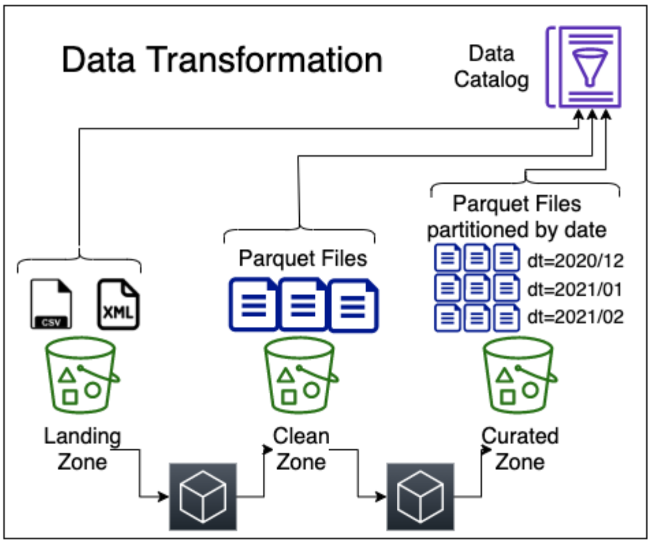


Notes:

- **CUSTOMER DATA**: stored in MySQL database. **System owner**: Database Team. **Data owner:** Marketing Team. **Data Load Frequency:** Daily. **Ingestion:** Investigate DMS for replication with Glue job to consolidate changes

- **SALES FORCE DATA:** Company has Sales Force SaaS subscription. Project needs opportunity data loaded from Sales Force. **System owner:** SalesForce Admin team. **Data owner:** Enterprise Sales Team. **Data Load Frequency:** Hourly. **Ingestion:** SaaS Integration Tool (AWS AppFlow or SalesForce dataloader.io are possibilities)

- **MOBILE APP:** Need to ingest metrics in real-time from companies mobile app used by sales team. **System owner:** AppDev team. **Data owner:** Enterprise Sales Team. **Data Load Frequency:** Near real-time. **Ingestion**: Streaming service (Kinesis or MSK are possibilities)



Notes:

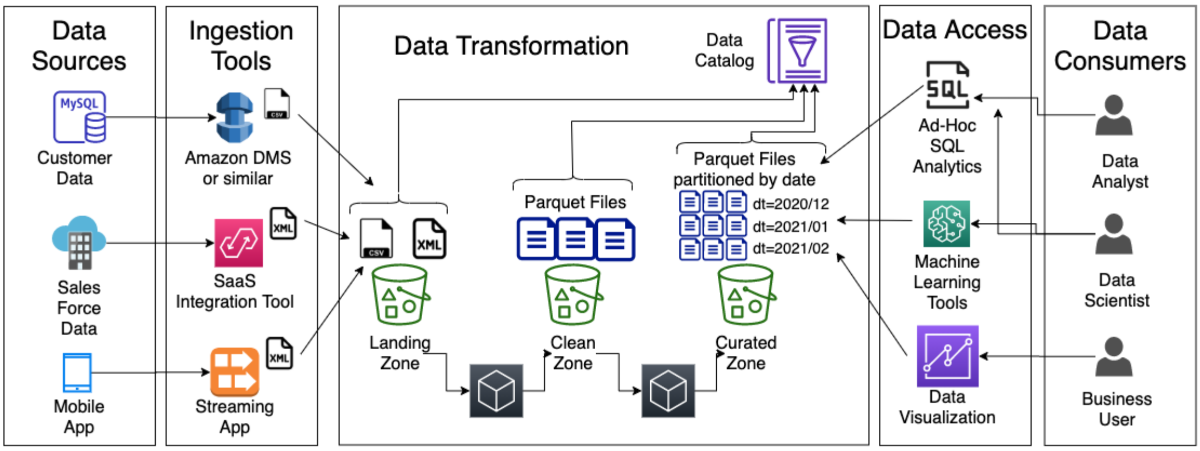
- Data is ingested into the landing zone in raw format (CSV and XML)

- Quality checks are performed and data standardized. Data files are also converted to Parquet format. Resulting files are written to the clean zone

- Data from relational database tables are denormalized based on use case requirements. New tables are enriched with additional data. Data is partitioned by year and month and written to the curated zone.

- As files are written into each zone, the data is catalogged and meta-data is written to the data catalog.

- The team does have previous experience with creating Spark ETL jobs written in PySpark



Data Sources

- **CUSTOMER DATA**: stored in MySQL database. **System owner**: Database Team. **Data owner:** Marketing Team. **Data Load Frequency:** Daily. **Ingestion:** Investigate DMS for replication with Glue job to consolidate changes

- **SALES FORCE DATA:** Company has Sales Force SaaS subscription. Project needs opportunity data loaded from Sales Force. **System owner:** SalesForce Admin team. **Data owner:** Enterprise Sales Team. **Data Load Frequency:** Hourly. **Ingestion:** SaaS Integration Tool (AWS AppFlow or SalesForce dataloader.io are possibilities)

- **MOBILE APP:** Need to ingest metrics in real-time from companies mobile app used by sales team. **System owner:** AppDev team. **Data owner:** Enterprise Sales Team. **Data Load Frequency:** Near real-time. **Ingestion**: Streaming service (Kinesis or MSK are possibilities)

Data Transformation

- Data is ingested into the landing zone in raw format (CSV and XML)

- Quality checks are performed and data standardized. Data files are also converted to Parquet format. Resulting files are written to the clean zone

- Data from relational database tables are denormalized based on use case requirements. New tables are enriched with additional data. Data is partitioned by year and month and written to the curated zone.

- As files are written into each zone, the data is catalogged and meta-data is written to the data catalog.

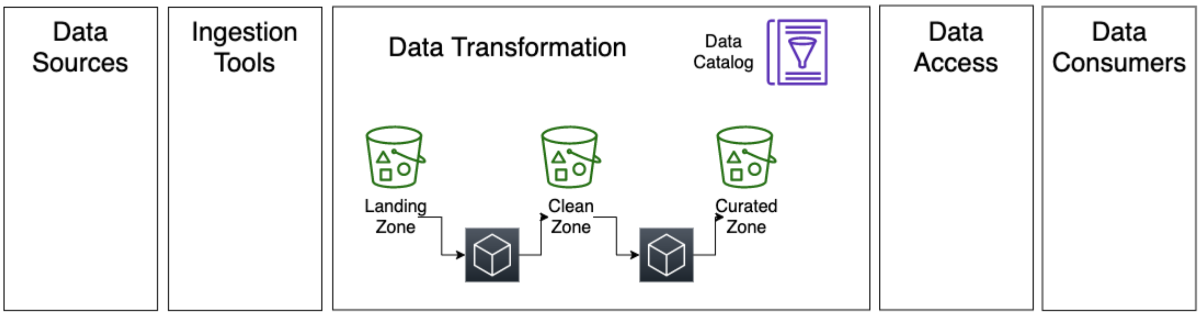
- The team does have previous experience with creating Spark ETL jobs written in PySpark

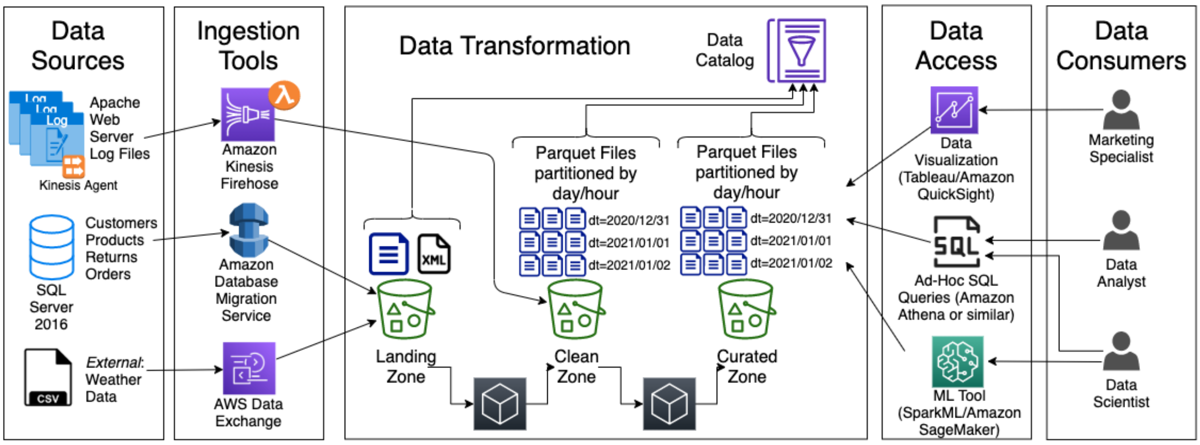
Data Consumers

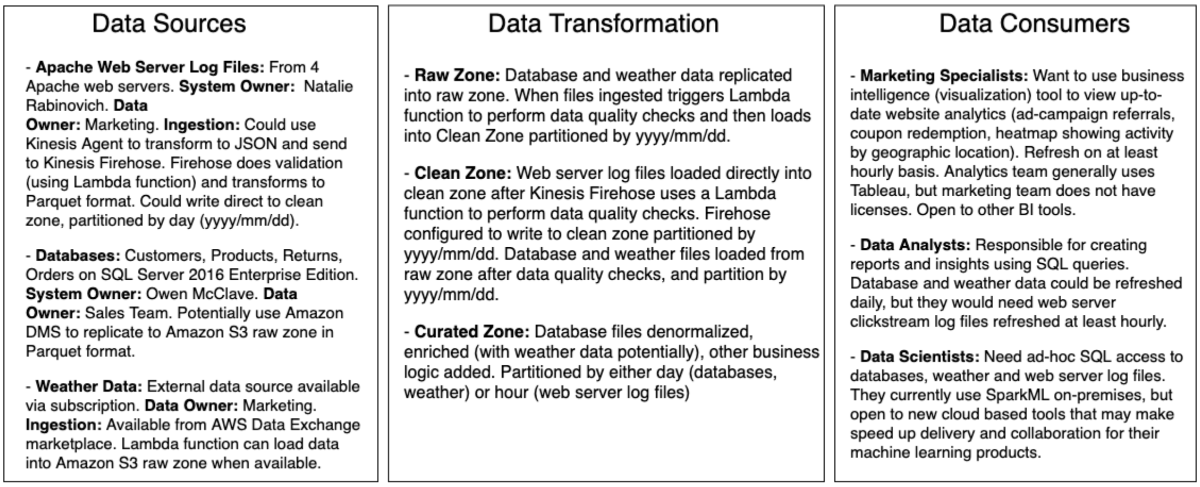
- **DATA ANALYSTS:** A team of 4 - 6 data analysts will be responsible for creating reports and drawing insights from the data that will be delivered to senior Sales Management. This team has experience in using SQL.

- **DATA SCIENTISTS:** A team of 2 - 3 data scientists will be tasked with creating Machine Learning models based on historical data that is part of this project. This team wants SQL access for exploring the data, as well as access to specialized machine learning tools

- **BUSINESS USERS:** This project will enable sales operations teams across the country. Total users approx 25 - 30. They want easy access to summarized data via a data visualization tool that lets them filter, drill-down, work with different graphs, etc. Some of this team have experience with Tableau for visualization, but we do not currently have enough licenses for all business users. Open to exploring alternate tools.







Data Sources

- **Apache Web Server Log Files:** From 4 Apache web servers. **System Owner:** Natalie Rabinovich. **Data Owner:** Marketing. **Ingestion:** Could use Kinesis Agent to transform to JSON and send to Kinesis Firehose. Firehose does validation (using Lambda function) and transforms to Parquet format. Could write direct to clean zone, partitioned by day (yyyy/mm/dd).

- **Databases:** Customers, Products, Returns, Orders on SQL Server 2016 Enterprise Edition. **System Owner:** Owen McClave. **Data Owner:** Sales Team. Potentially use Amazon DMS to replicate to Amazon S3 raw zone in Parquet format.

- **Weather Data:** External data source available via subscription. **Data Owner:** Marketing. **Ingestion:** Available from AWS Data Exchange marketplace. Lambda function can load data into Amazon S3 raw zone when available.

Data Transformation

- **Raw Zone:** Database and weather data replicated into raw zone. When files ingested triggers Lambda function to perform data quality checks and then loads into Clean Zone partitioned by yyyy/mm/dd.

- **Clean Zone:** Web server log files loaded directly into clean zone after Kinesis Firehose uses a Lambda function to perform data quality checks. Firehose configured to write to clean zone partitioned by yyyy/mm/dd. Database and weather files loaded from raw zone after data quality checks, and partition by yyyy/mm/dd.

- **Curated Zone:** Database files denormalized, enriched (with weather data potentially), other business logic added. Partitioned by either day (databases, weather) or hour (web server log files)

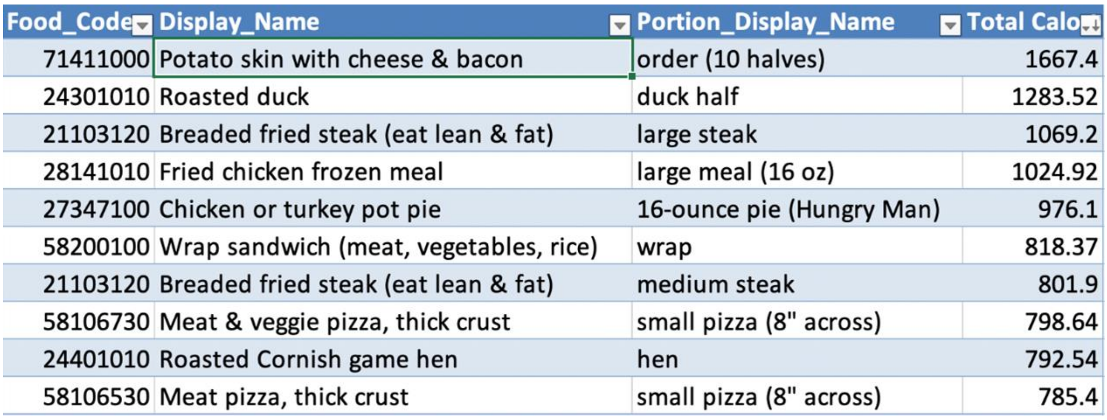
Data Consumers

- **Marketing Specialists:** Want to use business intelligence (visualization) tool to view up-to-date website analytics (ad-campaign referrals, coupon redemption, heatmap showing activity by geographic location). Refresh on at least hourly basis. Analytics team generally uses Tableau, but marketing team does not have licenses. Open to other BI tools.

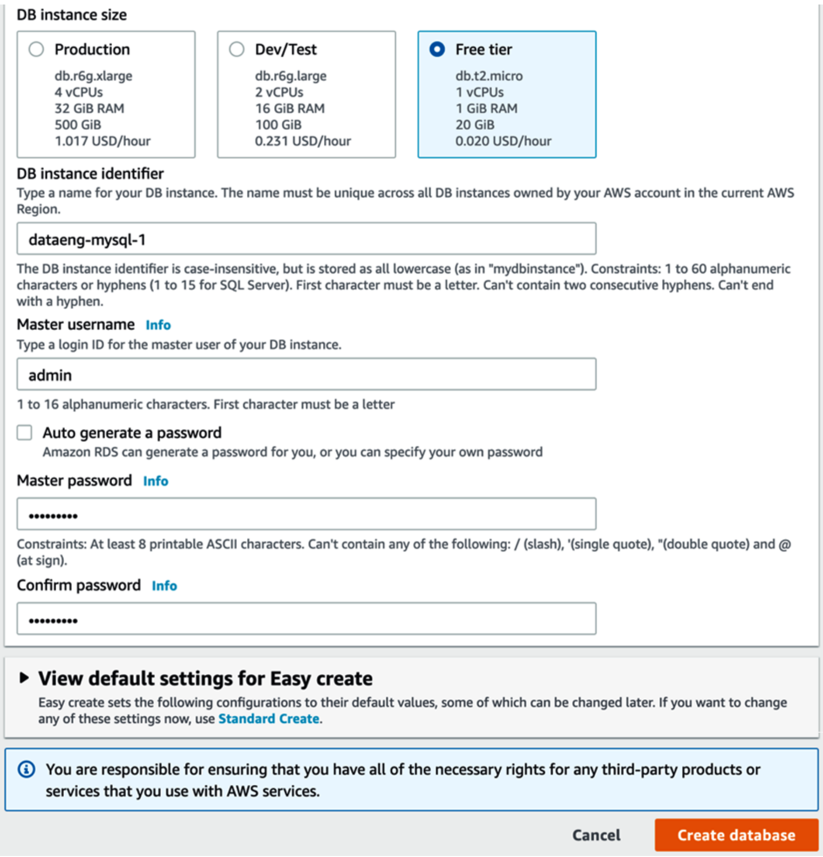
- **Data Analysts:** Responsible for creating reports and insights using SQL queries. Database and weather data could be refreshed daily, but they would need web server clickstream log files refreshed at least hourly.

- **Data Scientists:** Need ad-hoc SQL access to databases, weather and web server log files. They currently use SparkML on-premises, but open to new cloud based tools that may make speed up delivery and collaboration for their machine learning products.

Chapter 6: Ingesting Batch and Streaming Data



Food_Code	Display_Name	Portion_Display_Name	Total Calo
71411000	Potato skin with cheese & bacon	order (10 halves)	1667.4
24301010	Roasted duck	duck half	1283.52
21103120	Breaded fried steak (eat lean & fat)	large steak	1069.2
28141010	Fried chicken frozen meal	large meal (16 oz)	1024.92
27347100	Chicken or turkey pot pie	16-ounce pie (Hungry Man)	976.1
58200100	Wrap sandwich (meat, vegetables, rice)	wrap	818.37
21103120	Breaded fried steak (eat lean & fat)	medium steak	801.9
58106730	Meat & veggie pizza, thick crust	small pizza (8" across)	798.64
24401010	Roasted Cornish game hen	hen	792.54
58106530	Meat pizza, thick crust	small pizza (8" across)	785.4



DB instance size

○ Production	○ Dev/Test	● Free tier
db.r6g.xlarge 4 vCPUs 32 GiB RAM 500 GiB 1.017 USD/hour	db.r6g.large 2 vCPUs 16 GiB RAM 100 GiB 0.231 USD/hour	db.t2.micro 1 vCPUs 1 GiB RAM 20 GiB 0.020 USD/hour

DB instance identifier

Type a name for your DB instance. The name must be unique across all DB instances owned by your AWS account in the current AWS Region.

dataeng-mysql-1

The DB instance identifier is case-insensitive, but is stored as all lowercase (as in "mydbinstance"). Constraints: 1 to 60 alphanumeric characters or hyphens (1 to 15 for SQL Server). First character must be a letter. Can't contain two consecutive hyphens. Can't end with a hyphen.

Master username Info

Type a login ID for the master user of your DB instance.

admin

1 to 16 alphanumeric characters. First character must be a letter

□ Auto generate a password

Amazon RDS can generate a password for you, or you can specify your own password

Master password Info

•••••••••

Constraints: At least 8 printable ASCII characters. Can't contain any of the following: / (slash), '(single quote), "(double quote) and @ (at sign).

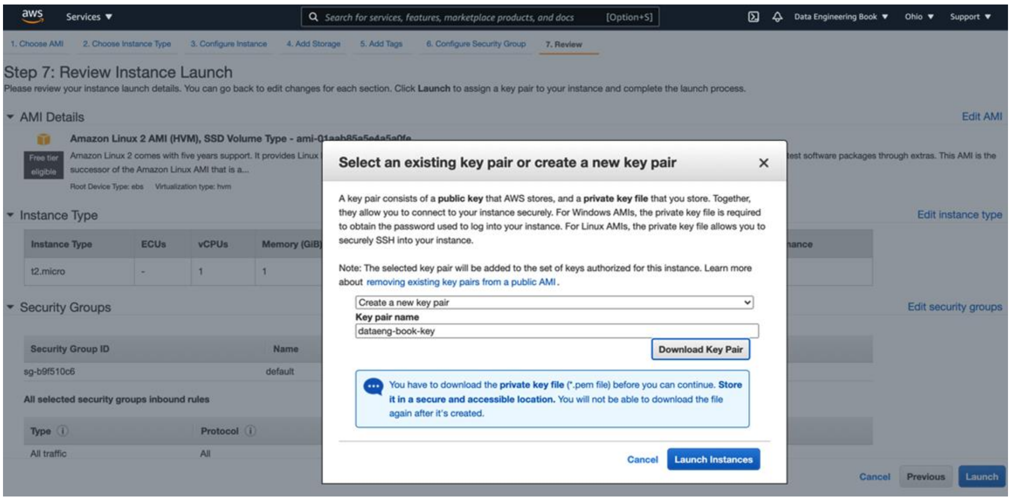
Confirm password Info

•••••••••

▸ **View default settings for Easy create**

Easy create sets the following configurations to their default values, some of which can be changed later. If you want to change any of these settings now, use Standard Create.

ⓘ You are responsible for ensuring that you have all of the necessary rights for any third-party products or services that you use with AWS services.



Step 7: Review Instance Launch

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

▾ AMI Details

Amazon Linux 2 AMI (HVM), SSD Volume Type - ami-01aab85a5e4a5a0fe

Amazon Linux 2 comes with five years support. It provides Linux … est software packages through extras. This AMI is the successor of the Amazon Linux AMI that is a...

Root Device Type: ebs Virtualization type: hvm

▾ Instance Type

Instance Type	ECUs	vCPUs	Memory (GiB)
t2.micro	-	1	1

▾ Security Groups

Security Group ID	Name
sg-b9f510c6	default

All selected security groups inbound rules

Type	Protocol
All traffic	All

Select an existing key pair or create a new key pair ×

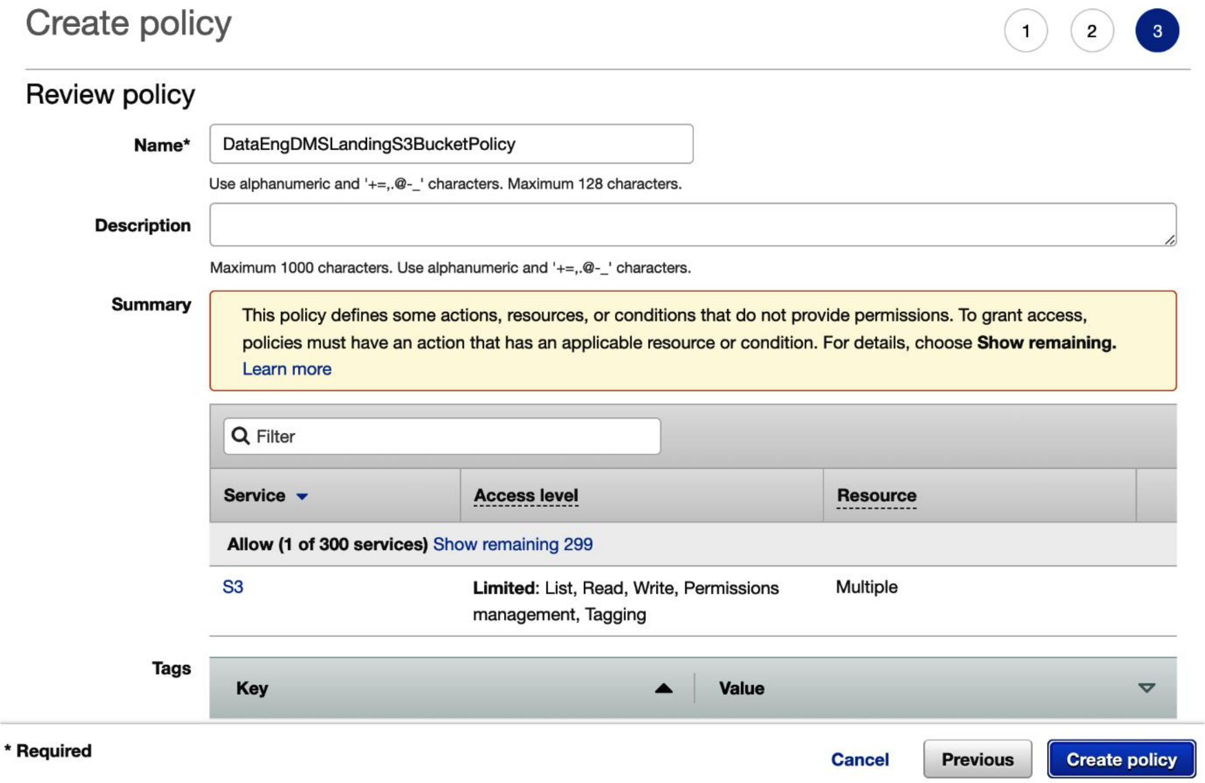
A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about removing existing key pairs from a public AMI.

Create a new key pair

Key pair name: dataeng-book-key

You have to download the **private key file** (*.pem file) before you can continue. **Store it in a secure and accessible location.** You will not be able to download the file again after it's created.



Create policy

Review policy

Name*: DataEngDMSLandingS3BucketPolicy

Use alphanumeric and '+=,.@-_' characters. Maximum 128 characters.

Description

Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Summary

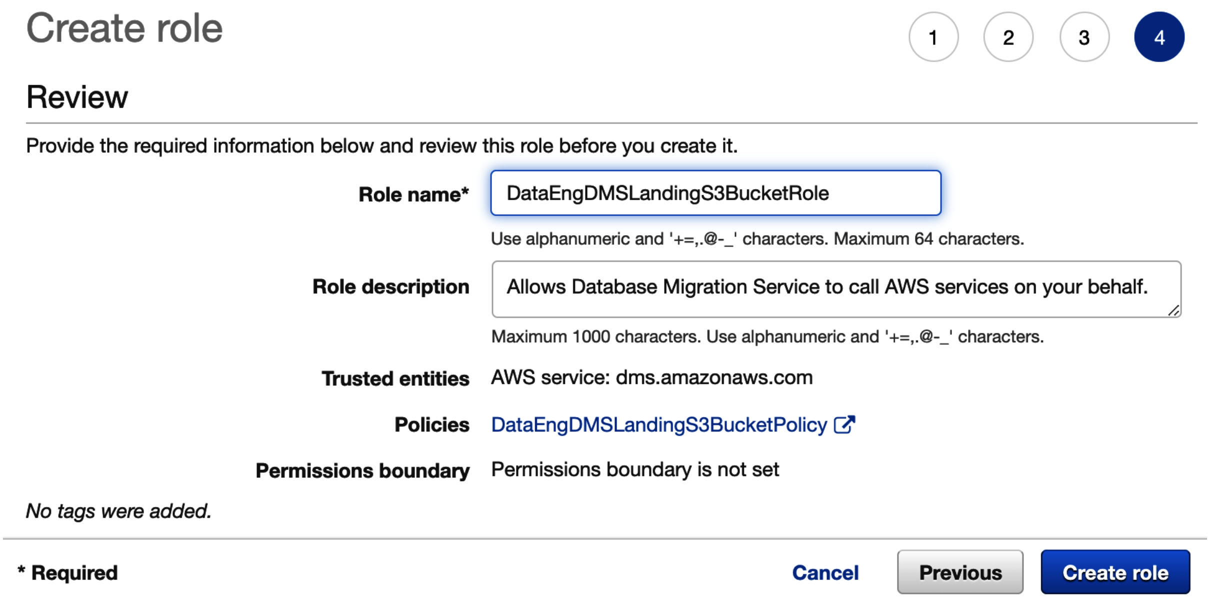
This policy defines some actions, resources, or conditions that do not provide permissions. To grant access, policies must have an action that has an applicable resource or condition. For details, choose **Show remaining.** Learn more

Service ▾	Access level	Resource
Allow (1 of 300 services) Show remaining 299		
S3	**Limited**: List, Read, Write, Permissions management, Tagging	Multiple

Tags

Key ▴	Value

*** Required**



Create role

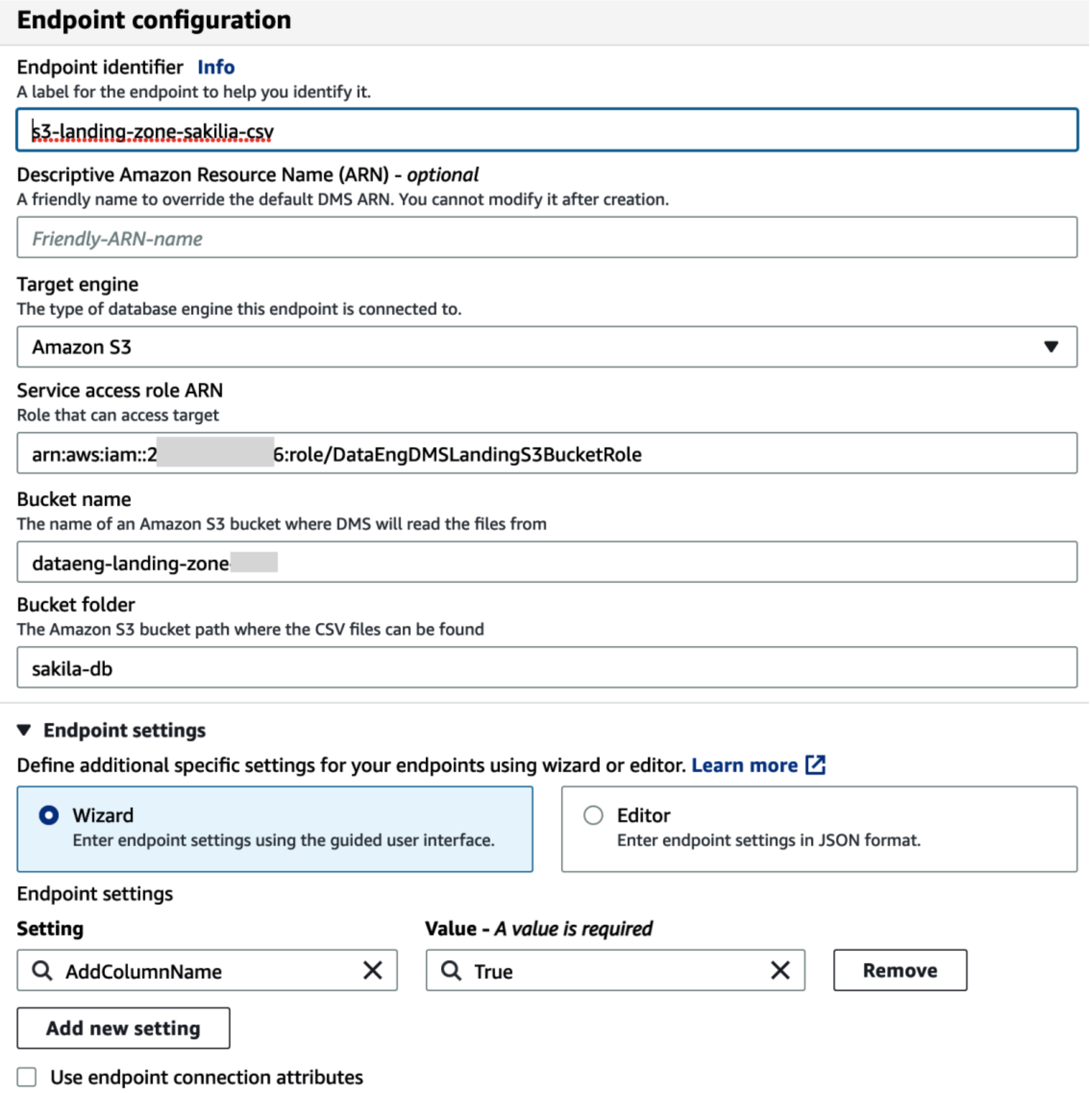
Review

Provide the required information below and review this role before you create it.

Role name*	DataEngDMSLandingS3BucketRole Use alphanumeric and '+=,.@-_' characters. Maximum 64 characters.
Role description	Allows Database Migration Service to call AWS services on your behalf. Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.
Trusted entities	AWS service: dms.amazonaws.com
Policies	DataEngDMSLandingS3BucketPolicy
Permissions boundary	Permissions boundary is not set

No tags were added.

*** Required**



Endpoint configuration

Endpoint identifier Info
A label for the endpoint to help you identify it.

s3-landing-zone-sakilia-csv

Descriptive Amazon Resource Name (ARN) - *optional*
A friendly name to override the default DMS ARN. You cannot modify it after creation.

Friendly-ARN-name

Target engine
The type of database engine this endpoint is connected to.

Amazon S3

Service access role ARN
Role that can access target

arn:aws:iam::2[redacted]6:role/DataEngDMSLandingS3BucketRole

Bucket name
The name of an Amazon S3 bucket where DMS will read the files from

dataeng-landing-zone[redacted]

Bucket folder
The Amazon S3 bucket path where the CSV files can be found

sakila-db

▼ Endpoint settings

Define additional specific settings for your endpoints using wizard or editor. Learn more

● Wizard
Enter endpoint settings using the guided user interface.

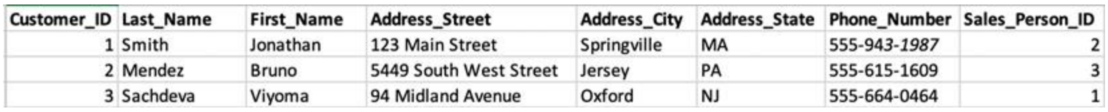
○ Editor
Enter endpoint settings in JSON format.

Endpoint settings

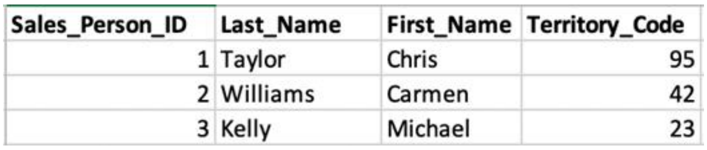
Setting	Value - *A value is required*
AddColumnName	True

□ Use endpoint connection attributes

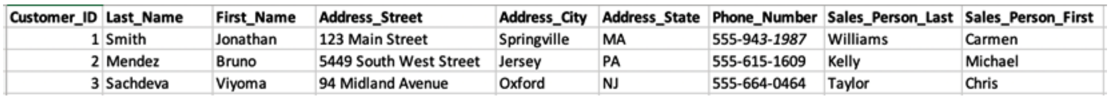
Chapter 7: Transforming Data to Optimize for Analytics



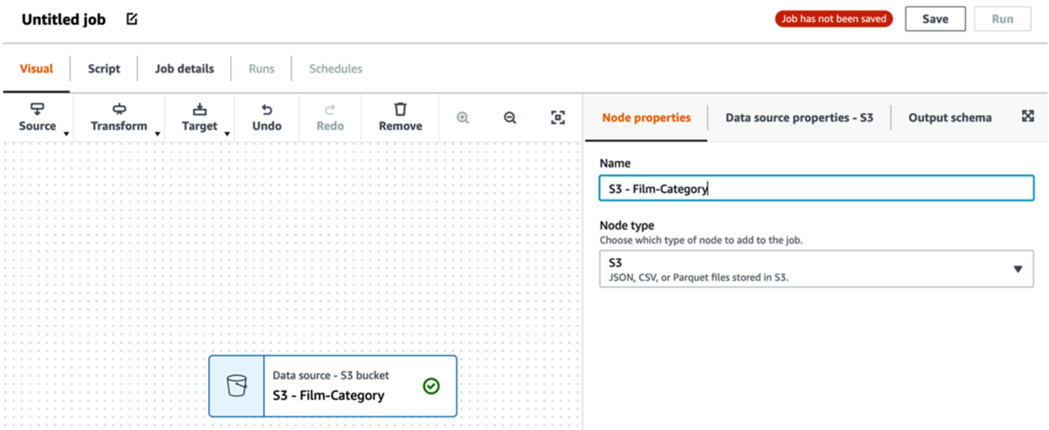
Customer_ID	Last_Name	First_Name	Address_Street	Address_City	Address_State	Phone_Number	Sales_Person_ID
1	Smith	Jonathan	123 Main Street	Springville	MA	555-943-*1987*	2
2	Mendez	Bruno	5449 South West Street	Jersey	PA	555-615-1609	3
3	Sachdeva	Viyoma	94 Midland Avenue	Oxford	NJ	555-664-0464	1



Sales_Person_ID	Last_Name	First_Name	Territory_Code
1	Taylor	Chris	95
2	Williams	Carmen	42
3	Kelly	Michael	23



Customer_ID	Last_Name	First_Name	Address_Street	Address_City	Address_State	Phone_Number	Sales_Person_Last	Sales_Person_First
1	Smith	Jonathan	123 Main Street	Springville	MA	555-943-*1987*	Williams	Carmen
2	Mendez	Bruno	5449 South West Street	Jersey	PA	555-615-1609	Kelly	Michael
3	Sachdeva	Viyoma	94 Midland Avenue	Oxford	NJ	555-664-0464	Taylor	Chris



Name

S3 - Film-Category

Node type

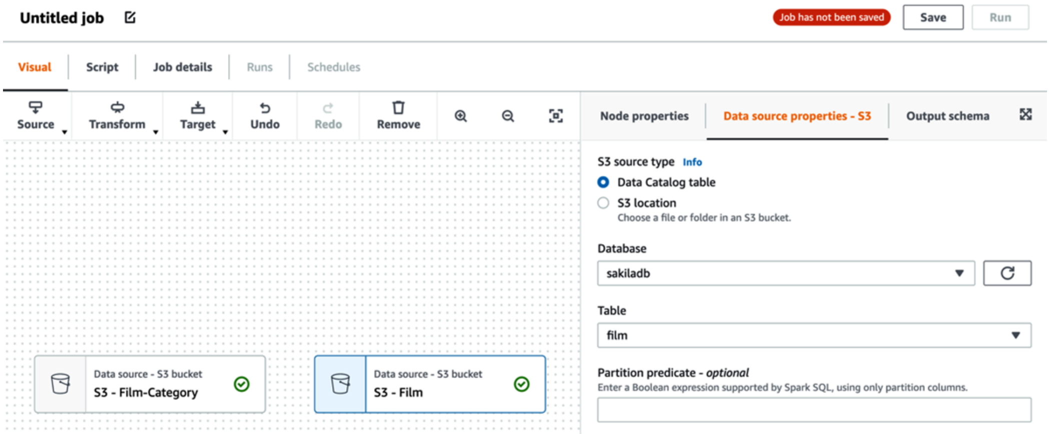
Choose which type of node to add to the job.

S3

JSON, CSV, or Parquet files stored in S3.

Data source - S3 bucket

S3 - Film-Category



S3 source type Info

● Data Catalog table

○ S3 location

Choose a file or folder in an S3 bucket.

Database

sakiladb

Table

film

Partition predicate - *optional*

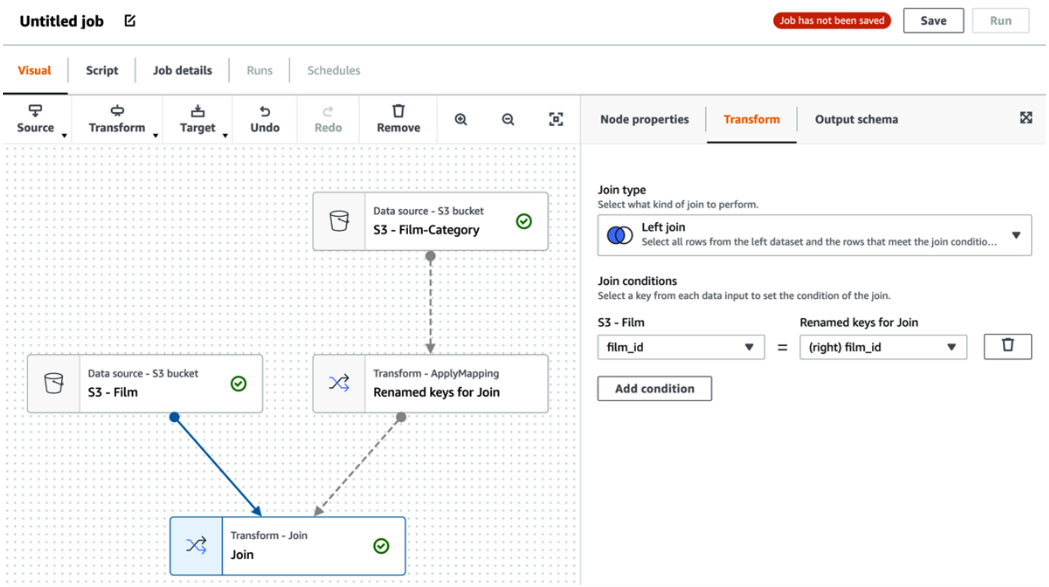
Enter a Boolean expression supported by Spark SQL, using only partition columns.

Data source - S3 bucket

S3 - Film-Category

Data source - S3 bucket

S3 - Film



Untitled job

Job has not been saved · Save · Run

Visual · Script · Job details · Runs · Schedules

Source · Transform · Target · Undo · Redo · Remove

Data source - S3 bucket
S3 - Film-Category

Data source - S3 bucket
S3 - Film

Transform - ApplyMapping
Renamed keys for Join

Transform - Join
Join

Node properties · Transform · Output schema

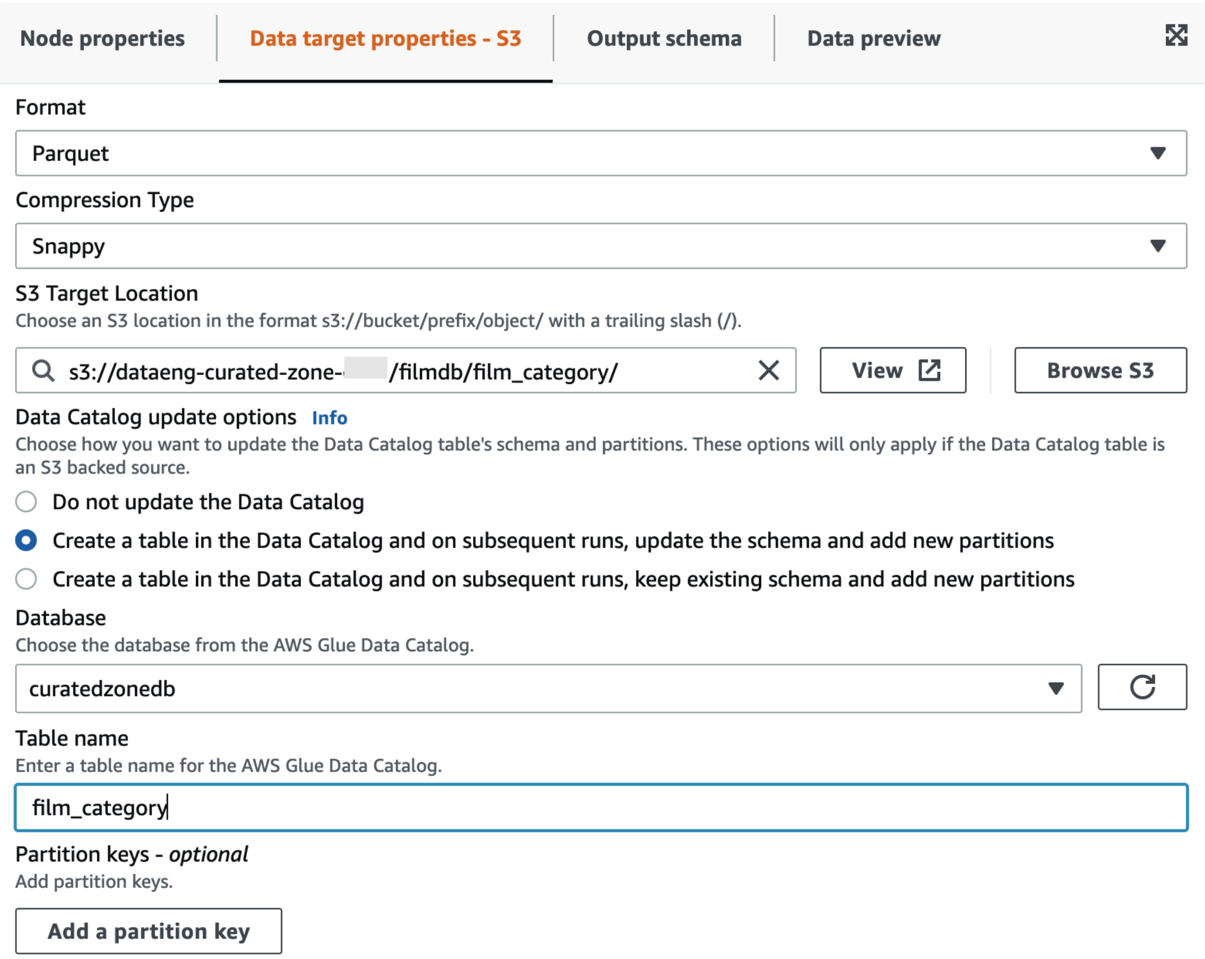
Join type
Select what kind of join to perform.

Left join
Select all rows from the left dataset and the rows that meet the join conditio...

Join conditions
Select a key from each data input to set the condition of the join.

S3 - Film		Renamed keys for Join
film_id	=	(right) film_id

Add condition



Node properties · Data target properties - S3 · Output schema · Data preview

Format

Parquet

Compression Type

Snappy

S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

s3://dataeng-curated-zone-[redacted]/filmdb/film_category/

View · Browse S3

Data Catalog update options Info

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

○ Do not update the Data Catalog

● Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions

○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database

Choose the database from the AWS Glue Data Catalog.

curatedzonedb

Table name

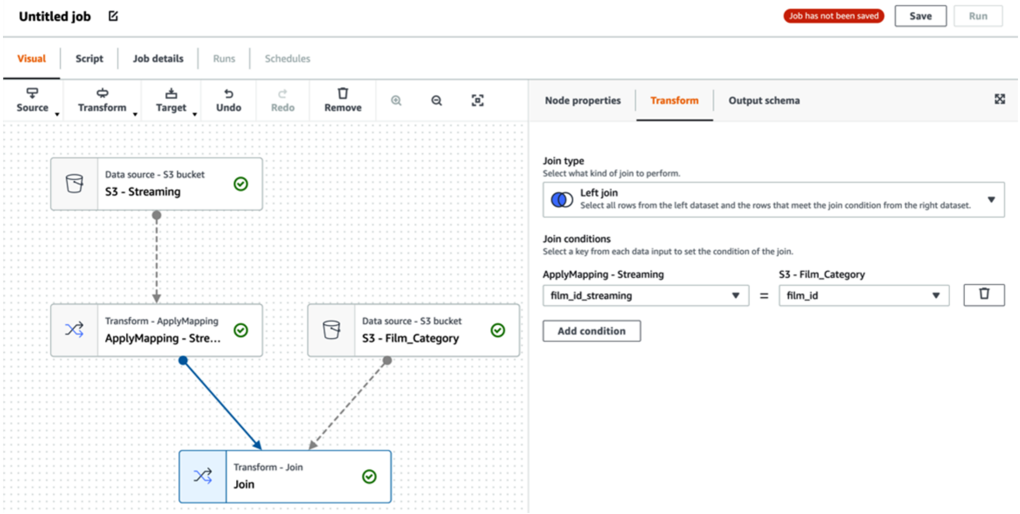
Enter a table name for the AWS Glue Data Catalog.

film_category

Partition keys - *optional*

Add partition keys.

Add a partition key

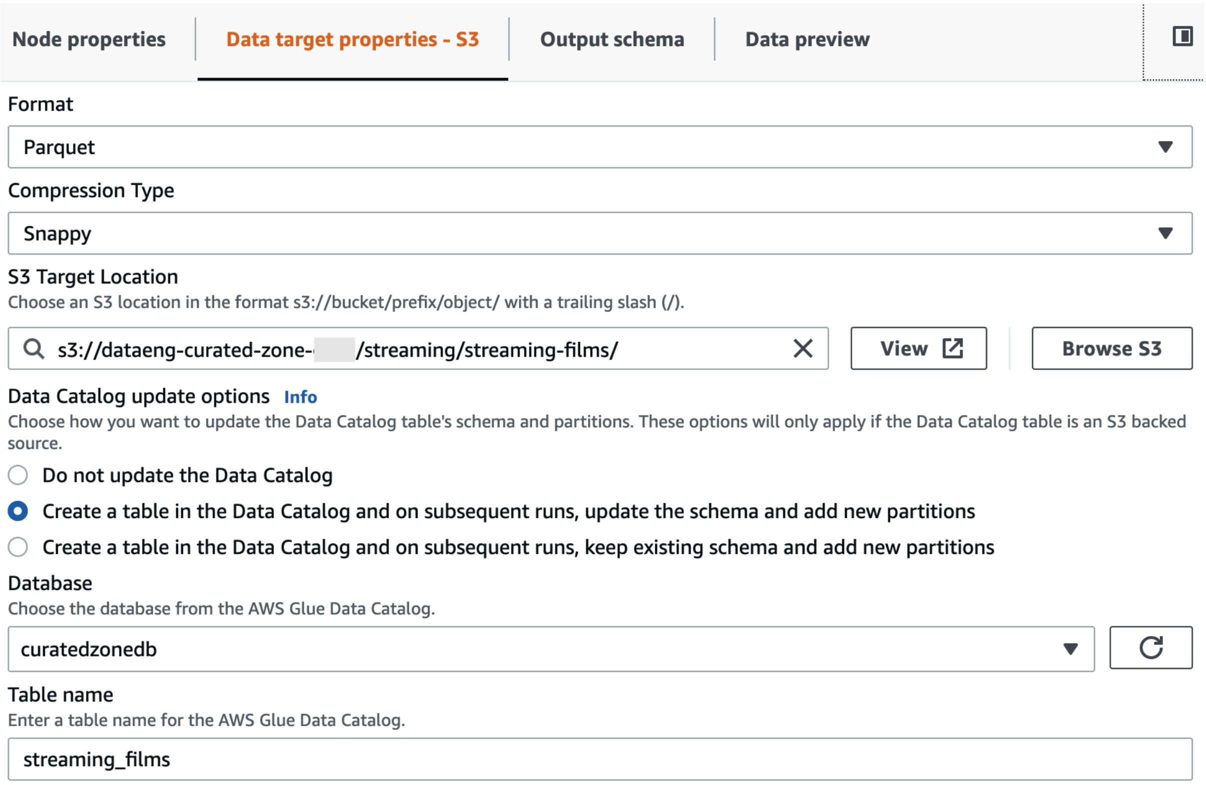


Join type
Select what kind of join to perform.
Left join
Select all rows from the left dataset and the rows that meet the join condition from the right dataset.

Join conditions
Select a key from each data input to set the condition of the join.

ApplyMapping - Streaming		S3 - Film_Category
film_id_streaming	=	film_id

Add condition



Format
Parquet

Compression Type
Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

s3://dataeng-curated-zone-[redacted]/streaming/streaming-films/

View

Browse S3

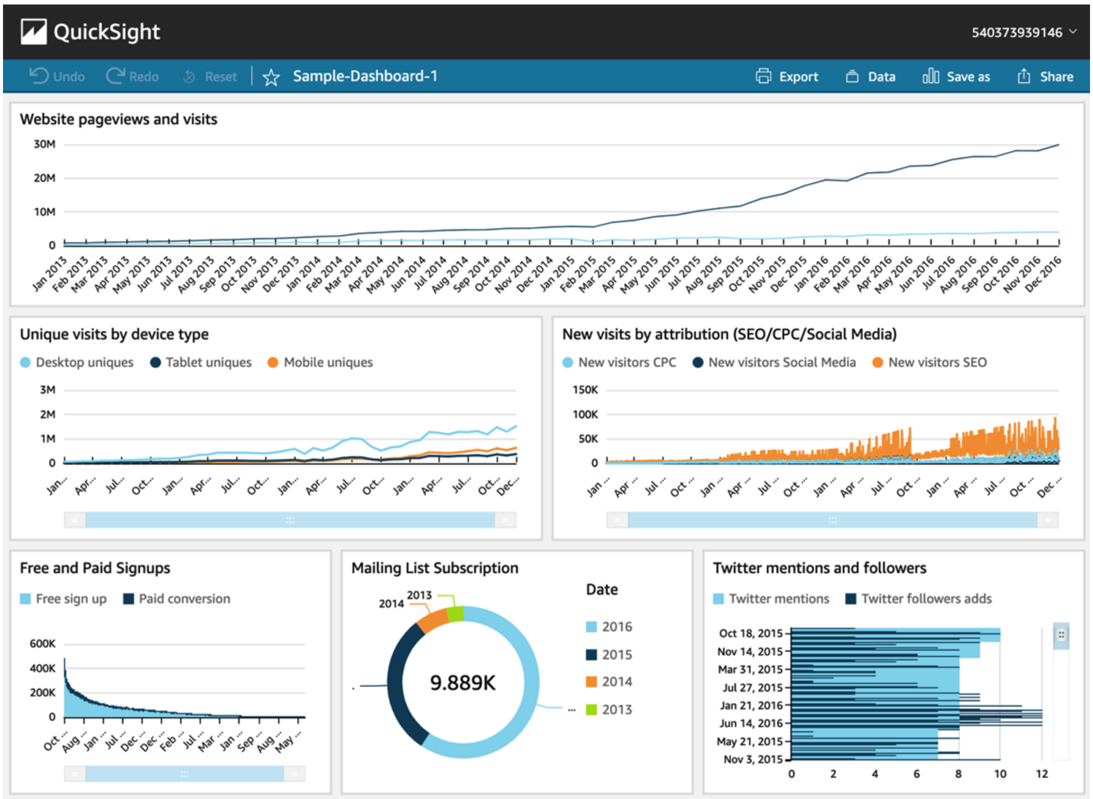
Data Catalog update options Info
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

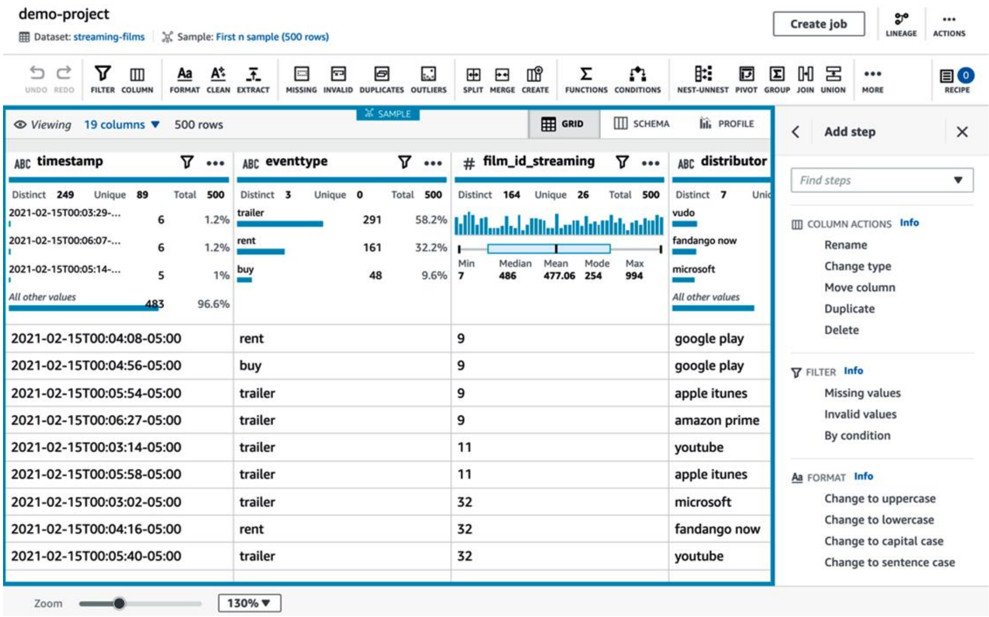
○ Do not update the Data Catalog
● Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database
Choose the database from the AWS Glue Data Catalog.
curatedzonedb

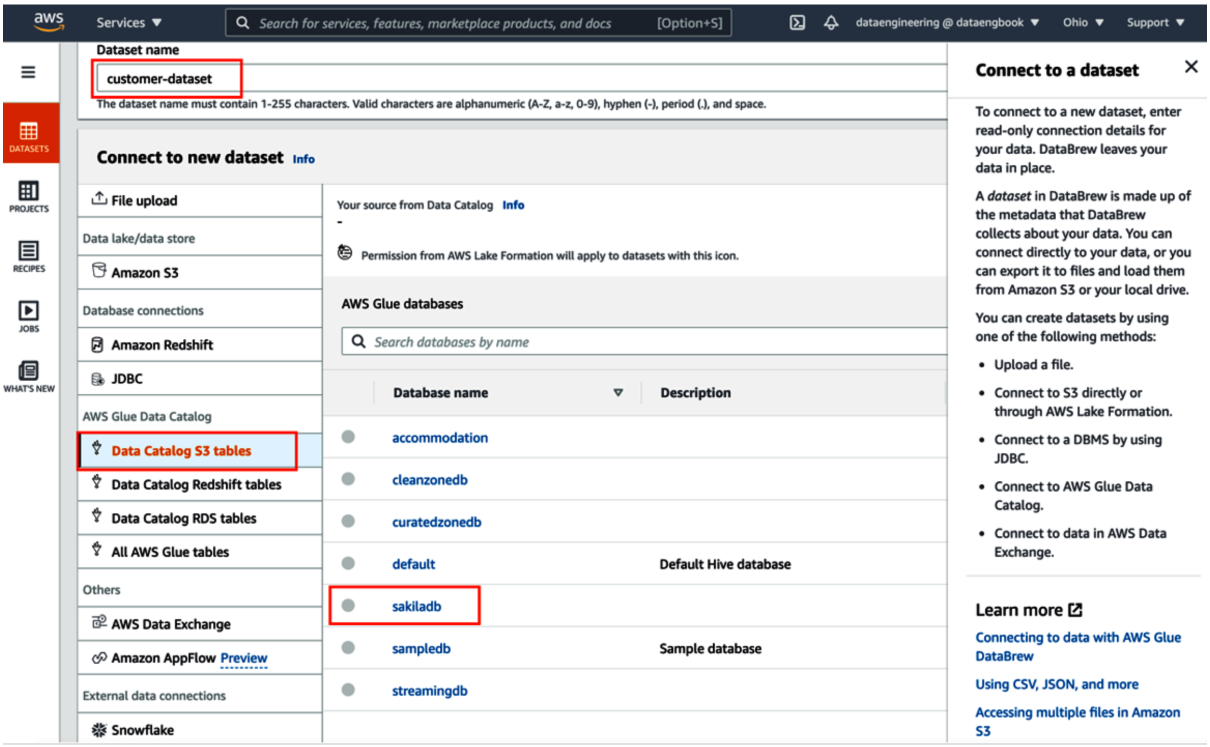
Table name
Enter a table name for the AWS Glue Data Catalog.
streaming_films

Chapter 8: Identifying and Enabling Data Consumers


QuickSight
540373939146
Undo
Redo
Reset
Sample-Dashboard-1
Export
Data
Save as
Share
Website pageviews and visits
30M
20M
10M
0
Jan 2013 Feb 2013 Mar 2013 Apr 2013 May 2013 Jun 2013 Jul 2013 Aug 2013 Sep 2013 Oct 2013 Nov 2013 Dec 2013 Jan 2014 Feb 2014 Mar 2014 Apr 2014 May 2014 Jun 2014 Jul 2014 Aug 2014 Sep 2014 Oct 2014 Nov 2014 Dec 2014 Jan 2015 Feb 2015 Mar 2015 Apr 2015 May 2015 Jun 2015 Jul 2015 Aug 2015 Sep 2015 Oct 2015 Nov 2015 Dec 2015 Jan 2016 Feb 2016 Mar 2016 Apr 2016 May 2016 Jun 2016 Jul 2016 Aug 2016 Sep 2016 Oct 2016 Nov 2016 Dec 2016
Unique visits by device type
Desktop uniques
Tablet uniques
Mobile uniques
3M
2M
1M
0
Jan... Apr... Jul... Oct... Jan... Apr... Jul... Oct... Jan... Apr... Jul... Oct... Jan... Apr... Jul... Oct... Dec...
New visits by attribution (SEO/CPC/Social Media)
New visitors CPC
New visitors Social Media
New visitors SEO
150K
100K
50K
0
Jan ... Apr ... Jul ... Oct ... Jan ... Apr ... Jul ... Oct ... Jan ... Apr ... Jul ... Oct ... Jan ... Apr ... Jul ... Oct ... Dec ...
Free and Paid Signups
Free sign up
Paid conversion
600K
400K
200K
0
Oct ... Aug ... Jan ... Jul ... Dec ... Dec ... Feb ... Jul ... Mar ... Jan ... Sep ... Aug ... May ...
Mailing List Subscription
2013
2014
9.889K
Date
2016
2015
2014
2013
Twitter mentions and followers
Twitter mentions
Twitter followers adds
Oct 18, 2015
Nov 14, 2015
Mar 31, 2015
Jul 27, 2015
Jan 21, 2016
Jun 14, 2016
May 21, 2015
Nov 3, 2015
0
2
4
6
8
10
12



ABC timestamp	ABC eventtype	# film_id_streaming	ABC distributor
Distinct 249 Unique 89 Total 500	Distinct 3 Unique 0 Total 500	Distinct 164 Unique 26 Total 500	Distinct 7 Uniq
2021-02-15T00:03:29-... 6 1.2%	trailer 291 58.2%	Min 7	vudo
2021-02-15T00:06:07-... 6 1.2%	rent 161 32.2%	Median 486	fandango now
2021-02-15T00:05:14-... 5 1%	buy 48 9.6%	Mean 477.06	microsoft
All other values 483 96.6%		Mode 254 Max 994	All other values
2021-02-15T00:04:08-05:00	rent	9	google play
2021-02-15T00:04:56-05:00	buy	9	google play
2021-02-15T00:05:54-05:00	trailer	9	apple itunes
2021-02-15T00:06:27-05:00	trailer	9	amazon prime
2021-02-15T00:03:14-05:00	trailer	11	youtube
2021-02-15T00:05:58-05:00	trailer	11	apple itunes
2021-02-15T00:03:02-05:00	trailer	32	microsoft
2021-02-15T00:04:16-05:00	rent	32	fandango now
2021-02-15T00:05:40-05:00	trailer	32	youtube



aws
Services ▼
Search for services, features, marketplace products, and docs [Option+S]
dataengineering @ dataengbook ▼
Ohio ▼
Support ▼
DATASETS
PROJECTS
RECIPES
JOBS
WHAT'S NEW

Dataset name

customer-dataset

The dataset name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.

Connect to new dataset Info

- File upload
- Data lake/data store
  - Amazon S3
- Database connections
  - Amazon Redshift
  - JDBC
- AWS Glue Data Catalog
  - Data Catalog S3 tables
  - Data Catalog Redshift tables
  - Data Catalog RDS tables
  - All AWS Glue tables
- Others
  - AWS Data Exchange
  - Amazon AppFlow Preview
- External data connections
  - Snowflake

Your source from Data Catalog Info

-

Permission from AWS Lake Formation will apply to datasets with this icon.

AWS Glue databases

Search databases by name

	Database name	Description
○	accommodation	
○	cleanzonedb	
○	curatedzonedb	
○	default	Default Hive database
○	sakiladb	
○	sampledb	Sample database
○	streamingdb	

Connect to a dataset ✕

To connect to a new dataset, enter read-only connection details for your data. DataBrew leaves your data in place.

A *dataset* in DataBrew is made up of the metadata that DataBrew collects about your data. You can connect directly to your data, or you can export it to files and load them from Amazon S3 or your local drive.

You can create datasets by using one of the following methods:

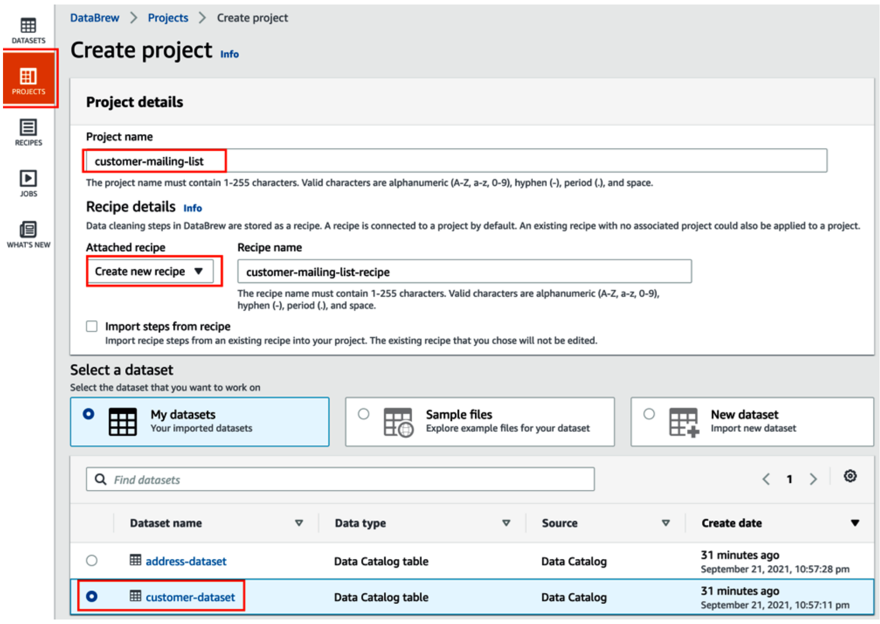
- Upload a file.
- Connect to S3 directly or through AWS Lake Formation.
- Connect to a DBMS by using JDBC.
- Connect to AWS Glue Data Catalog.
- Connect to data in AWS Data Exchange.

Learn more

Connecting to data with AWS Glue DataBrew

Using CSV, JSON, and more

Accessing multiple files in Amazon S3



DataBrew > Projects > Create project

Create project Info

Project details

Project name

customer-mailing-list

The project name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.

Recipe details Info

Data cleaning steps in DataBrew are stored as a recipe. A recipe is connected to a project by default. An existing recipe with no associated project could also be applied to a project.

Attached recipe: Create new recipe ▼

Recipe name: customer-mailing-list-recipe

The recipe name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.

□ Import steps from recipe

Import recipe steps from an existing recipe into your project. The existing recipe that you chose will not be edited.

Select a dataset

Select the dataset that you want to work on

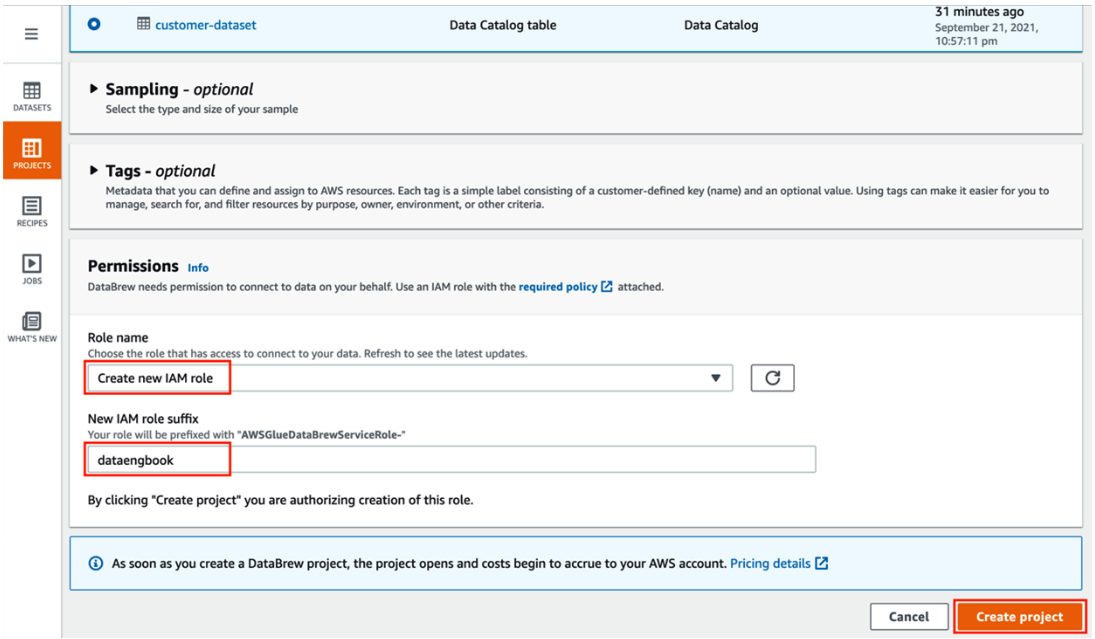
● My datasets — Your imported datasets

○ Sample files — Explore example files for your dataset

○ New dataset — Import new dataset

Find datasets

	Dataset name	Data type	Source	Create date
○	address-dataset	Data Catalog table	Data Catalog	31 minutes ago September 21, 2021, 10:57:28 pm
●	customer-dataset	Data Catalog table	Data Catalog	31 minutes ago September 21, 2021, 10:57:11 pm



● customer-dataset	Data Catalog table	Data Catalog	31 minutes ago September 21, 2021, 10:57:11 pm

▸ **Sampling** - *optional*
Select the type and size of your sample

▸ **Tags** - *optional*
Metadata that you can define and assign to AWS resources. Each tag is a simple label consisting of a customer-defined key (name) and an optional value. Using tags can make it easier for you to manage, search for, and filter resources by purpose, owner, environment, or other criteria.

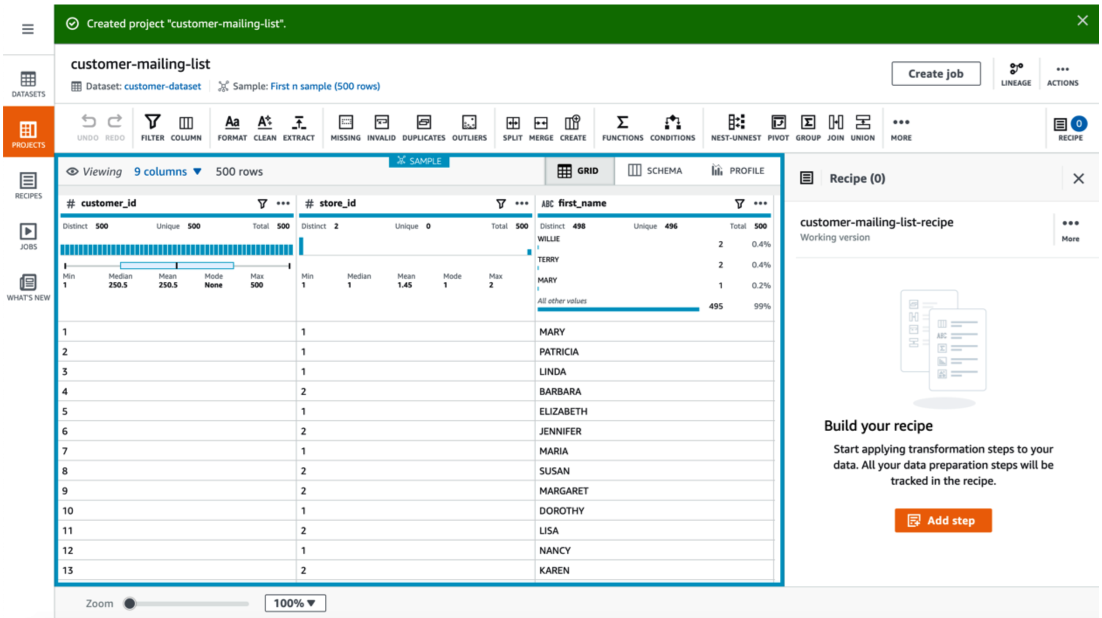
Permissions Info
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the required policy attached.

Role name
Choose the role that has access to connect to your data. Refresh to see the latest updates.
Create new IAM role

New IAM role suffix
Your role will be prefixed with "AWSGlueDataBrewServiceRole-"
dataengbook

By clicking "Create project" you are authorizing creation of this role.

As soon as you create a DataBrew project, the project opens and costs begin to accrue to your AWS account. Pricing details



Created project "customer-mailing-list".

customer-mailing-list
Dataset: customer-dataset   Sample: First n sample (500 rows)

Viewing 9 columns ▼ 500 rows

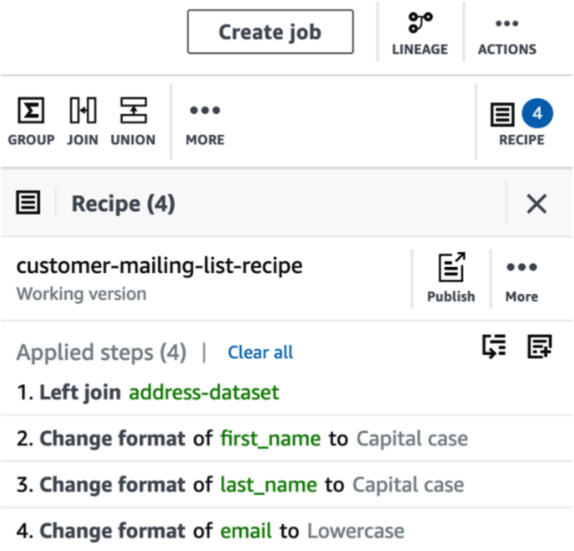
# customer_id	# store_id	ABC first_name
Distinct 500 Unique 500 Total 500 Min 1 Median 250.5 Mean 250.5 Mode None Max 500	Distinct 2 Unique 0 Total 500 Min 1 Median 1 Mean 1.45 Mode 1 Max 2	Distinct 498 Unique 496 Total 500 WILLIE 2 0.4% TERRY 2 0.4% MARY 1 0.2% All other values 495 99%
1	1	MARY
2	1	PATRICIA
3	1	LINDA
4	2	BARBARA
5	1	ELIZABETH
6	2	JENNIFER
7	1	MARIA
8	2	SUSAN
9	2	MARGARET
10	1	DOROTHY
11	2	LISA
12	1	NANCY
13	2	KAREN

Recipe (0)

customer-mailing-list-recipe
Working version

Build your recipe
Start applying transformation steps to your data. All your data preparation steps will be tracked in the recipe.

Add step

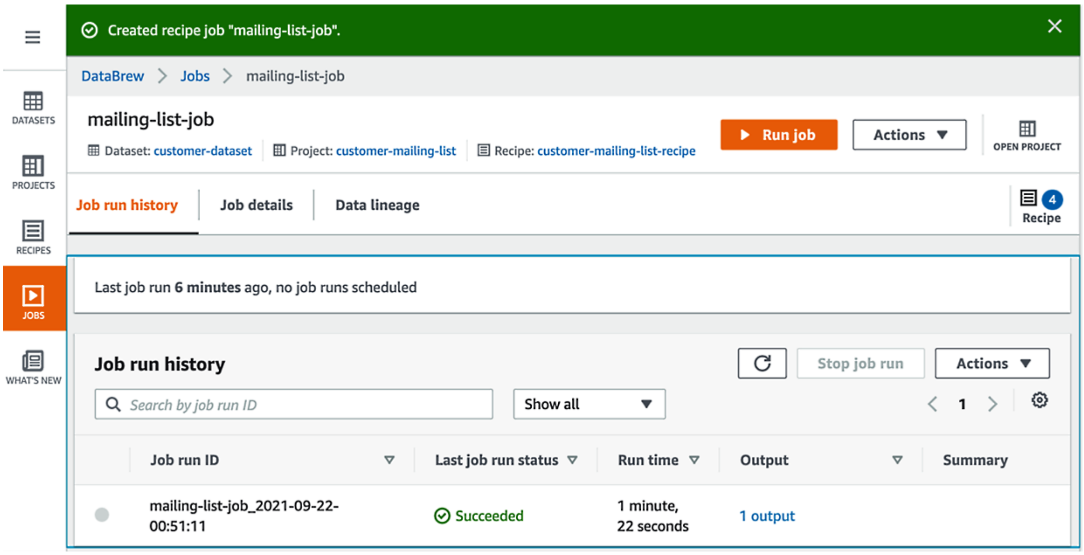


Recipe (4)

customer-mailing-list-recipe
Working version

Applied steps (4) | Clear all

1. Left join address-dataset
2. Change format of first_name to Capital case
3. Change format of last_name to Capital case
4. Change format of email to Lowercase



Created recipe job "mailing-list-job".

DataBrew > Jobs > mailing-list-job

mailing-list-job

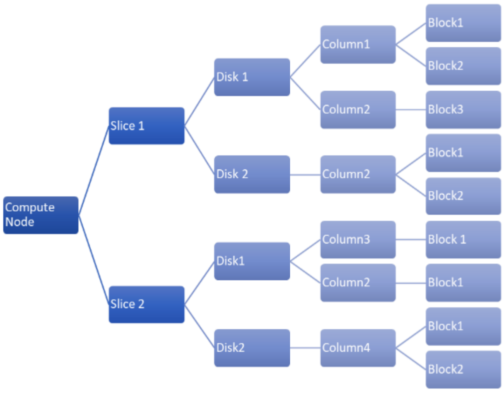
Dataset: customer-dataset | Project: customer-mailing-list | Recipe: customer-mailing-list-recipe

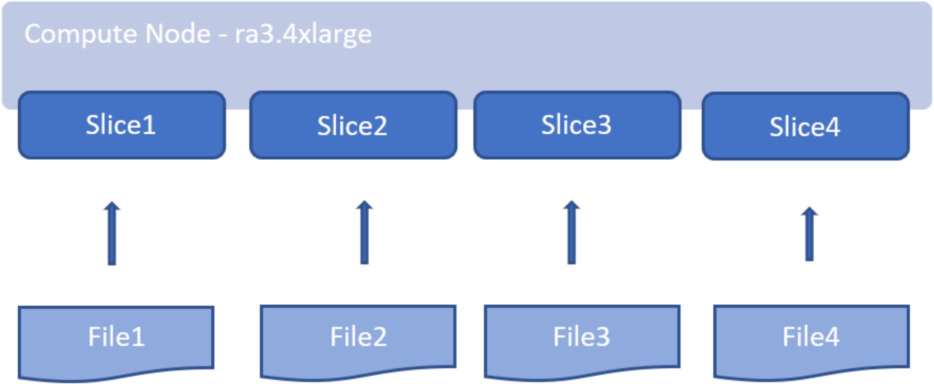
Last job run **6 minutes** ago, no job runs scheduled

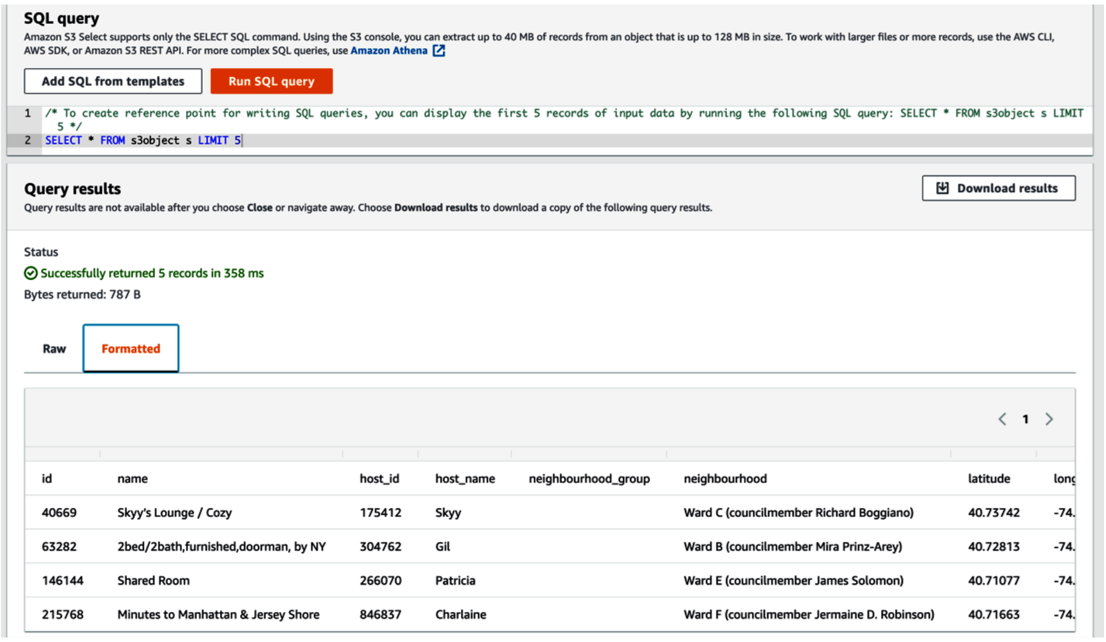
Job run history

Job run ID	Last job run status	Run time	Output	Summary
mailing-list-job_2021-09-22-00:51:11	✓ Succeeded	1 minute, 22 seconds	1 output	

Chapter 9: Loading Data into a Data Mart







SQL query

Amazon S3 Select supports only the SELECT SQL command. Using the S3 console, you can extract up to 40 MB of records from an object that is up to 128 MB in size. To work with larger files or more records, use the AWS CLI, AWS SDK, or Amazon S3 REST API. For more complex SQL queries, use Amazon Athena

```
/* To create reference point for writing SQL queries, you can display the first 5 records of input data by running the following SQL query: SELECT * FROM s3object s LIMIT
  5 */
SELECT * FROM s3object s LIMIT 5
```

Query results

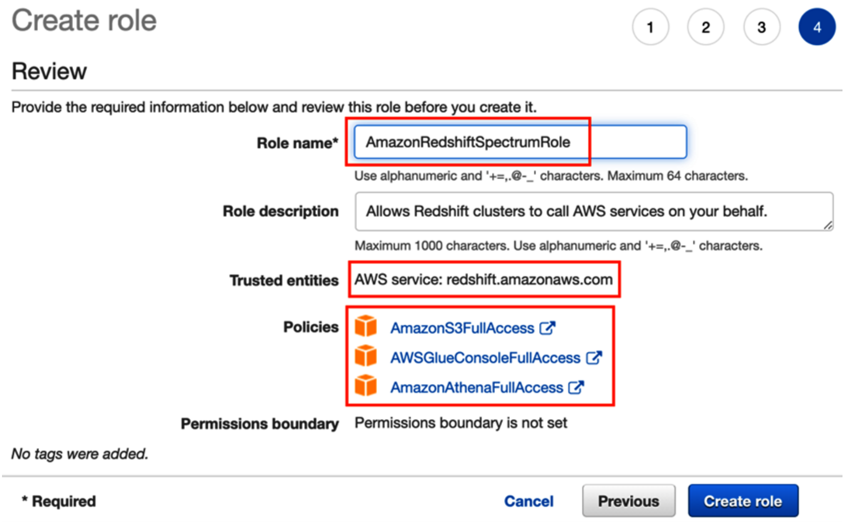
Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Status

Successfully returned 5 records in 358 ms

Bytes returned: 787 B

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	long
40669	Skyy's Lounge / Cozy	175412	Skyy		Ward C (councilmember Richard Boggiano)	40.73742	-74.
63282	2bed/2bath,furnished,doorman, by NY	304762	Gil		Ward B (councilmember Mira Prinz-Arey)	40.72813	-74.
146144	Shared Room	266070	Patricia		Ward E (councilmember James Solomon)	40.71077	-74.
215768	Minutes to Manhattan & Jersey Shore	846837	Charlaine		Ward F (councilmember Jermaine D. Robinson)	40.71663	-74.



Create role

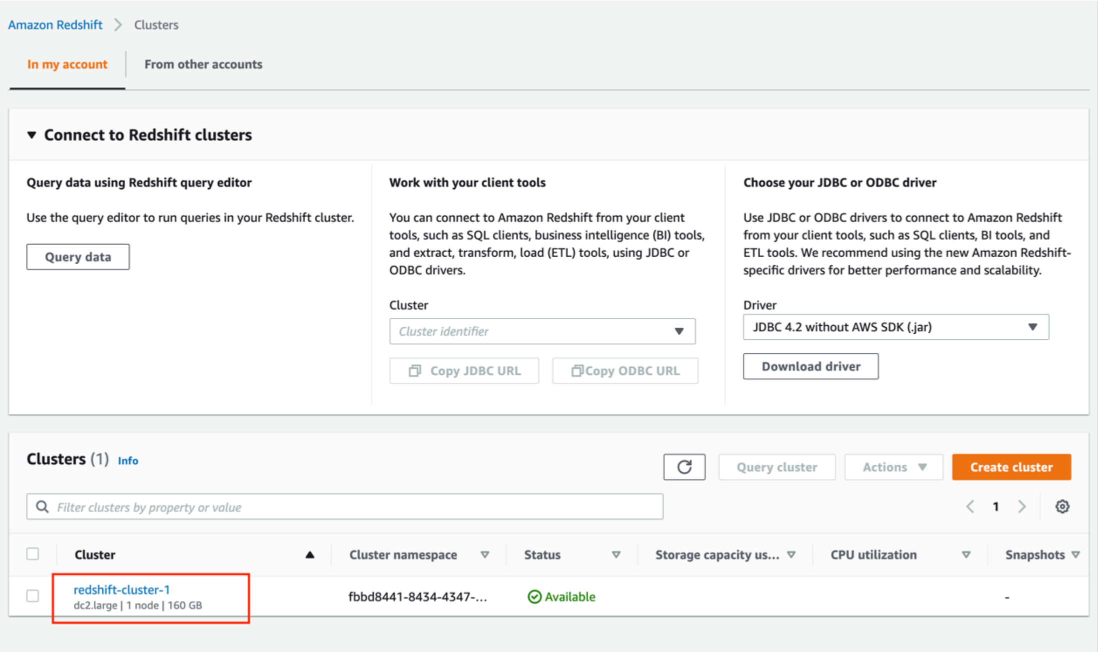
Review

Provide the required information below and review this role before you create it.

Role name*	AmazonRedshiftSpectrumRole Use alphanumeric and '+=,.@-_' characters. Maximum 64 characters.
Role description	Allows Redshift clusters to call AWS services on your behalf. Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.
Trusted entities	AWS service: redshift.amazonaws.com
Policies	AmazonS3FullAccess AWSGlueConsoleFullAccess AmazonAthenaFullAccess
Permissions boundary	Permissions boundary is not set

No tags were added.

* Required



Connect to Redshift clusters

Query data using Redshift query editor

Use the query editor to run queries in your Redshift cluster.

Work with your client tools

You can connect to Amazon Redshift from your client tools, such as SQL clients, business intelligence (BI) tools, and extract, transform, load (ETL) tools, using JDBC or ODBC drivers.

Cluster: Cluster identifier

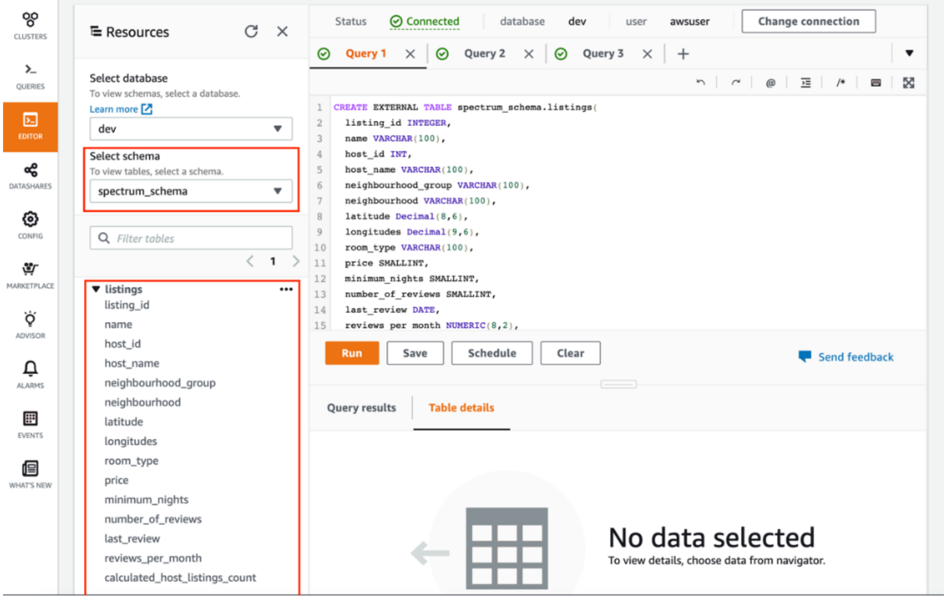
Choose your JDBC or ODBC driver

Use JDBC or ODBC drivers to connect to Amazon Redshift from your client tools, such as SQL clients, BI tools, and ETL tools. We recommend using the new Amazon Redshift-specific drivers for better performance and scalability.

Driver: JDBC 4.2 without AWS SDK (.jar)

Clusters (1) Info

□	Cluster	Cluster namespace	Status	Storage capacity us...	CPU utilization	Snapshots
□	redshift-cluster-1 dc2.large | 1 node | 160 GB	fbbd8441-8434-4347-...	✓ Available			-



Resources

Select database
To view schemas, select a database.
Learn more
dev

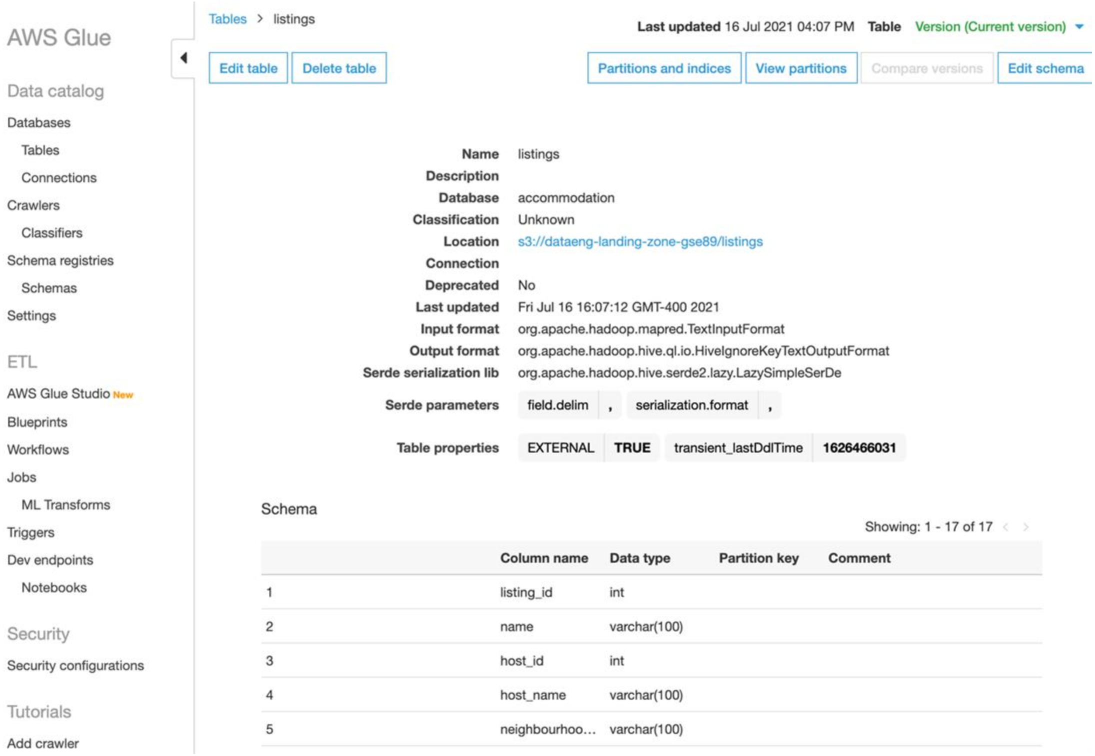
Select schema
To view tables, select a schema.
spectrum_schema

listings
- listing_id
- name
- host_id
- host_name
- neighbourhood_group
- neighbourhood
- latitude
- longitudes
- room_type
- price
- minimum_nights
- number_of_reviews
- last_review
- reviews_per_month
- calculated_host_listings_count

Status ✓ Connected database dev user awsuser

```
CREATE EXTERNAL TABLE spectrum_schema.listings(
  listing_id INTEGER,
  name VARCHAR(100),
  host_id INT,
  host_name VARCHAR(100),
  neighbourhood_group VARCHAR(100),
  neighbourhood VARCHAR(100),
  latitude Decimal(8,6),
  longitudes Decimal(9,6),
  room_type VARCHAR(100),
  price SMALLINT,
  minimum_nights SMALLINT,
  number_of_reviews SMALLINT,
  last_review DATE,
  reviews per month NUMERIC(8,2),
```

No data selected
To view details, choose data from navigator.



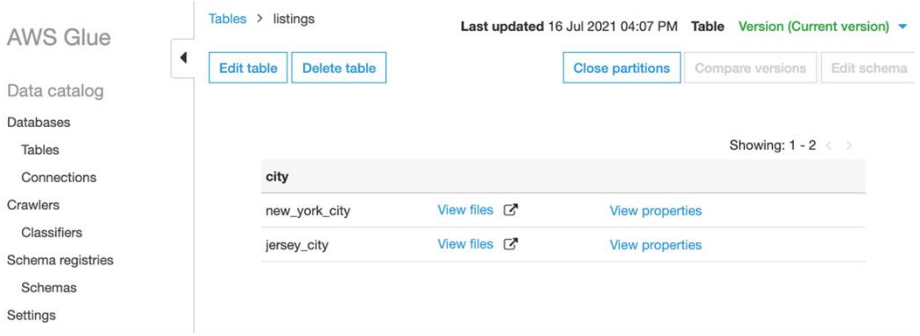
Last updated 16 Jul 2021 04:07 PM **Table** Version (Current version)

Name	listings
Description	
Database	accommodation
Classification	Unknown
Location	s3://dataeng-landing-zone-gse89/listings
Connection	
Deprecated	No
Last updated	Fri Jul 16 16:07:12 GMT-400 2021
Input format	org.apache.hadoop.mapred.TextInputFormat
Output format	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters	field.delim , serialization.format ,
Table properties	EXTERNAL **TRUE** transient_lastDdlTime **1626466031**

Schema

Showing: 1 - 17 of 17

	Column name	Data type	Partition key	Comment
1	listing_id	int		
2	name	varchar(100)		
3	host_id	int		
4	host_name	varchar(100)		
5	neighbourhoo…	varchar(100)		

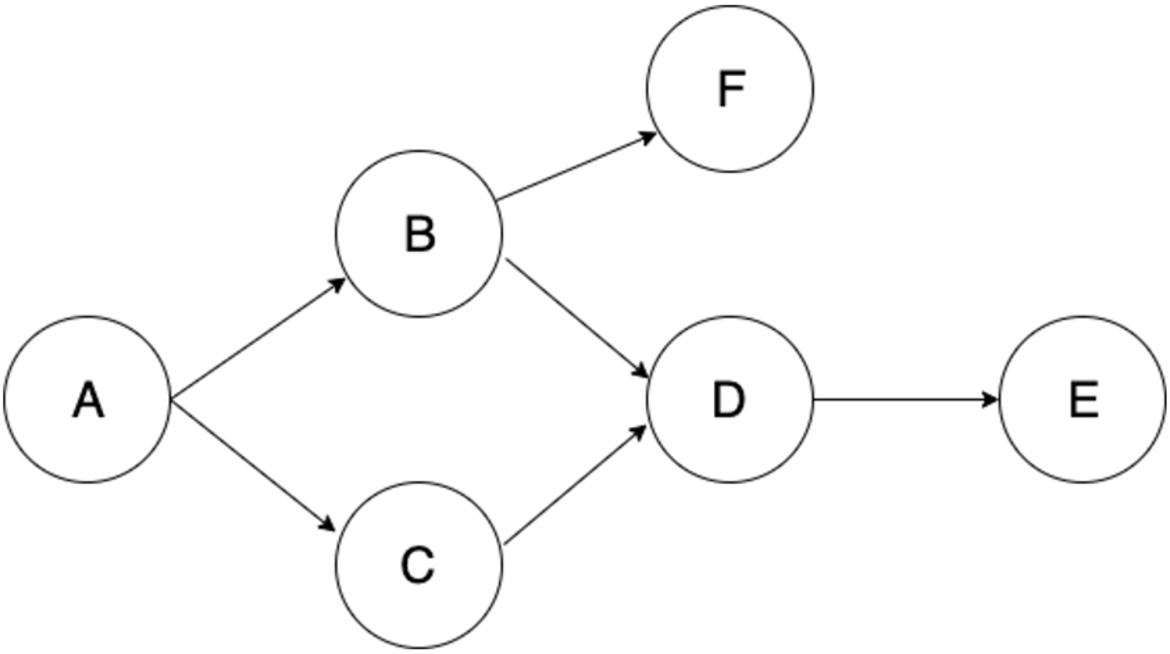


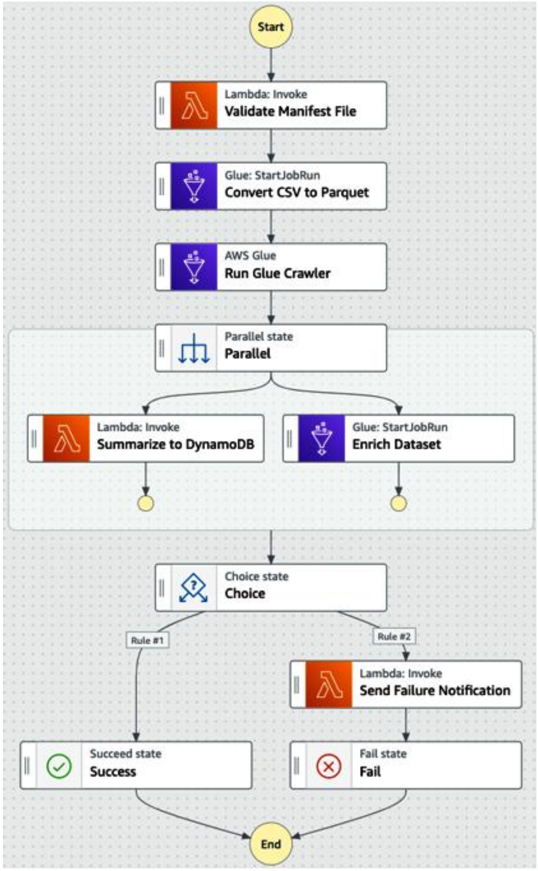
Last updated 16 Jul 2021 04:07 PM **Table** Version (Current version)

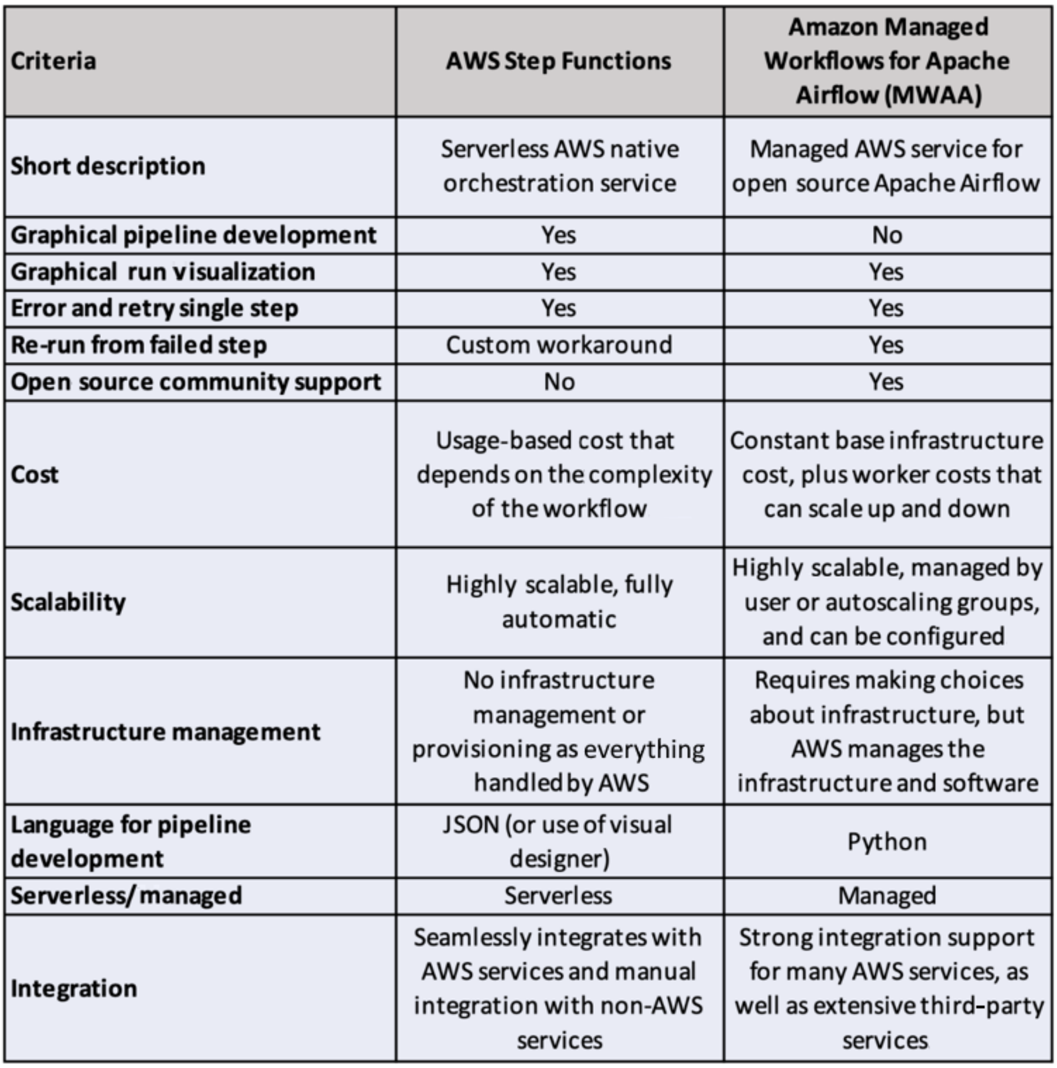
Showing: 1 - 2

city		
new_york_city	View files	View properties
jersey_city	View files	View properties

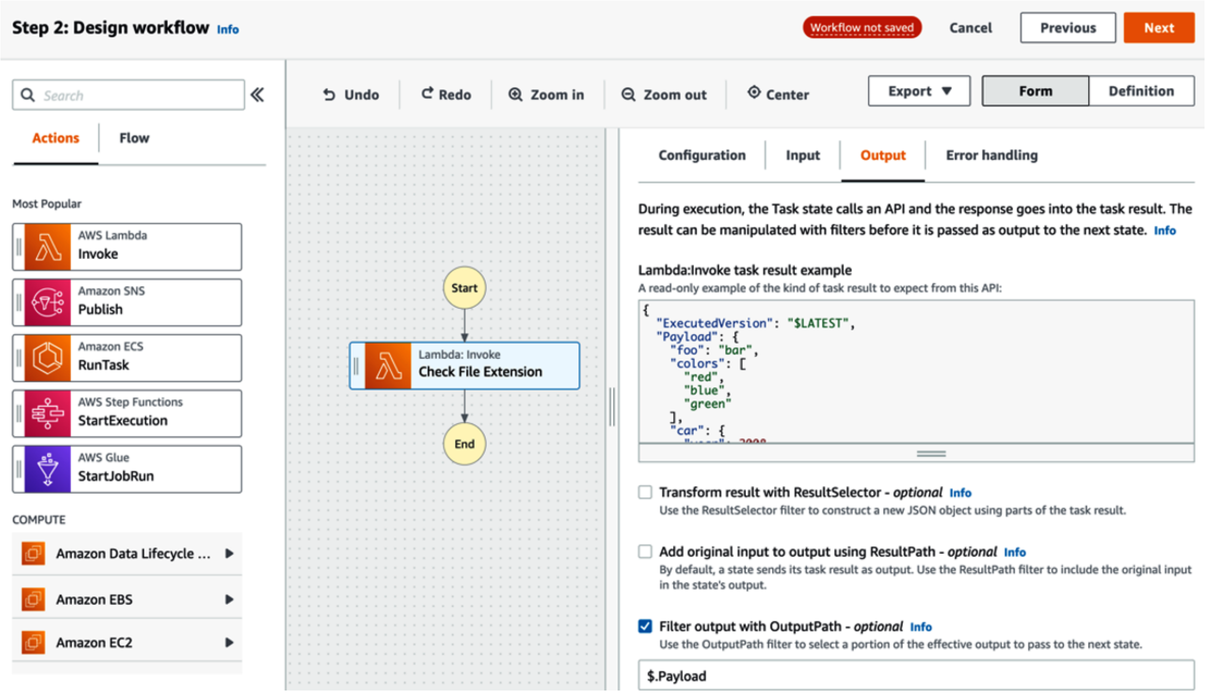
Chapter 10: Orchestrating the Data Pipeline







Criteria	AWS Step Functions	Amazon Managed Workflows for Apache Airflow (MWAA)
Short description	Serverless AWS native orchestration service	Managed AWS service for open source Apache Airflow
Graphical pipeline development	Yes	No
Graphical run visualization	Yes	Yes
Error and retry single step	Yes	Yes
Re-run from failed step	Custom workaround	Yes
Open source community support	No	Yes
Cost	Usage-based cost that depends on the complexity of the workflow	Constant base infrastructure cost, plus worker costs that can scale up and down
Scalability	Highly scalable, fully automatic	Highly scalable, managed by user or autoscaling groups, and can be configured
Infrastructure management	No infrastructure management or provisioning as everything handled by AWS	Requires making choices about infrastructure, but AWS manages the infrastructure and software
Language for pipeline development	JSON (or use of visual designer)	Python
Serverless/ managed	Serverless	Managed
Integration	Seamlessly integrates with AWS services and manual integration with non-AWS services	Strong integration support for many AWS services, as well as extensive third-party services



During execution, the Task state calls an API and the response goes into the task result. The result can be manipulated with filters before it is passed as output to the next state. Info

Lambda:Invoke task result example

A read-only example of the kind of task result to expect from this API:

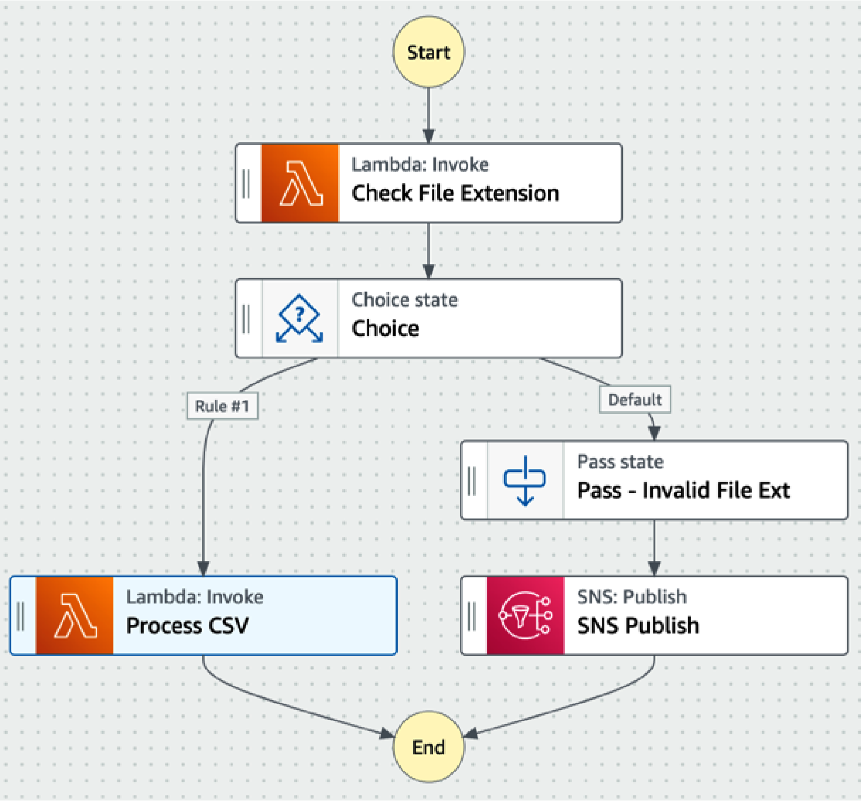
```
{
  "ExecutedVersion": "$LATEST",
  "Payload": {
    "foo": "bar",
    "colors": [
      "red",
      "blue",
      "green"
    ],
    "car": {
```

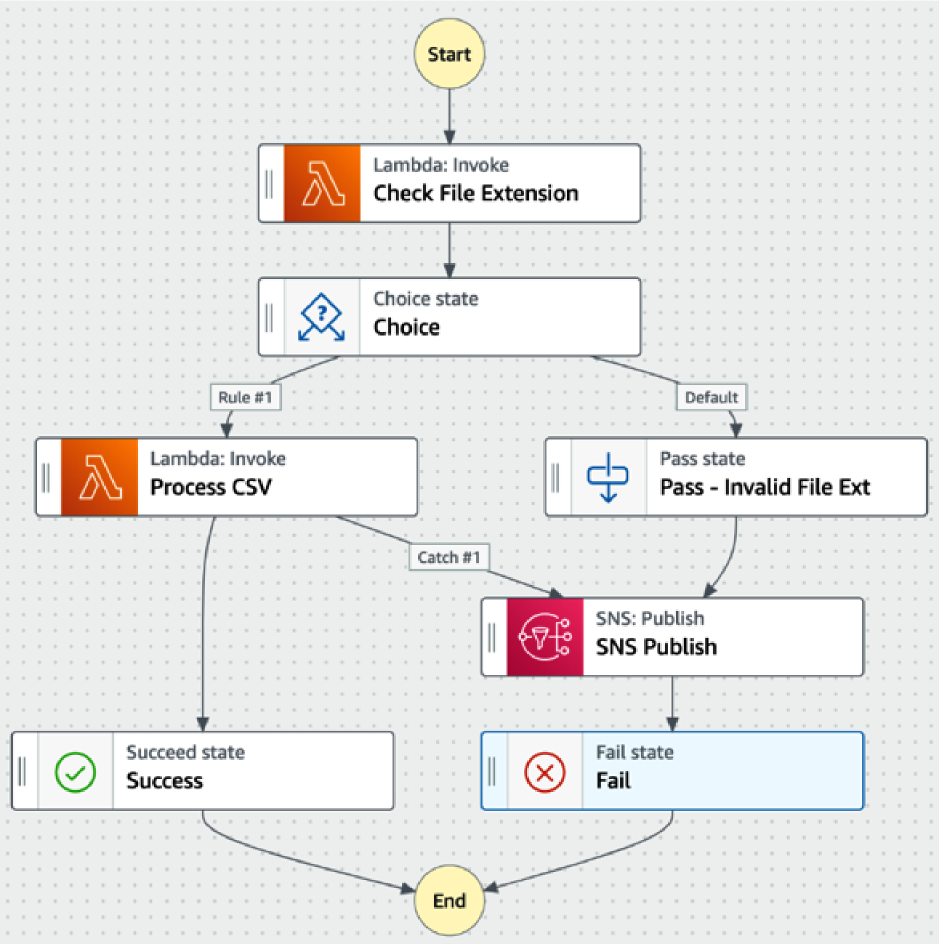
□ Transform result with ResultSelector - *optional* Info
Use the ResultSelector filter to construct a new JSON object using parts of the task result.

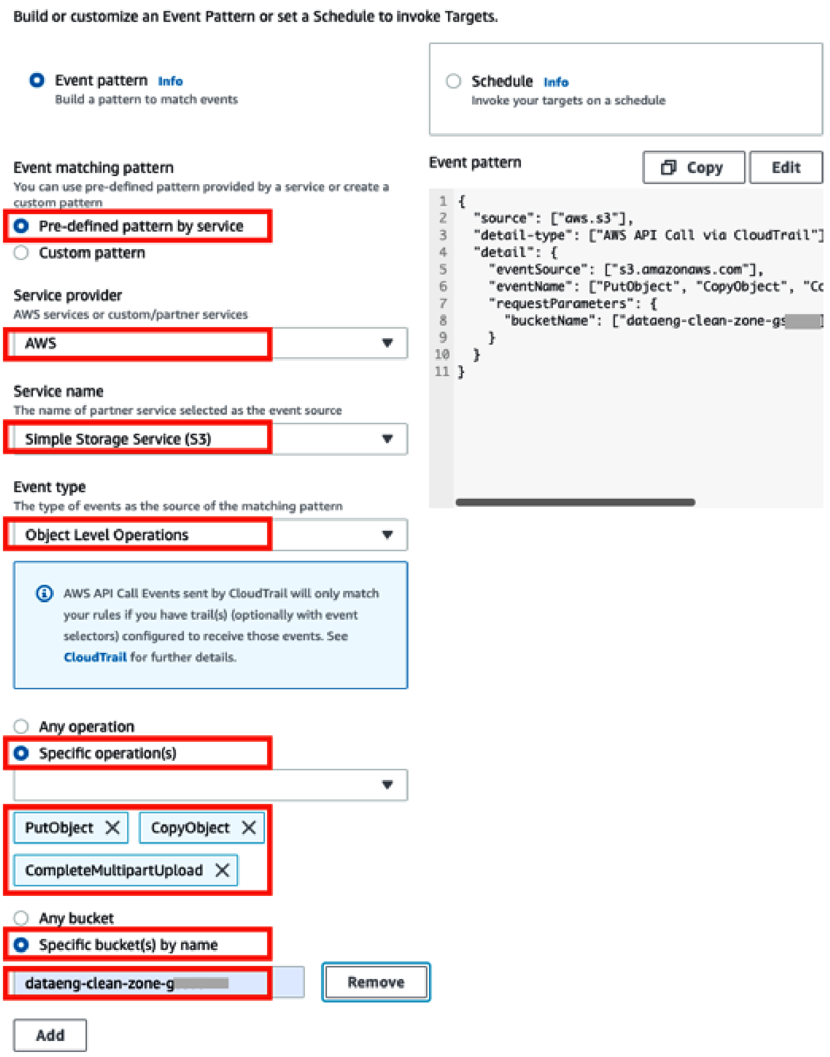
□ Add original input to output using ResultPath - *optional* Info
By default, a state sends its task result as output. Use the ResultPath filter to include the original input in the state's output.

☑ Filter output with OutputPath - *optional* Info
Use the OutputPath filter to select a portion of the effective output to pass to the next state.

$.Payload


Start
Lambda: Invoke
Check File Extension
Choice state
Choice
Rule #1
Default
Pass state
Pass - Invalid File Ext
Lambda: Invoke
Process CSV
SNS: Publish
SNS Publish
End


Start
Lambda: Invoke
Check File Extension
Choice state
Choice
Rule #1
Default
Lambda: Invoke
Process CSV
Pass state
Pass - Invalid File Ext
Catch #1
SNS: Publish
SNS Publish
Succeed state
Success
Fail state
Fail
End



Build or customize an Event Pattern or set a Schedule to invoke Targets.

● Event pattern Info
Build a pattern to match events

○ Schedule Info
Invoke your targets on a schedule

Event matching pattern
You can use pre-defined pattern provided by a service or create a custom pattern

● Pre-defined pattern by service
○ Custom pattern

Service provider
AWS services or custom/partner services
AWS

Service name
The name of partner service selected as the event source
Simple Storage Service (S3)

Event type
The type of events as the source of the matching pattern
Object Level Operations

AWS API Call Events sent by CloudTrail will only match your rules if you have trail(s) (optionally with event selectors) configured to receive those events. See CloudTrail for further details.

○ Any operation
● Specific operation(s)

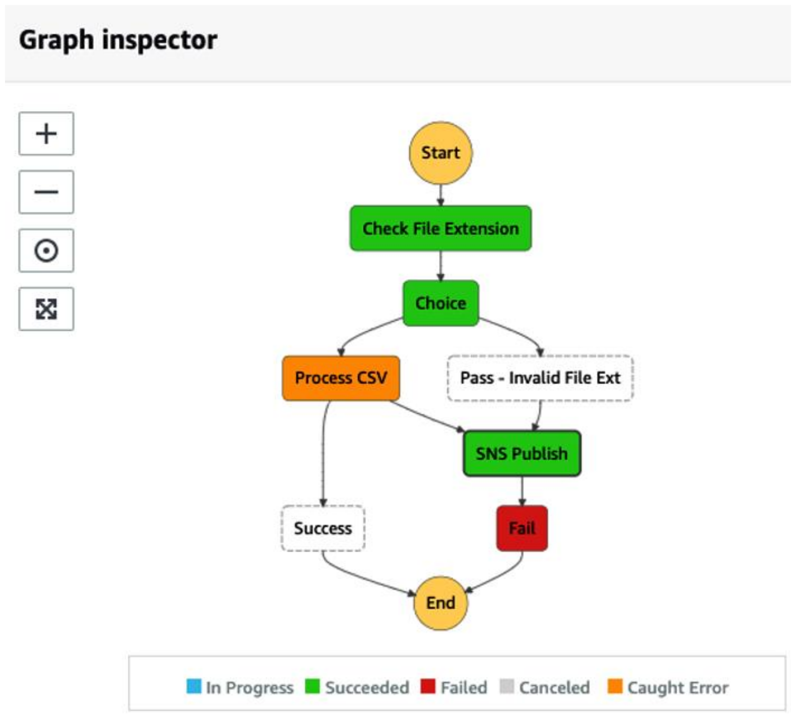
PutObject ✕ CopyObject ✕
CompleteMultipartUpload ✕

○ Any bucket
● Specific bucket(s) by name

dataeng-clean-zone-g[redacted]

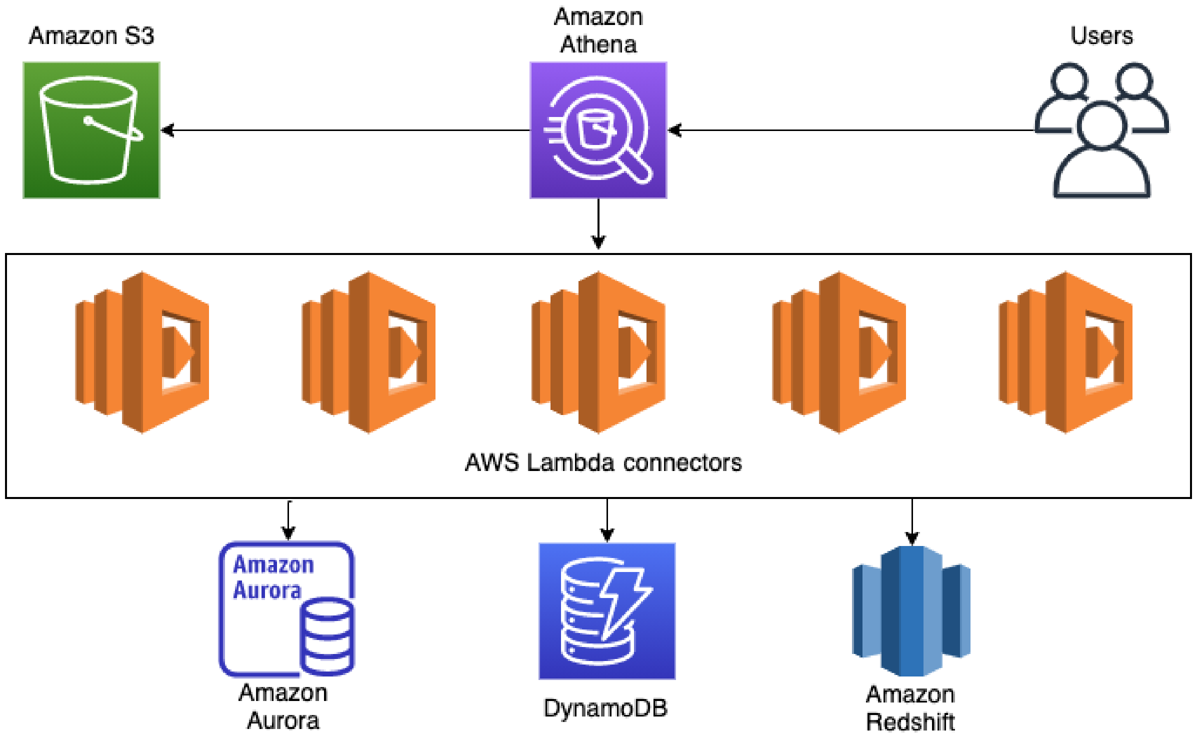
Event pattern

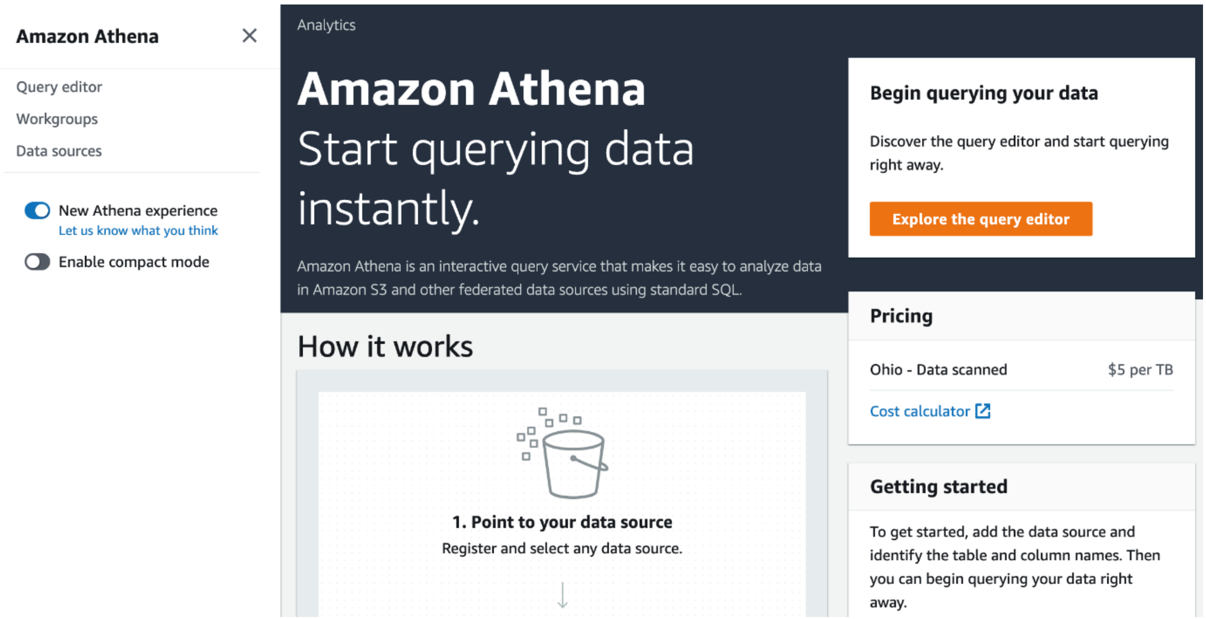
```
{
  "source": ["aws.s3"],
  "detail-type": ["AWS API Call via CloudTrail"]
  "detail": {
    "eventSource": ["s3.amazonaws.com"],
    "eventName": ["PutObject", "CopyObject", "Co
    "requestParameters": {
      "bucketName": ["dataeng-clean-zone-gs[redacted]]
    }
  }
}
```



Graph inspector

Chapter 11: Ad Hoc Queries with Amazon Athena





Amazon Athena

Start querying data instantly.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 and other federated data sources using standard SQL.

Begin querying your data

Discover the query editor and start querying right away.

Pricing

Ohio - Data scanned	$5 per TB

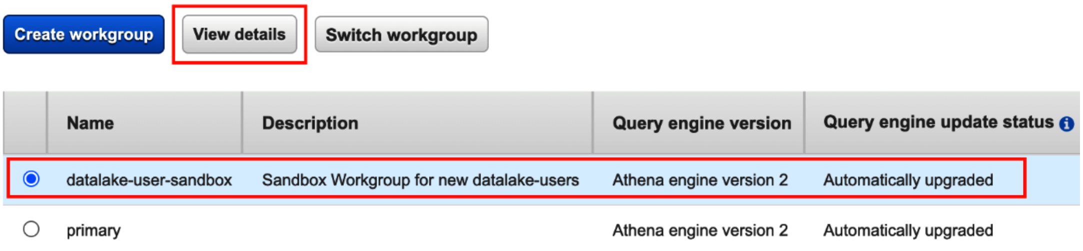
Getting started

To get started, add the data source and identify the table and column names. Then you can begin querying your data right away.

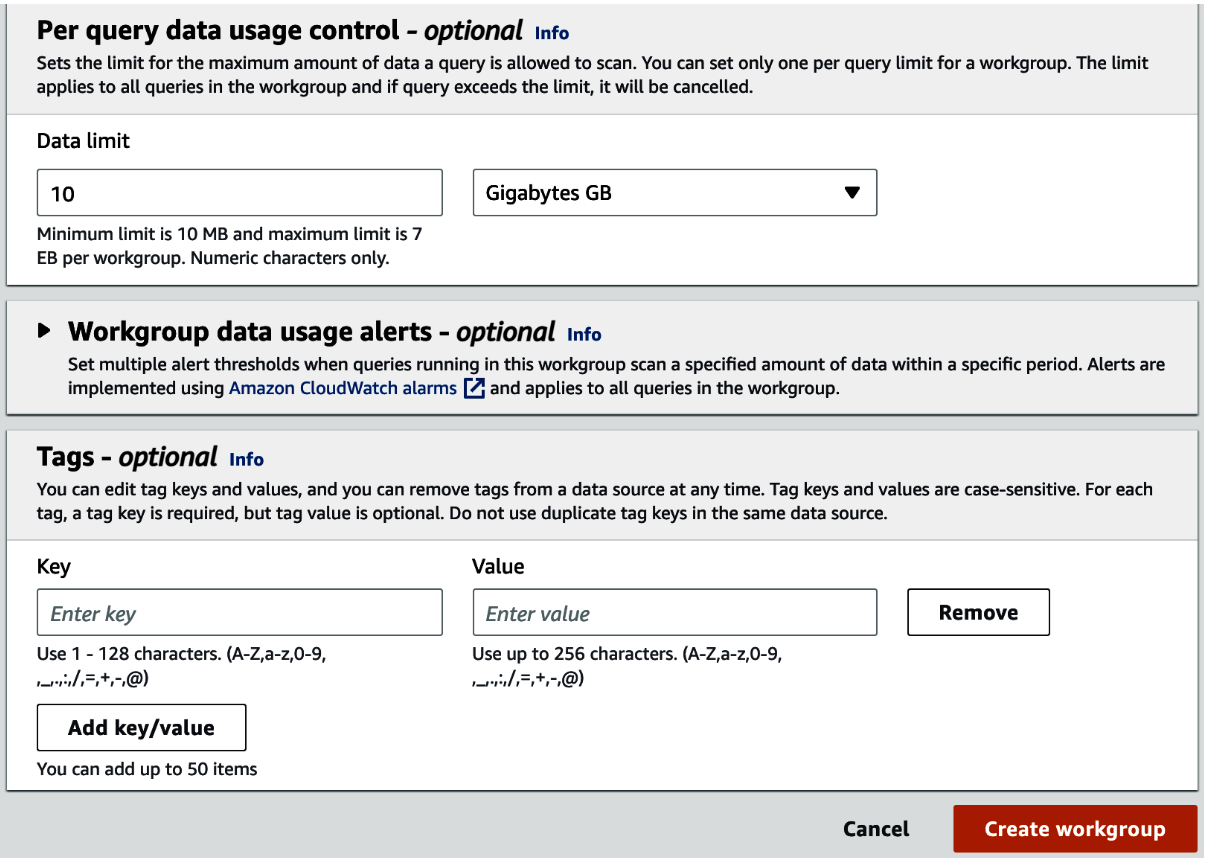
How it works

1. Point to your data source

Register and select any data source.



	Name	Description	Query engine version	Query engine update status
●	datalake-user-sandbox	Sandbox Workgroup for new datalake-users	Athena engine version 2	Automatically upgraded
○	primary		Athena engine version 2	Automatically upgraded



Per query data usage control - *optional* Info

Sets the limit for the maximum amount of data a query is allowed to scan. You can set only one per query limit for a workgroup. The limit applies to all queries in the workgroup and if query exceeds the limit, it will be cancelled.

Data limit

10

Gigabytes GB ▼

Minimum limit is 10 MB and maximum limit is 7 EB per workgroup. Numeric characters only.

▶ **Workgroup data usage alerts - *optional*** Info

Set multiple alert thresholds when queries running in this workgroup scan a specified amount of data within a specific period. Alerts are implemented using Amazon CloudWatch alarms and applies to all queries in the workgroup.

Tags - *optional* Info

You can edit tag keys and values, and you can remove tags from a data source at any time. Tag keys and values are case-sensitive. For each tag, a tag key is required, but tag value is optional. Do not use duplicate tag keys in the same data source.

Key

Enter key

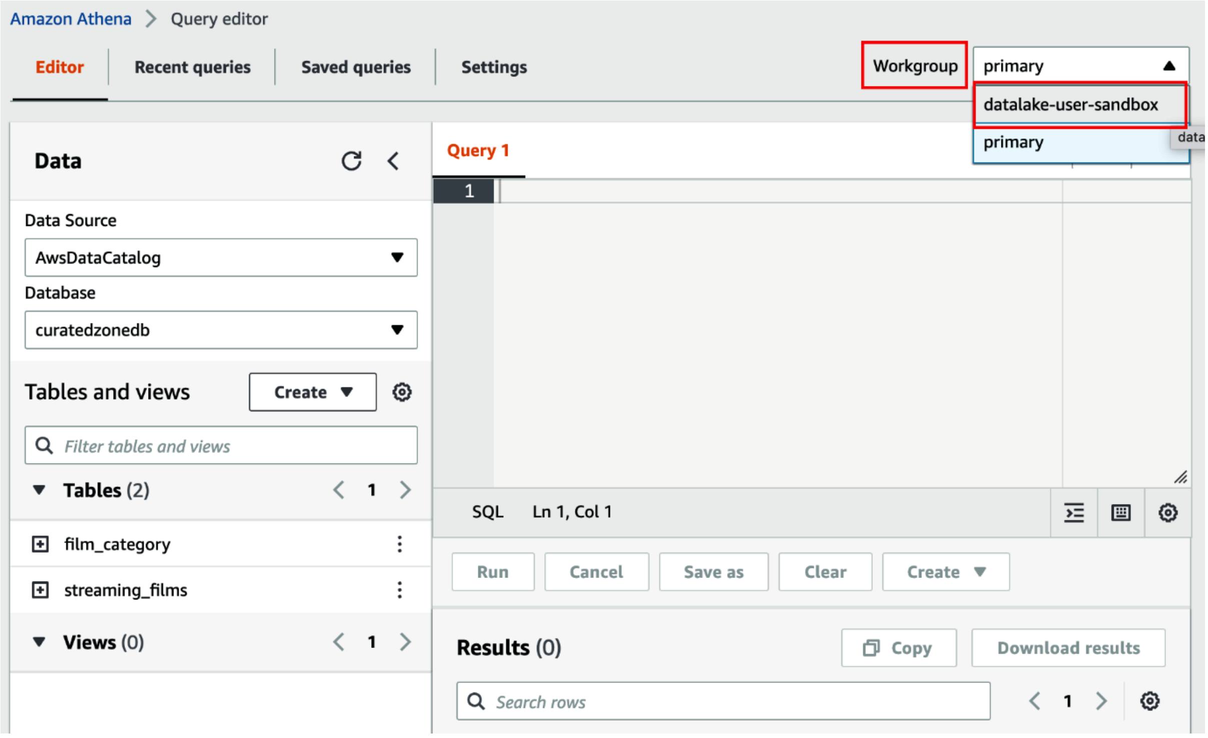
Use 1 - 128 characters. (A-Z,a-z,0-9, ,_.:/,=,+,-,@)

Value

Enter value

Use up to 256 characters. (A-Z,a-z,0-9, ,_.:/,=,+,-,@)

You can add up to 50 items

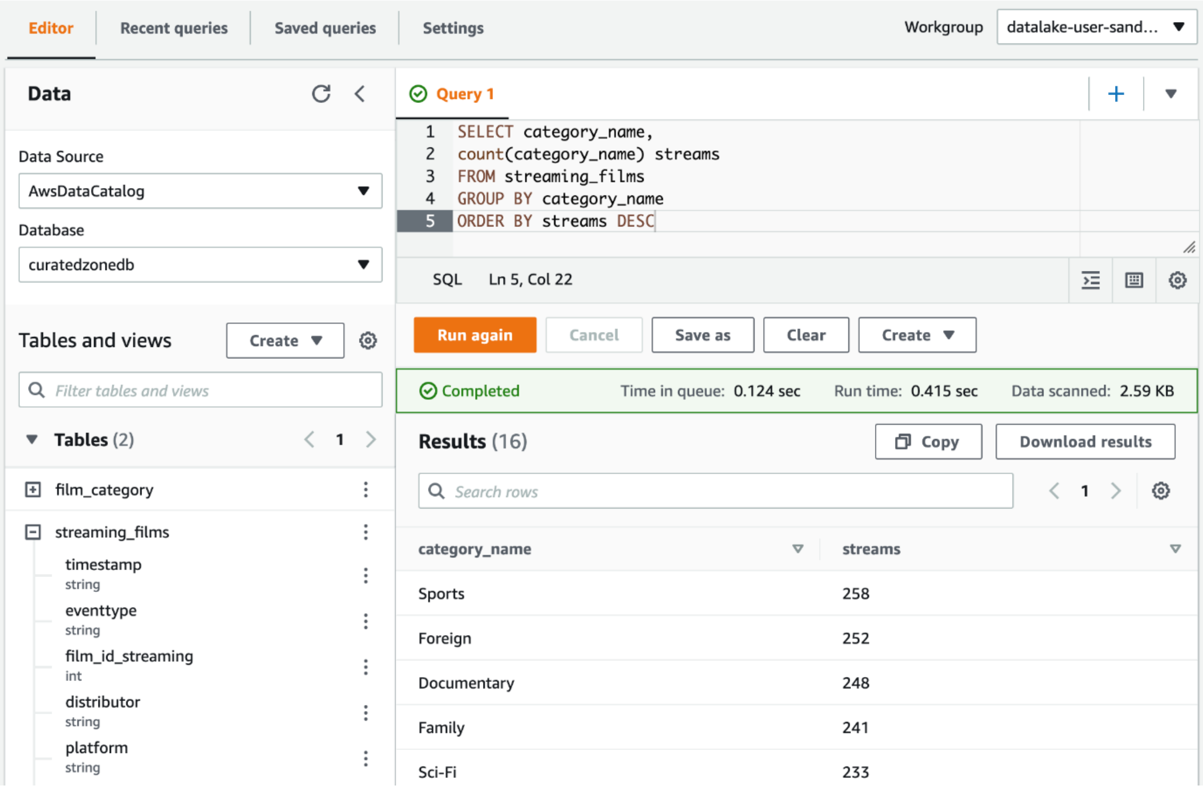


Workgroup: primary ▲

Data

Data Source: AwsDataCatalog ▼

Database: curatedzonedb ▼



Data Source: AwsDataCatalog

Database: curatedzonedb

streaming_films:

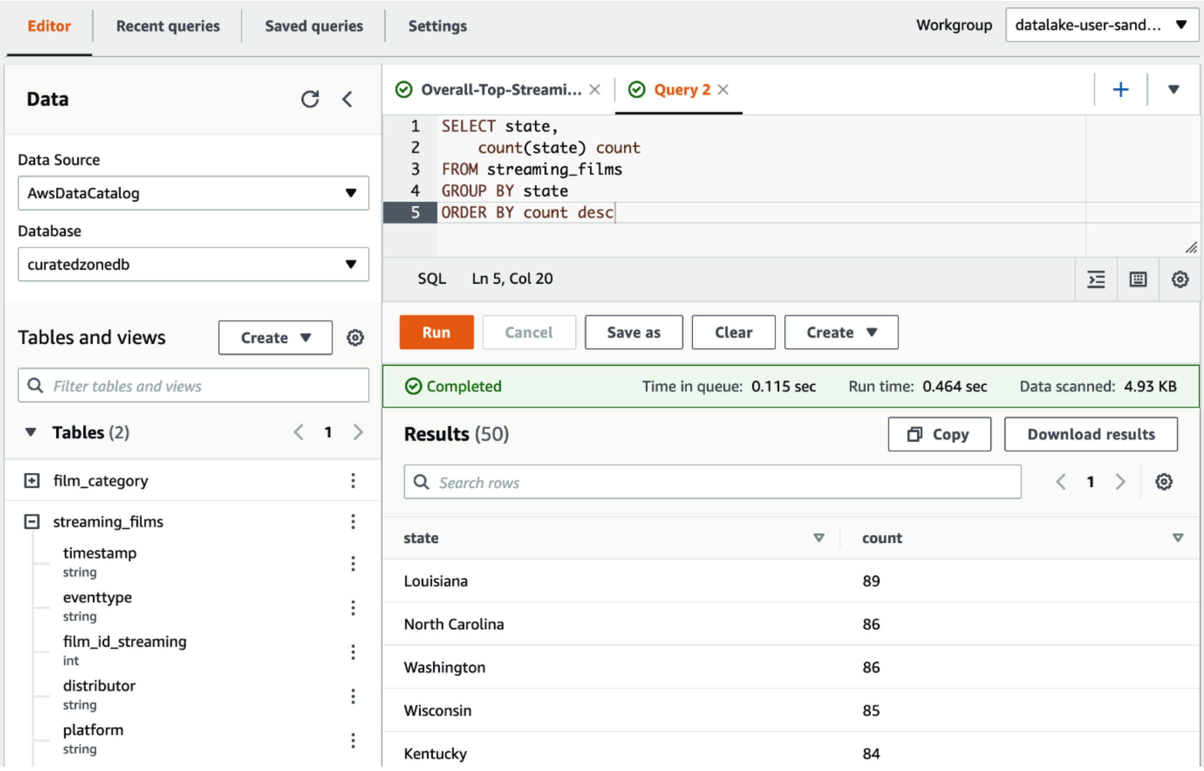
timestamp	string
eventtype	string
film_id_streaming	int
distributor	string
platform	string

```
SELECT category_name,
count(category_name) streams
FROM streaming_films
GROUP BY category_name
ORDER BY streams DESC
```

Completed   Time in queue: 0.124 sec   Run time: 0.415 sec   Data scanned: 2.59 KB

Results (16)

category_name	streams
Sports	258
Foreign	252
Documentary	248
Family	241
Sci-Fi	233



Data Source: AwsDataCatalog

Database: curatedzonedb

streaming_films:

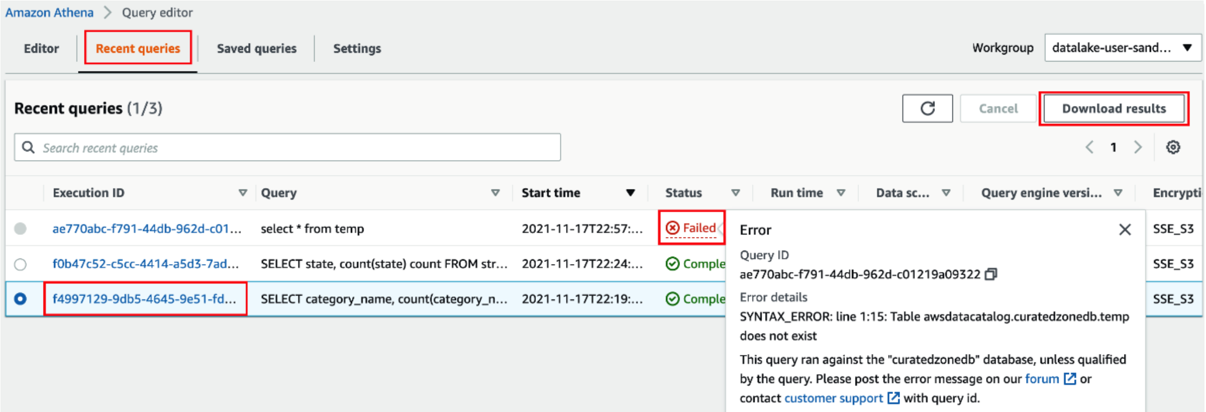
timestamp	string
eventtype	string
film_id_streaming	int
distributor	string
platform	string

```
SELECT state,
    count(state) count
FROM streaming_films
GROUP BY state
ORDER BY count desc
```

Completed   Time in queue: 0.115 sec   Run time: 0.464 sec   Data scanned: 4.93 KB

Results (50)

state	count
Louisiana	89
North Carolina	86
Washington	86
Wisconsin	85
Kentucky	84



Recent queries (1/3)

	Execution ID	Query	Start time	Status	Run time	Data sc...	Query engine versi...	Encrypti
○	ae770abc-f791-44db-962d-c01...	select * from temp	2021-11-17T22:57:...	Failed				SSE_S3
○	f0b47c52-c5cc-4414-a5d3-7ad...	SELECT state, count(state) count FROM str...	2021-11-17T22:24:...	Comple				SSE_S3
●	f4997129-9db5-4645-9e51-fd...	SELECT category_name, count(category_n...	2021-11-17T22:19:...	Comple				SSE_S3

Error

Query ID

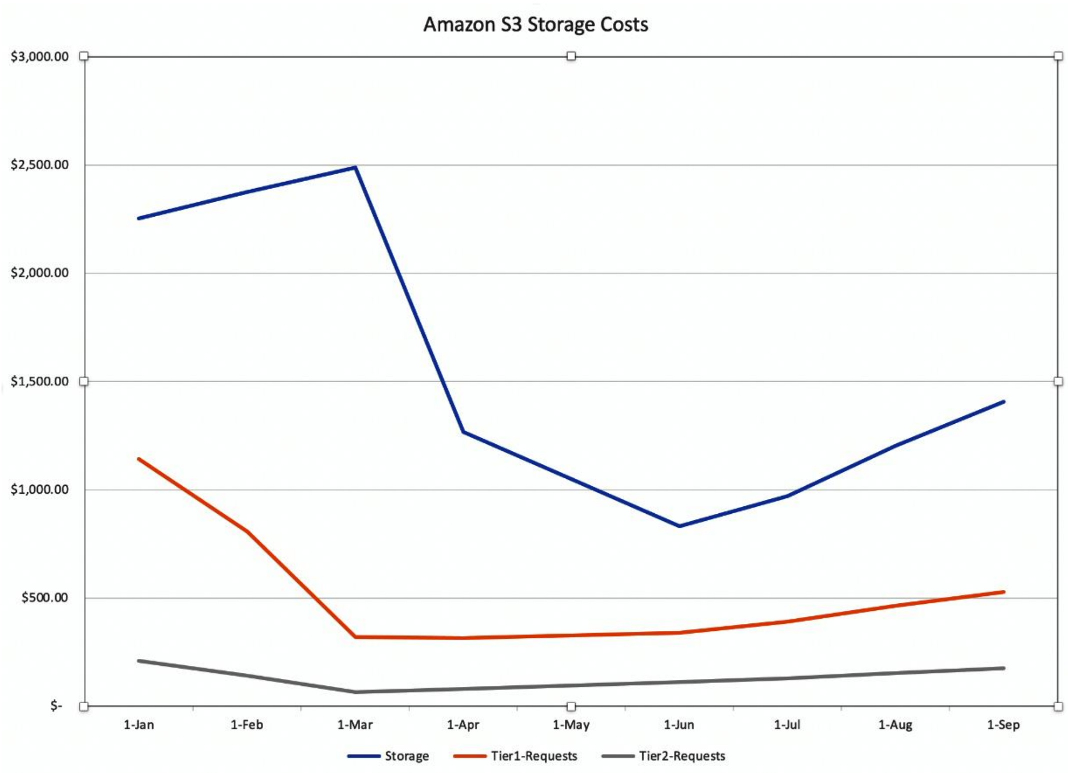
ae770abc-f791-44db-962d-c01219a09322

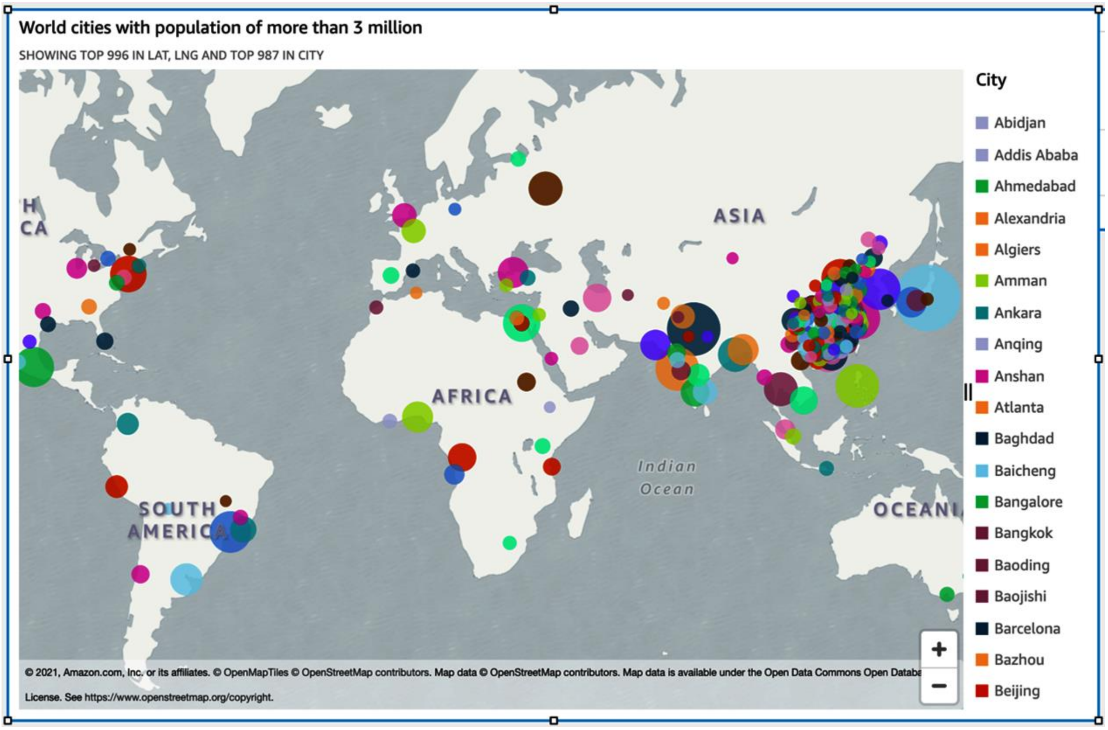
Error details

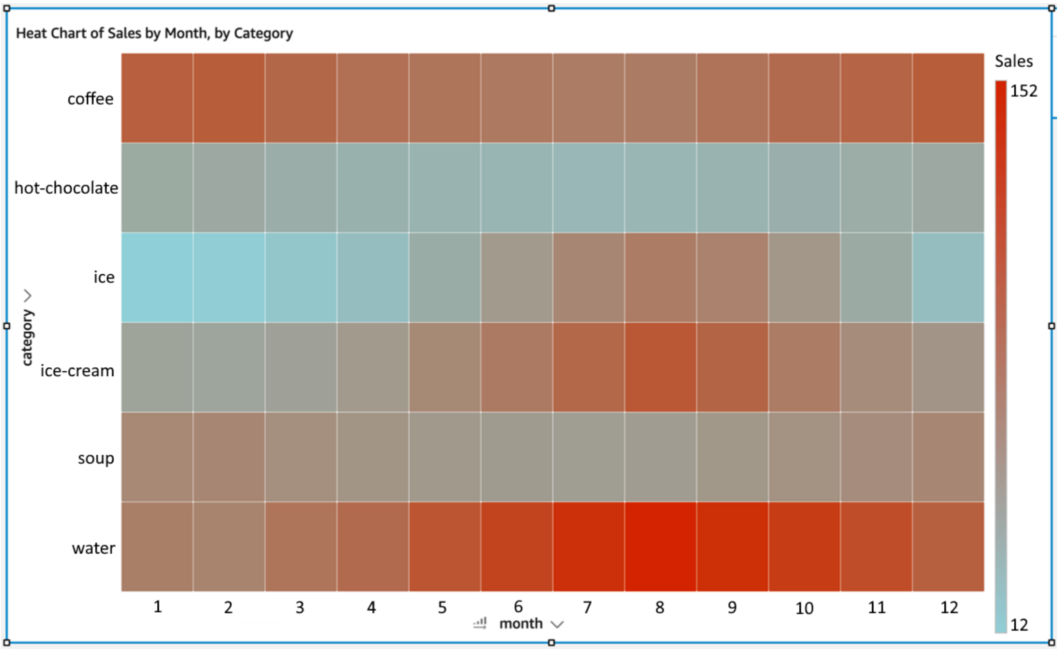
SYNTAX_ERROR: line 1:15: Table awsdatacatalog.curatedzonedb.temp does not exist

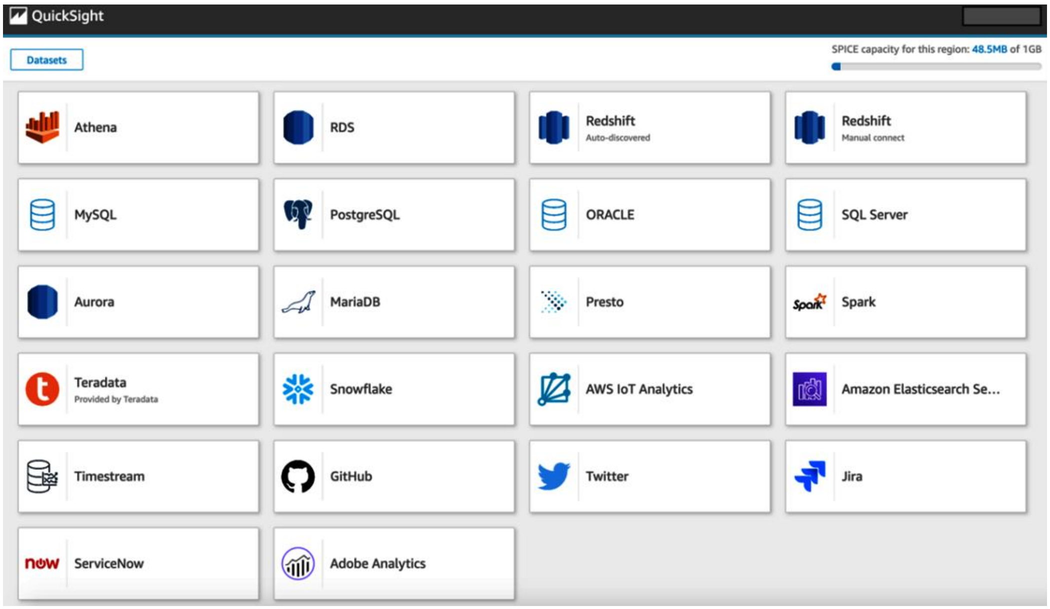
This query ran against the "curatedzonedb" database, unless qualified by the query. Please post the error message on our forum or contact customer support with query id.

Chapter 12: Visualizing Data with Amazon QuickSight

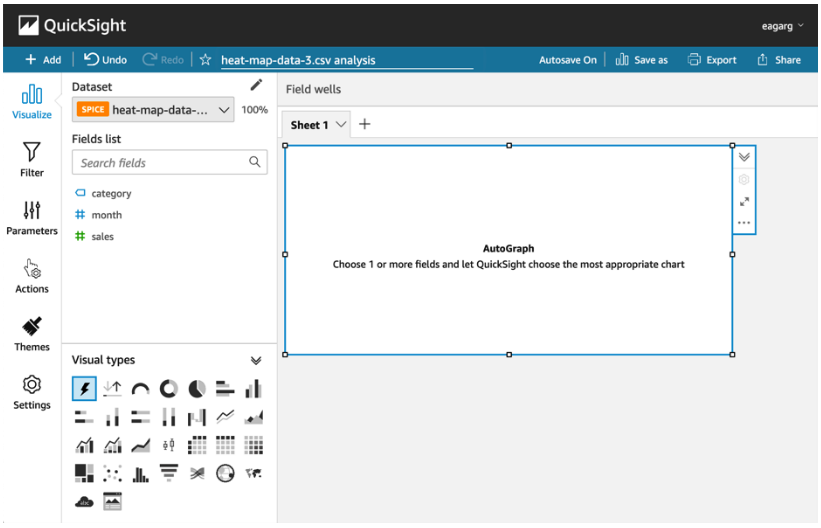



World cities with population of more than 3 million
SHOWING TOP 996 IN LAT, LNG AND TOP 987 IN CITY
ASIA
AFRICA
SOUTH AMERICA
Indian Ocean
OCEANI
City
Abidjan
Addis Ababa
Ahmedabad
Alexandria
Algiers
Amman
Ankara
Anqing
Anshan
Atlanta
Baghdad
Baicheng
Bangalore
Bangkok
Baoding
Baojishi
Barcelona
Bazhou
Beijing
+
−
© 2021, Amazon.com, Inc. or its affiliates. © OpenMapTiles © OpenStreetMap contributors. Map data © OpenStreetMap contributors. Map data is available under the Open Data Commons Open Databa
License. See https://www.openstreetmap.org/copyright.





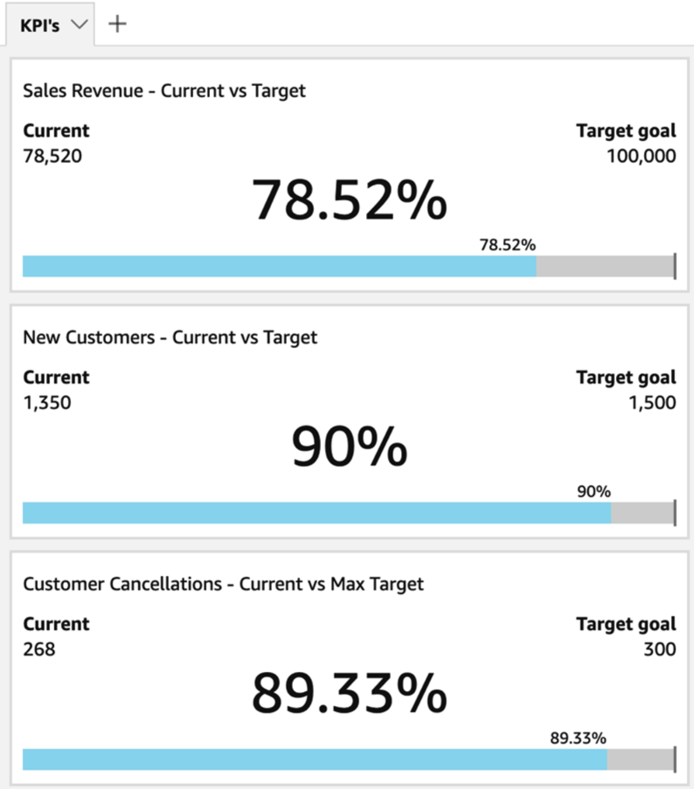
SPICE capacity for this region: 48.5MB of 1GB



- category
- month
- sales

AutoGraph

Choose 1 or more fields and let QuickSight choose the most appropriate chart



Sales Revenue - Current vs Target

Current 78,520 | **Target goal** 100,000

78.52%

New Customers - Current vs Target

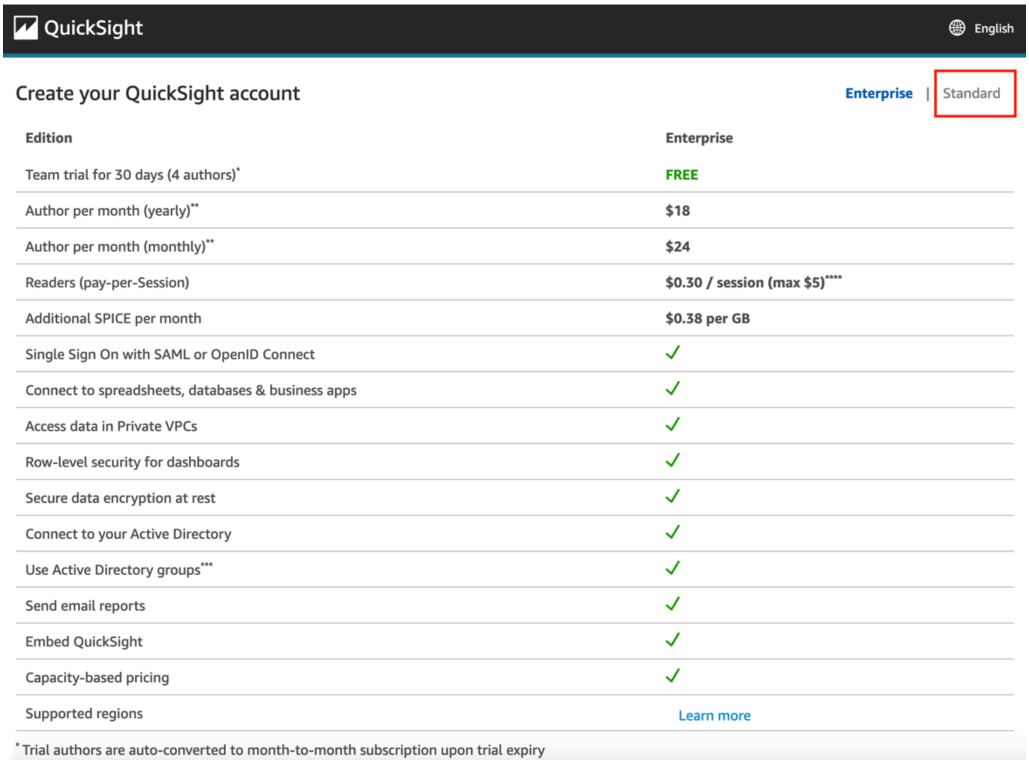
Current 1,350 | **Target goal** 1,500

90%

Customer Cancellations - Current vs Max Target

Current 268 | **Target goal** 300

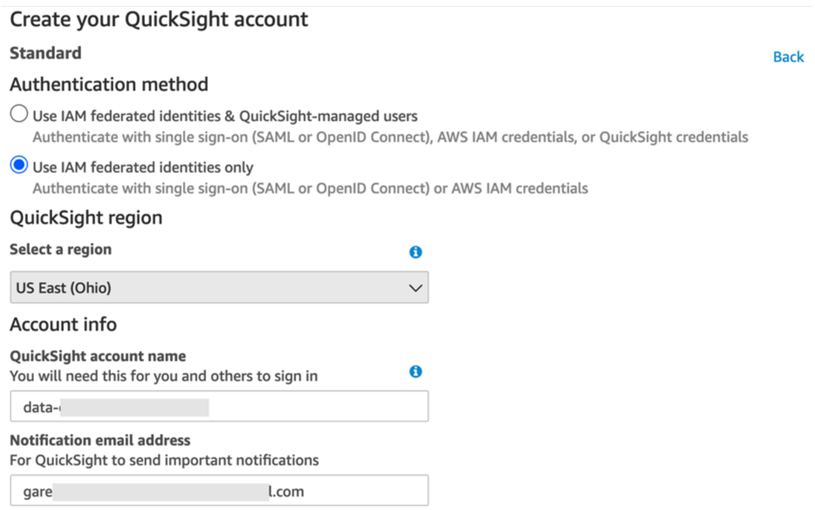
89.33%



Create your QuickSight account

Edition	Enterprise
Team trial for 30 days (4 authors)*	FREE
Author per month (yearly)**	$18
Author per month (monthly)**	$24
Readers (pay-per-Session)	$0.30 / session (max $5)****
Additional SPICE per month	$0.38 per GB
Single Sign On with SAML or OpenID Connect	✓
Connect to spreadsheets, databases & business apps	✓
Access data in Private VPCs	✓
Row-level security for dashboards	✓
Secure data encryption at rest	✓
Connect to your Active Directory	✓
Use Active Directory groups***	✓
Send email reports	✓
Embed QuickSight	✓
Capacity-based pricing	✓
Supported regions	Learn more

* Trial authors are auto-converted to month-to-month subscription upon trial expiry



Create your QuickSight account

Standard

Back

Authentication method

○ Use IAM federated identities & QuickSight-managed users
Authenticate with single sign-on (SAML or OpenID Connect), AWS IAM credentials, or QuickSight credentials

● Use IAM federated identities only
Authenticate with single sign-on (SAML or OpenID Connect) or AWS IAM credentials

QuickSight region

Select a region: US East (Ohio)

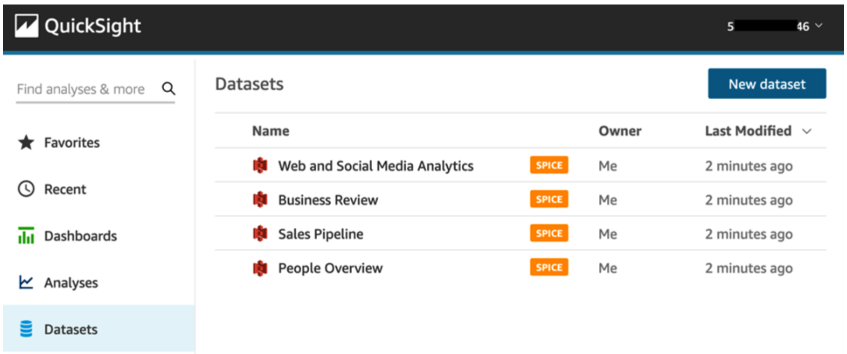
Account info

QuickSight account name
You will need this for you and others to sign in

data-[redacted]

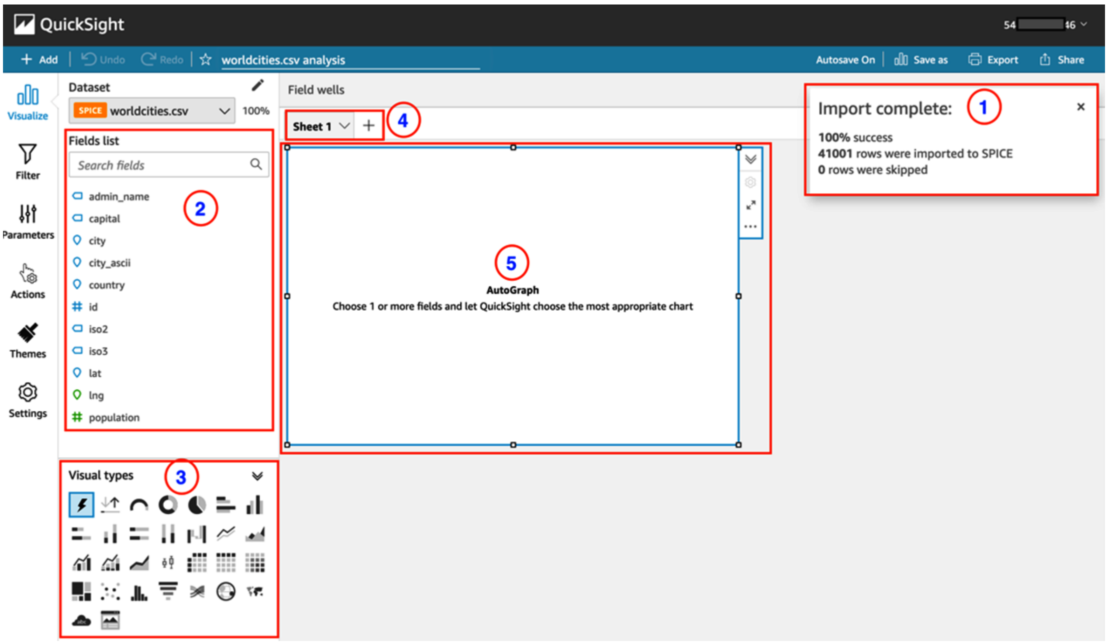
Notification email address
For QuickSight to send important notifications

gare[redacted]l.com



Datasets

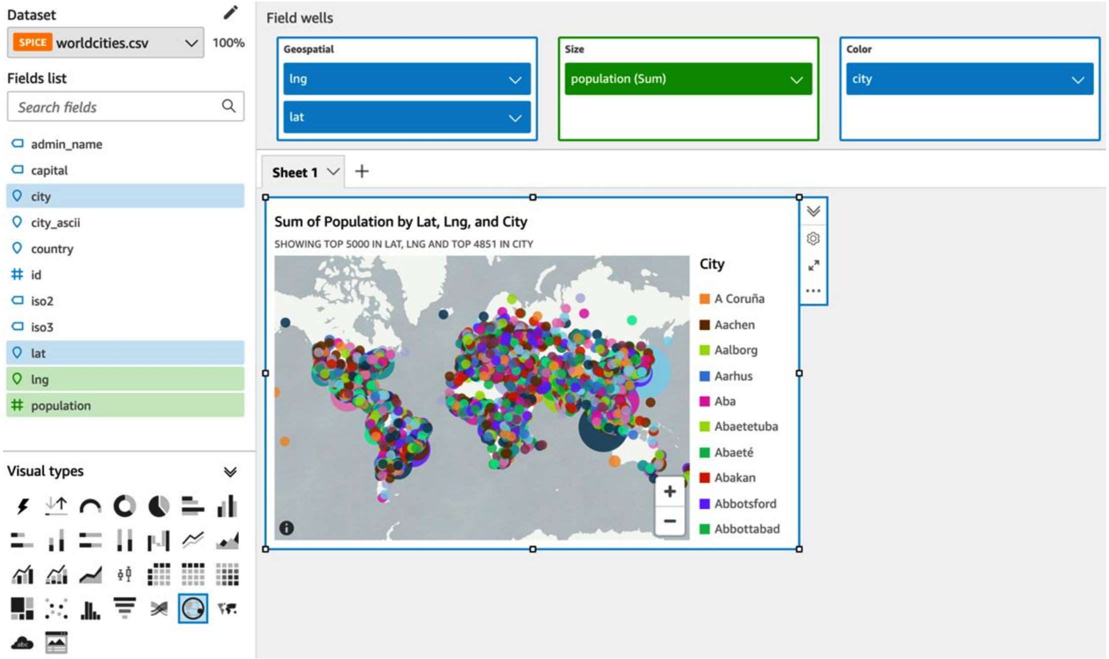
Name		Owner	Last Modified
Web and Social Media Analytics	SPICE	Me	2 minutes ago
Business Review	SPICE	Me	2 minutes ago
Sales Pipeline	SPICE	Me	2 minutes ago
People Overview	SPICE	Me	2 minutes ago



Import complete:
100% success
41001 rows were imported to SPICE
0 rows were skipped

Fields list
Search fields
- admin_name
- capital
- city
- city_ascii
- country
- id
- iso2
- iso3
- lat
- lng
- population

AutoGraph
Choose 1 or more fields and let QuickSight choose the most appropriate chart



Dataset: SPICE worldcities.csv 100%

Fields list

- admin_name
- capital
- city
- city_ascii
- country
- id
- iso2
- iso3
- lat
- lng
- population

Field wells

Geospatial: lng, lat

Size: population (Sum)

Color: city

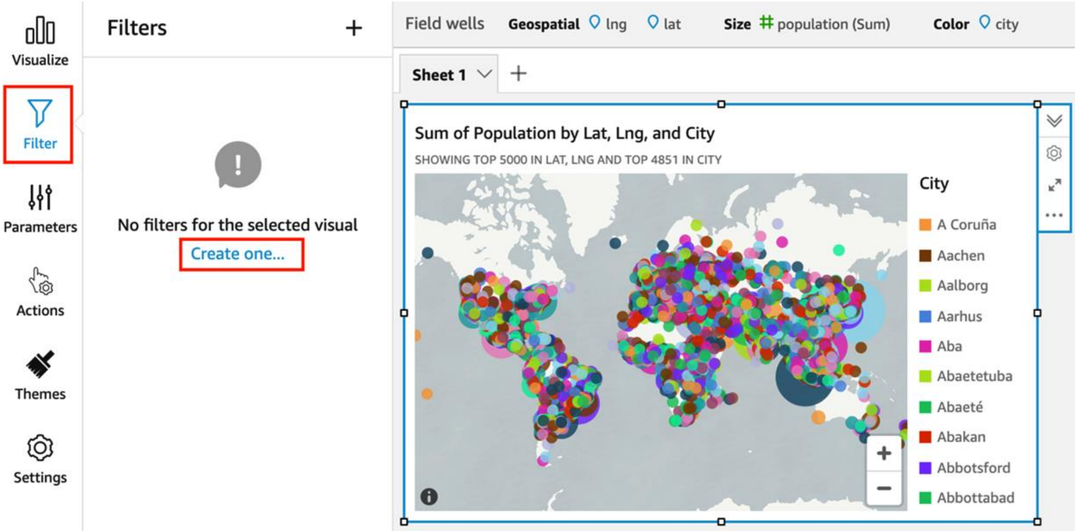
Sheet 1

Sum of Population by Lat, Lng, and City

SHOWING TOP 5000 IN LAT, LNG AND TOP 4851 IN CITY

City

- 🟧 A Coruña
- 🟫 Aachen
- 🟩 Aalborg
- 🟦 Aarhus
- 🟪 Aba
- 🟩 Abaetetuba
- 🟩 Abaeté
- 🟥 Abakan
- 🟪 Abbotsford
- 🟩 Abbottabad



Filters

No filters for the selected visual

Create one...

Field wells Geospatial lng lat Size population (Sum) Color city

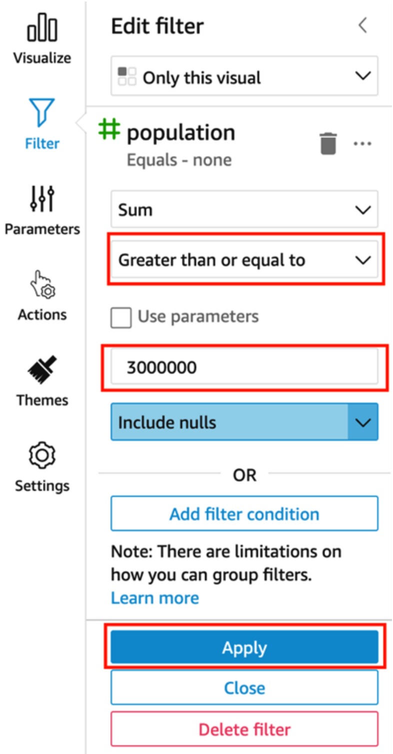
Sheet 1

Sum of Population by Lat, Lng, and City

SHOWING TOP 5000 IN LAT, LNG AND TOP 4851 IN CITY

City

- 🟧 A Coruña
- 🟫 Aachen
- 🟩 Aalborg
- 🟦 Aarhus
- 🟪 Aba
- 🟩 Abaetetuba
- 🟩 Abaeté
- 🟥 Abakan
- 🟪 Abbotsford
- 🟩 Abbottabad



Edit filter

Only this visual

population
Equals - none

Sum

Greater than or equal to

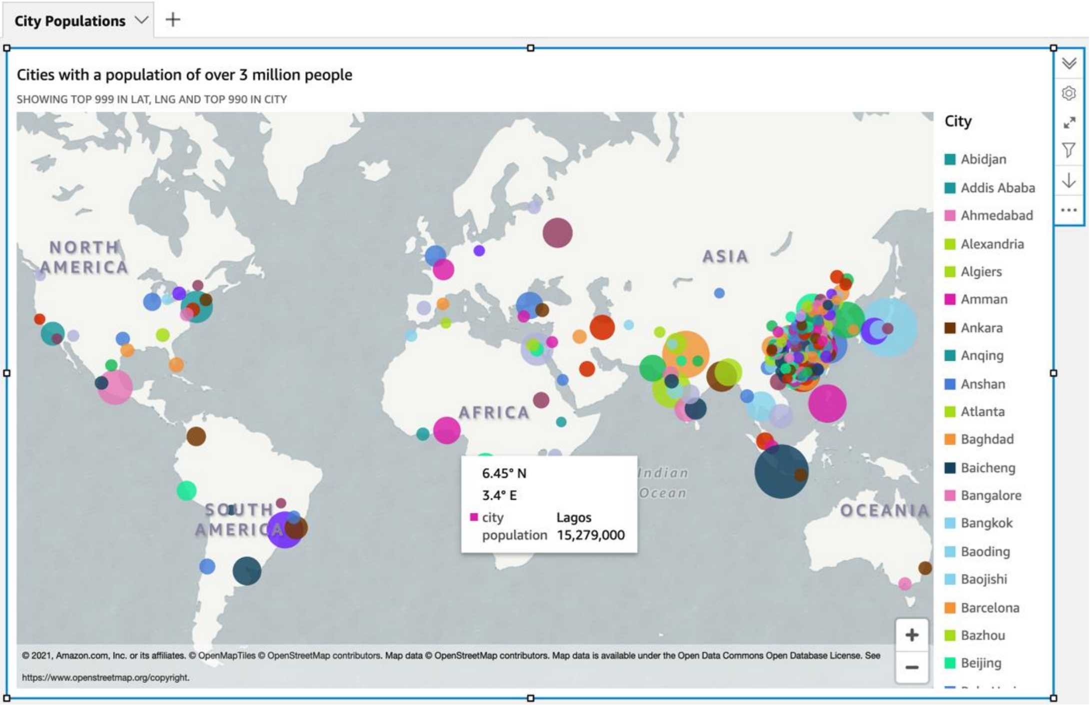
☐ Use parameters

3000000

Include nulls

OR

Note: There are limitations on how you can group filters.
Learn more



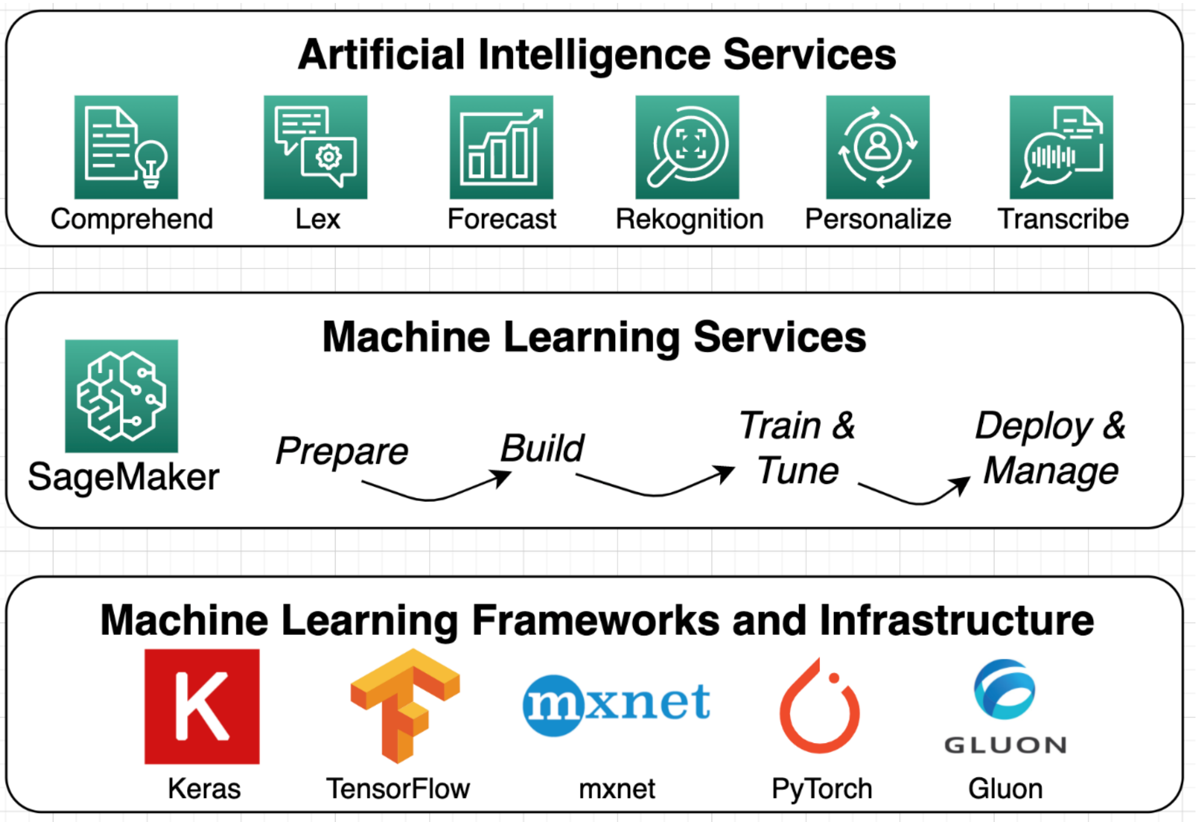
Cities with a population of over 3 million people

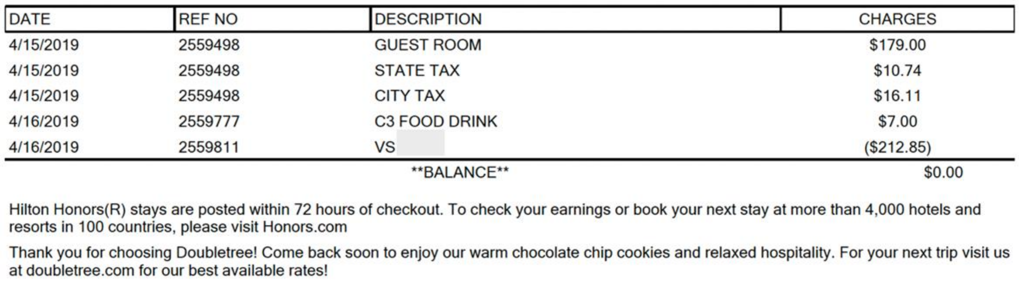
SHOWING TOP 999 IN LAT, LNG AND TOP 990 IN CITY

6.45° N
3.4° E
city Lagos
population 15,279,000

© 2021, Amazon.com, Inc. or its affiliates. © OpenMapTiles © OpenStreetMap contributors. Map data © OpenStreetMap contributors. Map data is available under the Open Data Commons Open Database License. See https://www.openstreetmap.org/copyright.

Chapter 13: Enabling Artificial Intelligence and Machine Learning

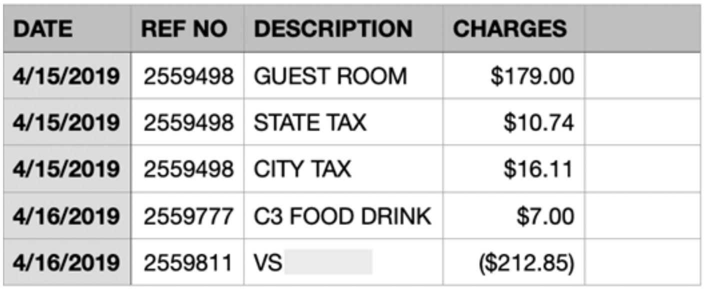




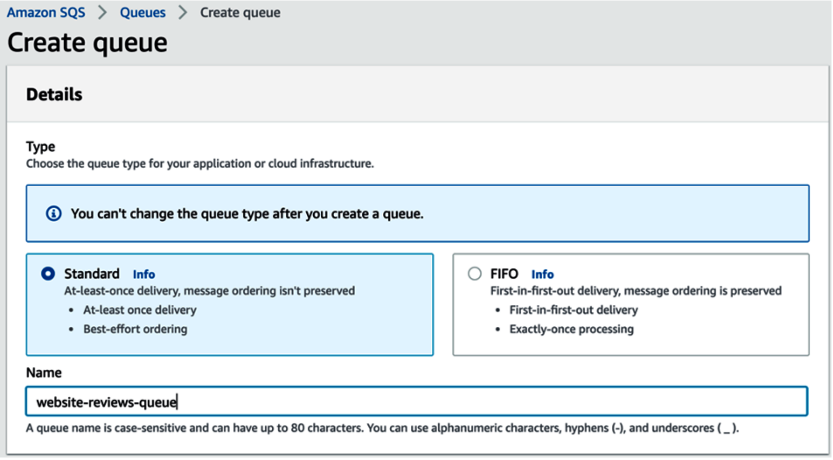
DATE	REF NO	DESCRIPTION	CHARGES
4/15/2019	2559498	GUEST ROOM	$179.00
4/15/2019	2559498	STATE TAX	$10.74
4/15/2019	2559498	CITY TAX	$16.11
4/16/2019	2559777	C3 FOOD DRINK	$7.00
4/16/2019	2559811	VS [redacted]	($212.85)
		BALANCE	$0.00

Hilton Honors(R) stays are posted within 72 hours of checkout. To check your earnings or book your next stay at more than 4,000 hotels and resorts in 100 countries, please visit Honors.com

Thank you for choosing Doubletree! Come back soon to enjoy our warm chocolate chip cookies and relaxed hospitality. For your next trip visit us at doubletree.com for our best available rates!



DATE	REF NO	DESCRIPTION	CHARGES	
4/15/2019	2559498	GUEST ROOM	$179.00	
4/15/2019	2559498	STATE TAX	$10.74	
4/15/2019	2559498	CITY TAX	$16.11	
4/16/2019	2559777	C3 FOOD DRINK	$7.00	
4/16/2019	2559811	VS [redacted]	($212.85)	



Amazon SQS > Queues > Create queue

Create queue

Details

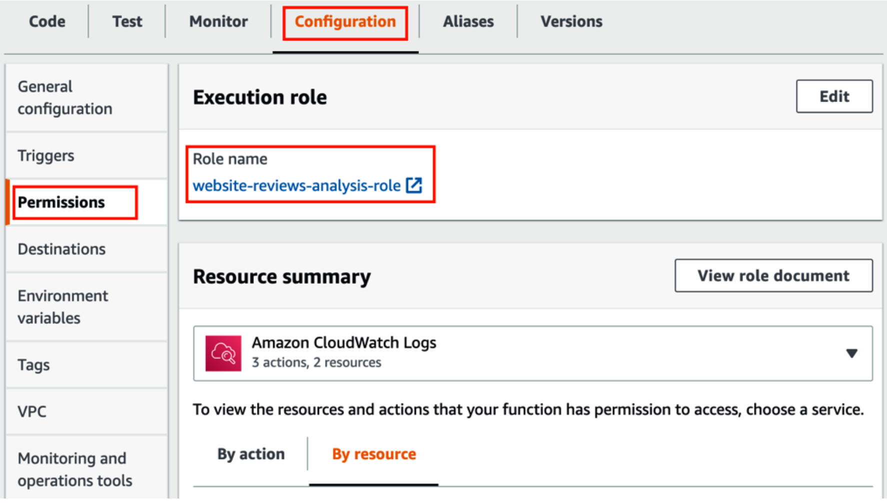
Type
Choose the queue type for your application or cloud infrastructure.

ⓘ You can't change the queue type after you create a queue.

● Standard Info
At-least-once delivery, message ordering isn't preserved
- At-least once delivery
- Best-effort ordering

○ FIFO Info
First-in-first-out delivery, message ordering is preserved
- First-in-first-out delivery
- Exactly-once processing

Name
website-reviews-queue
A queue name is case-sensitive and can have up to 80 characters. You can use alphanumeric characters, hyphens (-), and underscores (_).



Execution role Edit

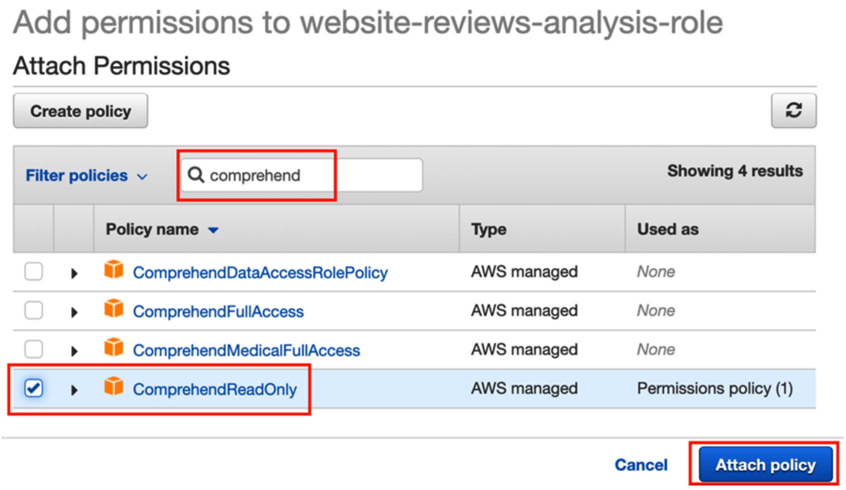
Role name
website-reviews-analysis-role

Resource summary View role document

Amazon CloudWatch Logs
3 actions, 2 resources

To view the resources and actions that your function has permission to access, choose a service.

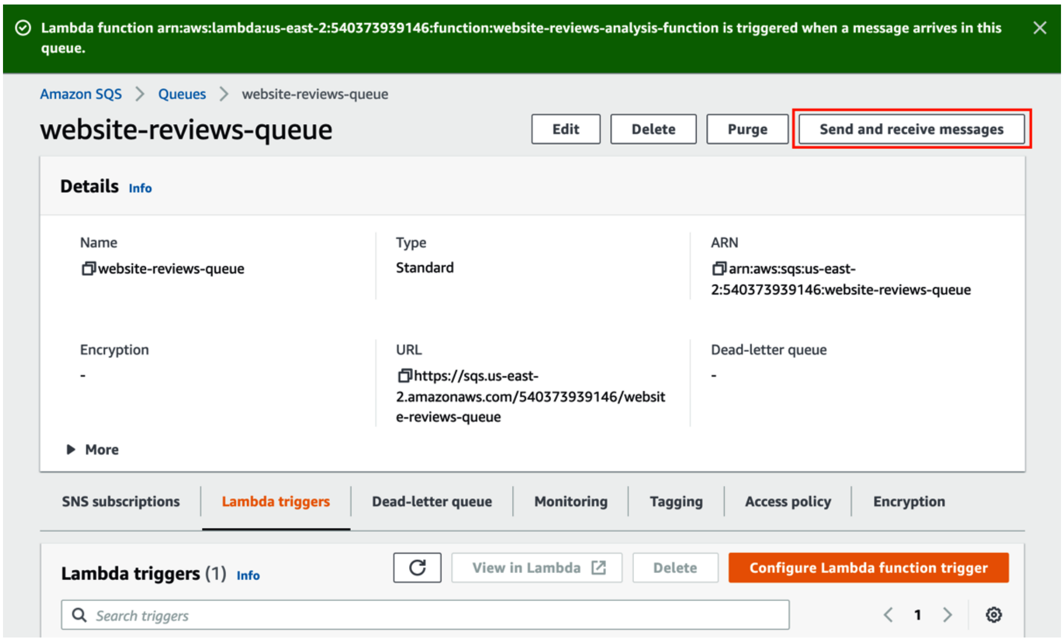
By action | By resource



Add permissions to website-reviews-analysis-role

Attach Permissions

	Policy name	Type	Used as
□	ComprehendDataAccessRolePolicy	AWS managed	None
□	ComprehendFullAccess	AWS managed	None
□	ComprehendMedicalFullAccess	AWS managed	None
☑	ComprehendReadOnly	AWS managed	Permissions policy (1)



Lambda function arn:aws:lambda:us-east-2:540373939146:function:website-reviews-analysis-function is triggered when a message arrives in this queue.

Amazon SQS > Queues > website-reviews-queue

website-reviews-queue

Edit
Delete
Purge
Send and receive messages

Details Info

Name: website-reviews-queue

Type: Standard

ARN: arn:aws:sqs:us-east-2:540373939146:website-reviews-queue

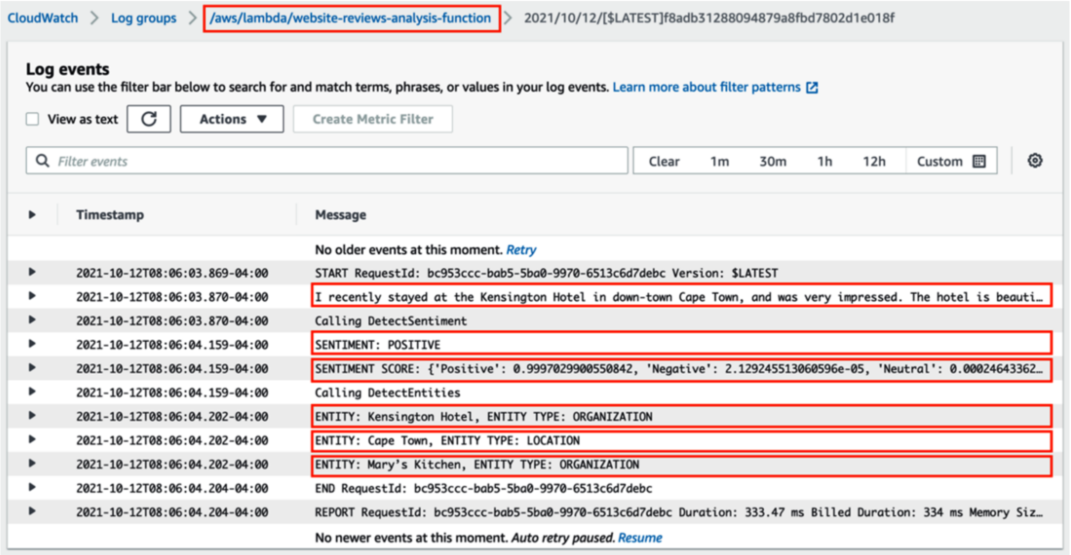
Encryption: -

URL: https://sqs.us-east-2.amazonaws.com/540373939146/website-reviews-queue

Dead-letter queue: -

Lambda triggers (1) Info

View in Lambda
Delete
Configure Lambda function trigger
Search triggers
1



CloudWatch > Log groups > /aws/lambda/website-reviews-analysis-function > 2021/10/12/[$LATEST]f8adb31288094879a8fbd7802d1e018f

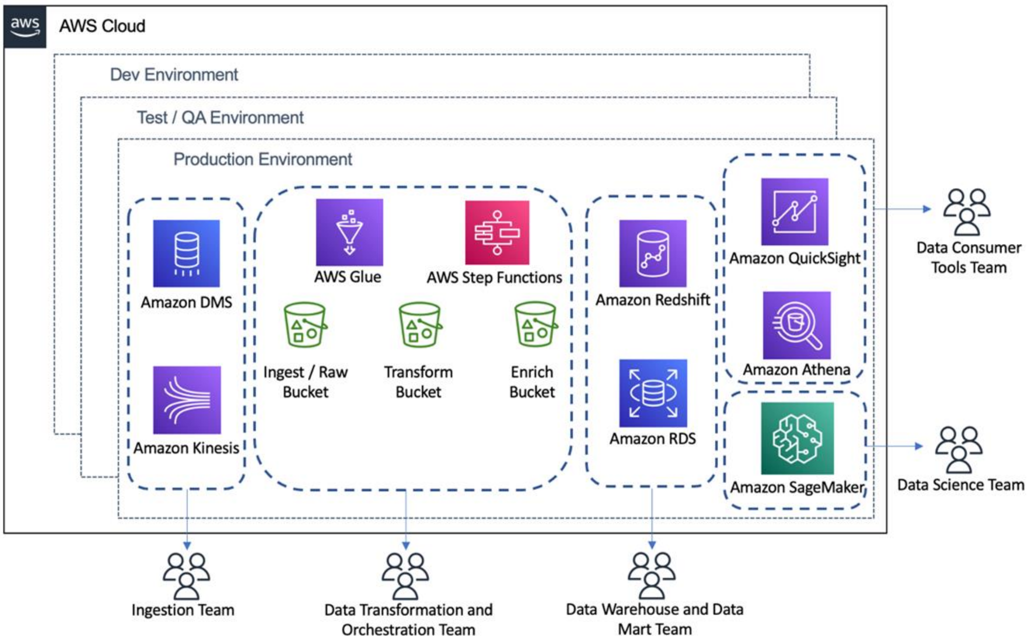
Log events

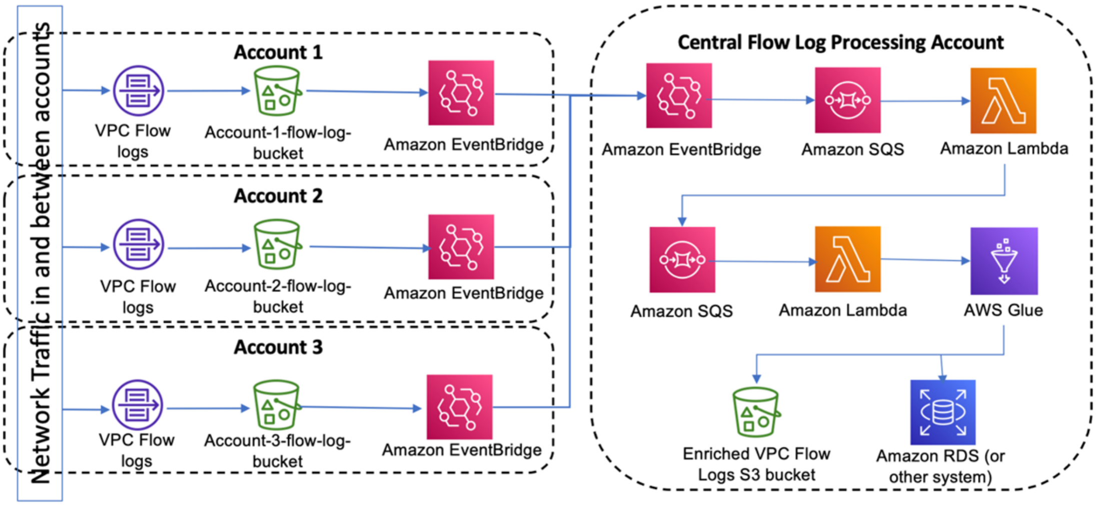
You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

View as text
Actions
Create Metric Filter
Filter events
Clear
1m
30m
1h
12h
Custom

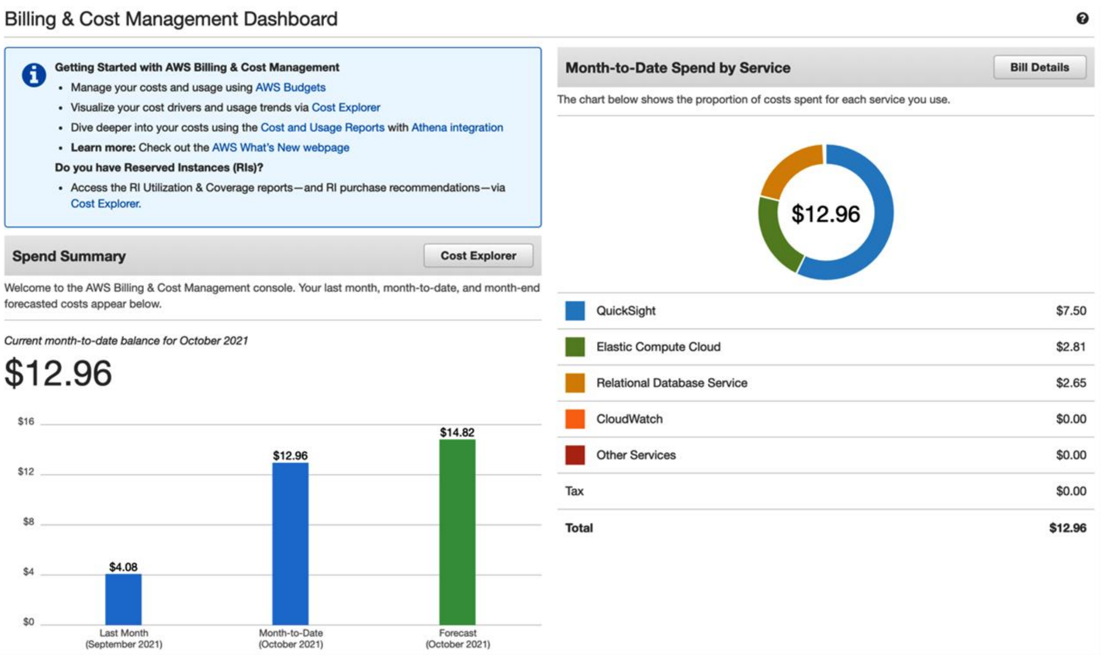
```
Timestamp                       Message
                                No older events at this moment. Retry
2021-10-12T08:06:03.869-04:00   START RequestId: bc953ccc-bab5-5ba0-9970-6513c6d7debc Version: $LATEST
2021-10-12T08:06:03.870-04:00   I recently stayed at the Kensington Hotel in down-town Cape Town, and was very impressed. The hotel is beauti…
2021-10-12T08:06:03.870-04:00   Calling DetectSentiment
2021-10-12T08:06:04.159-04:00   SENTIMENT: POSITIVE
2021-10-12T08:06:04.159-04:00   SENTIMENT SCORE: {'Positive': 0.9997029900550842, 'Negative': 2.129245513060596e-05, 'Neutral': 0.00024643362…
2021-10-12T08:06:04.159-04:00   Calling DetectEntities
2021-10-12T08:06:04.202-04:00   ENTITY: Kensington Hotel, ENTITY TYPE: ORGANIZATION
2021-10-12T08:06:04.202-04:00   ENTITY: Cape Town, ENTITY TYPE: LOCATION
2021-10-12T08:06:04.202-04:00   ENTITY: Mary's Kitchen, ENTITY TYPE: ORGANIZATION
2021-10-12T08:06:04.204-04:00   END RequestId: bc953ccc-bab5-5ba0-9970-6513c6d7debc
2021-10-12T08:06:04.204-04:00   REPORT RequestId: bc953ccc-bab5-5ba0-9970-6513c6d7debc Duration: 333.47 ms Billed Duration: 334 ms Memory Siz…
                                No newer events at this moment. Auto retry paused. Resume
```

Chapter 14: Wrapping Up the First Part of Your Learning Journey





Network Traffic in and between accounts
Account 1
VPC Flow logs
Account-1-flow-log-bucket
Amazon EventBridge
Account 2
VPC Flow logs
Account-2-flow-log-bucket
Amazon EventBridge
Account 3
VPC Flow logs
Account-3-flow-log-bucket
Amazon EventBridge
Central Flow Log Processing Account
Amazon EventBridge
Amazon SQS
Amazon Lambda
Amazon SQS
Amazon Lambda
AWS Glue
Enriched VPC Flow Logs S3 bucket
Amazon RDS (or other system)



Billing & Cost Management Dashboard

Getting Started with AWS Billing & Cost Management

- Manage your costs and usage using AWS Budgets
- Visualize your cost drivers and usage trends via Cost Explorer
- Dive deeper into your costs using the Cost and Usage Reports with Athena integration
- **Learn more:** Check out the AWS What's New webpage

Do you have Reserved Instances (RIs)?

- Access the RI Utilization & Coverage reports—and RI purchase recommendations—via Cost Explorer.

Spend Summary

Welcome to the AWS Billing & Cost Management console. Your last month, month-to-date, and month-end forecasted costs appear below.

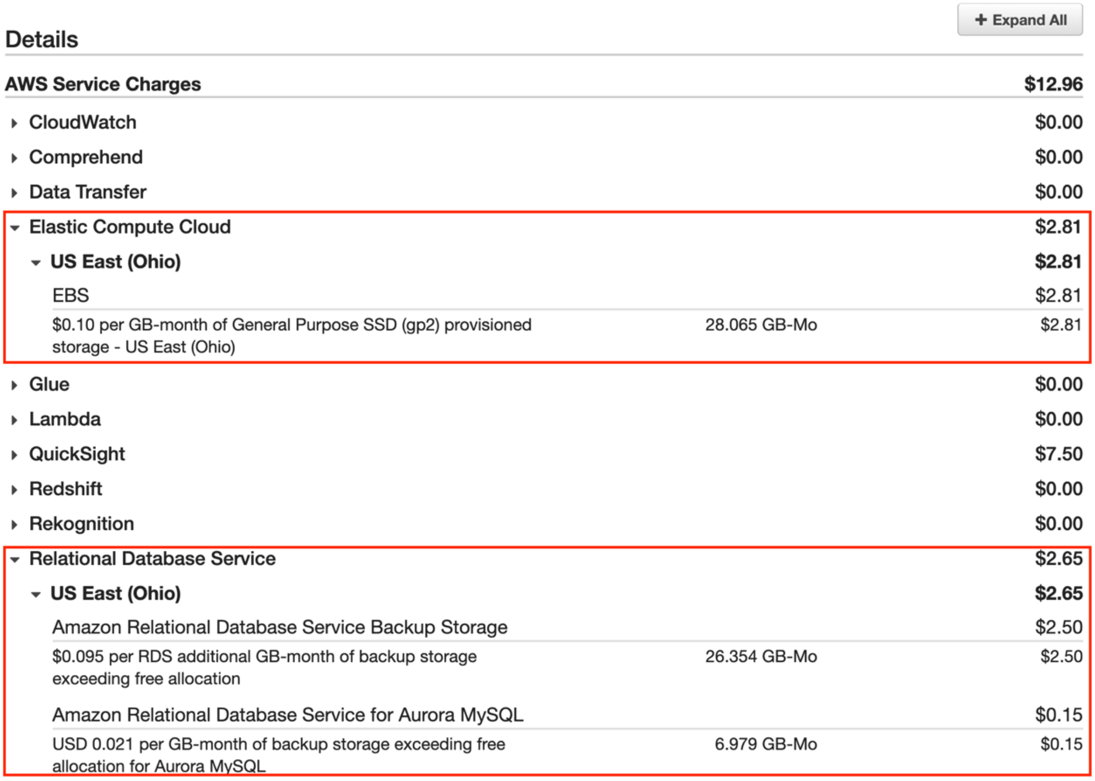
Current month-to-date balance for October 2021

$12.96

Month-to-Date Spend by Service

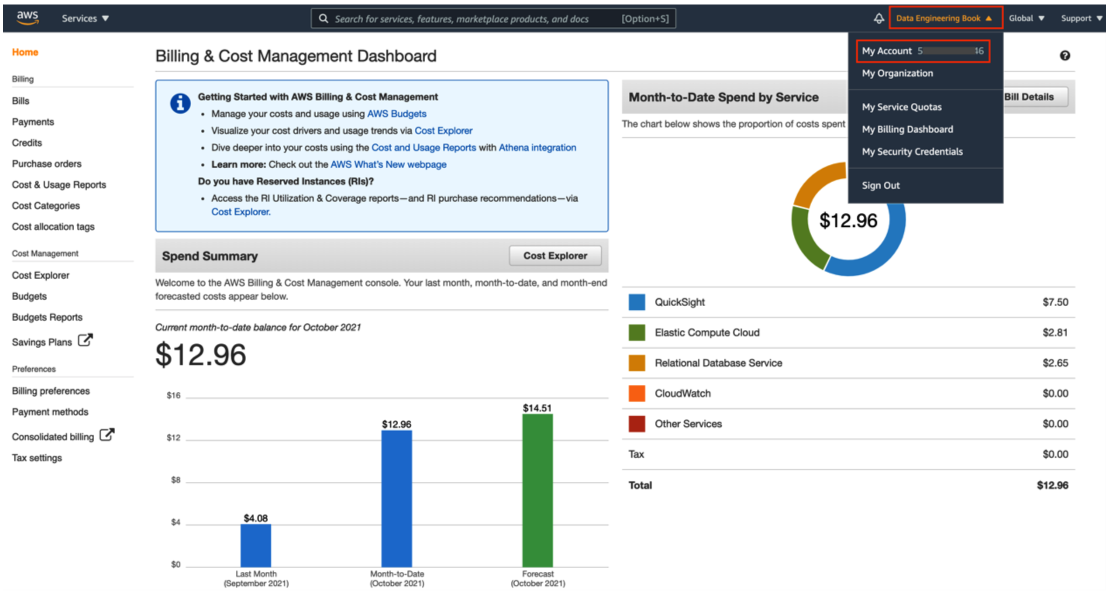
The chart below shows the proportion of costs spent for each service you use.

🟦 QuickSight	$7.50
🟩 Elastic Compute Cloud	$2.81
🟧 Relational Database Service	$2.65
🟧 CloudWatch	$0.00
🟥 Other Services	$0.00
Tax	$0.00
Total	**$12.96**



Details

AWS Service Charges		**$12.96**
CloudWatch		$0.00
Comprehend		$0.00
Data Transfer		$0.00
Elastic Compute Cloud		$2.81
US East (Ohio)		**$2.81**
EBS		$2.81
$0.10 per GB-month of General Purpose SSD (gp2) provisioned storage - US East (Ohio)	28.065 GB-Mo	$2.81
Glue		$0.00
Lambda		$0.00
QuickSight		$7.50
Redshift		$0.00
Rekognition		$0.00
Relational Database Service		$2.65
US East (Ohio)		**$2.65**
Amazon Relational Database Service Backup Storage		$2.50
$0.095 per RDS additional GB-month of backup storage exceeding free allocation	26.354 GB-Mo	$2.50
Amazon Relational Database Service for Aurora MySQL		$0.15
USD 0.021 per GB-month of backup storage exceeding free allocation for Aurora MySQL	6.979 GB-Mo	$0.15



Billing & Cost Management Dashboard

Getting Started with AWS Billing & Cost Management

- Manage your costs and usage using AWS Budgets
- Visualize your cost drivers and usage trends via Cost Explorer
- Dive deeper into your costs using the Cost and Usage Reports with Athena integration
- **Learn more:** Check out the AWS What's New webpage

Do you have Reserved Instances (RIs)?

- Access the RI Utilization & Coverage reports—and RI purchase recommendations—via Cost Explorer.

Spend Summary

Welcome to the AWS Billing & Cost Management console. Your last month, month-to-date, and month-end forecasted costs appear below.

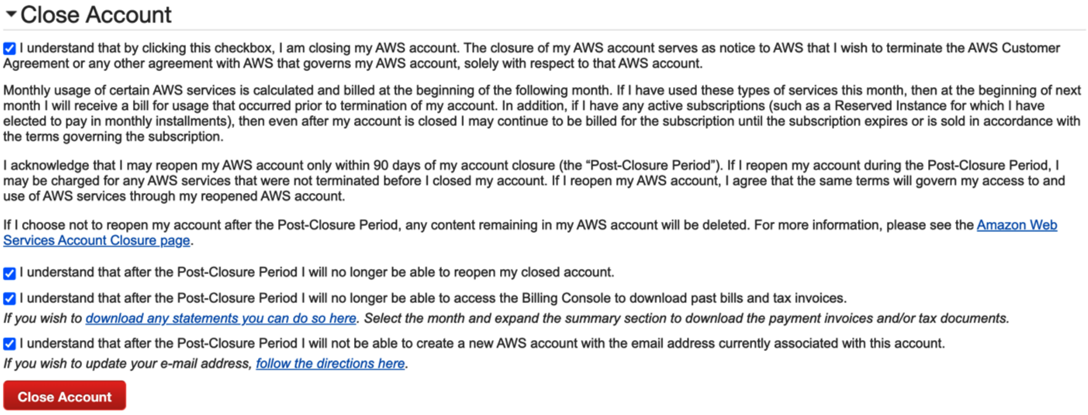
Current month-to-date balance for October 2021

$12.96

Month-to-Date Spend by Service

The chart below shows the proportion of costs spent

🟦	QuickSight	$7.50
🟩	Elastic Compute Cloud	$2.81
🟧	Relational Database Service	$2.65
🟧	CloudWatch	$0.00
🟥	Other Services	$0.00
Tax		$0.00
Total		**$12.96**



▾ Close Account

☑ I understand that by clicking this checkbox, I am closing my AWS account. The closure of my AWS account serves as notice to AWS that I wish to terminate the AWS Customer Agreement or any other agreement with AWS that governs my AWS account, solely with respect to that AWS account.

Monthly usage of certain AWS services is calculated and billed at the beginning of the following month. If I have used these types of services this month, then at the beginning of next month I will receive a bill for usage that occurred prior to termination of my account. In addition, if I have any active subscriptions (such as a Reserved Instance for which I have elected to pay in monthly installments), then even after my account is closed I may continue to be billed for the subscription until the subscription expires or is sold in accordance with the terms governing the subscription.

I acknowledge that I may reopen my AWS account only within 90 days of my account closure (the "Post-Closure Period"). If I reopen my account during the Post-Closure Period, I may be charged for any AWS services that were not terminated before I closed my account. If I reopen my AWS account, I agree that the same terms will govern my access to and use of AWS services through my reopened AWS account.

If I choose not to reopen my account after the Post-Closure Period, any content remaining in my AWS account will be deleted. For more information, please see the Amazon Web Services Account Closure page.

☑ I understand that after the Post-Closure Period I will no longer be able to reopen my closed account.

☑ I understand that after the Post-Closure Period I will no longer be able to access the Billing Console to download past bills and tax invoices.
If you wish to download any statements you can do so here. Select the month and expand the summary section to download the payment invoices and/or tax documents.

☑ I understand that after the Post-Closure Period I will not be able to create a new AWS account with the email address currently associated with this account.
If you wish to update your e-mail address, follow the directions here.