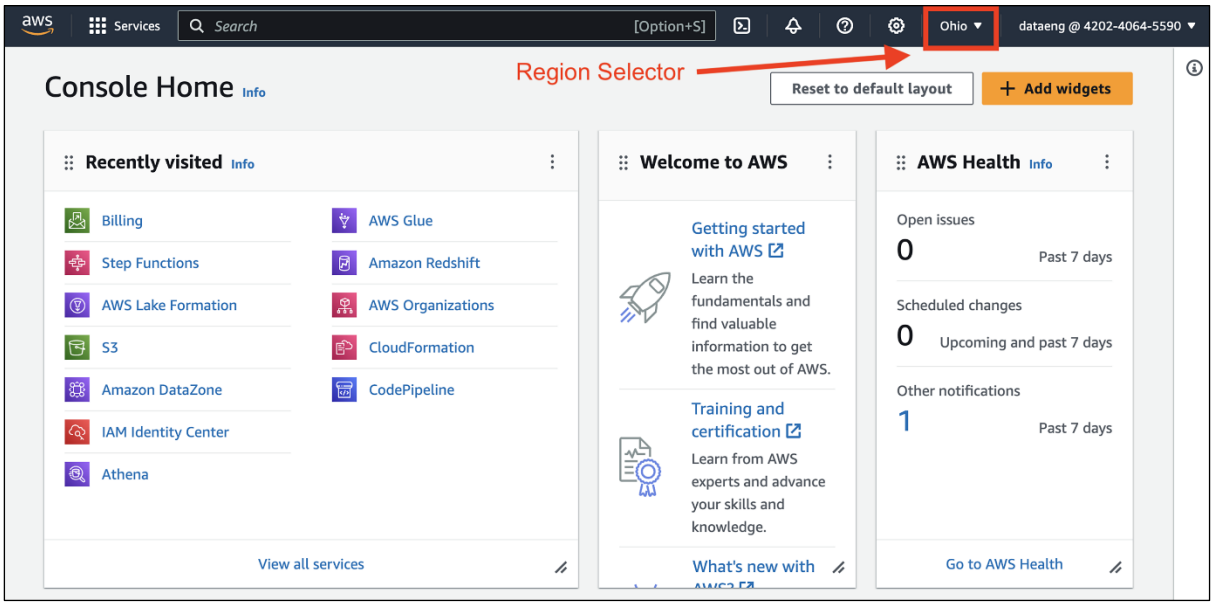
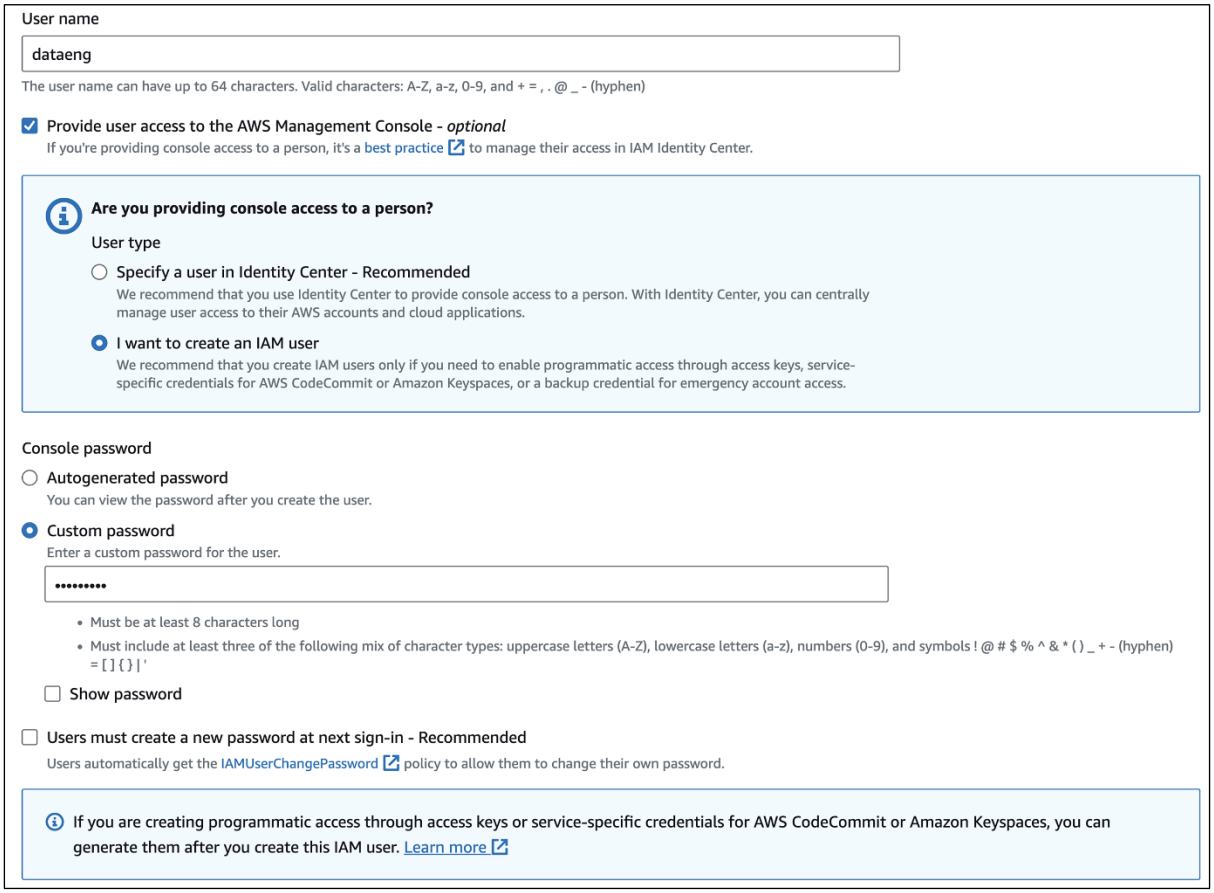
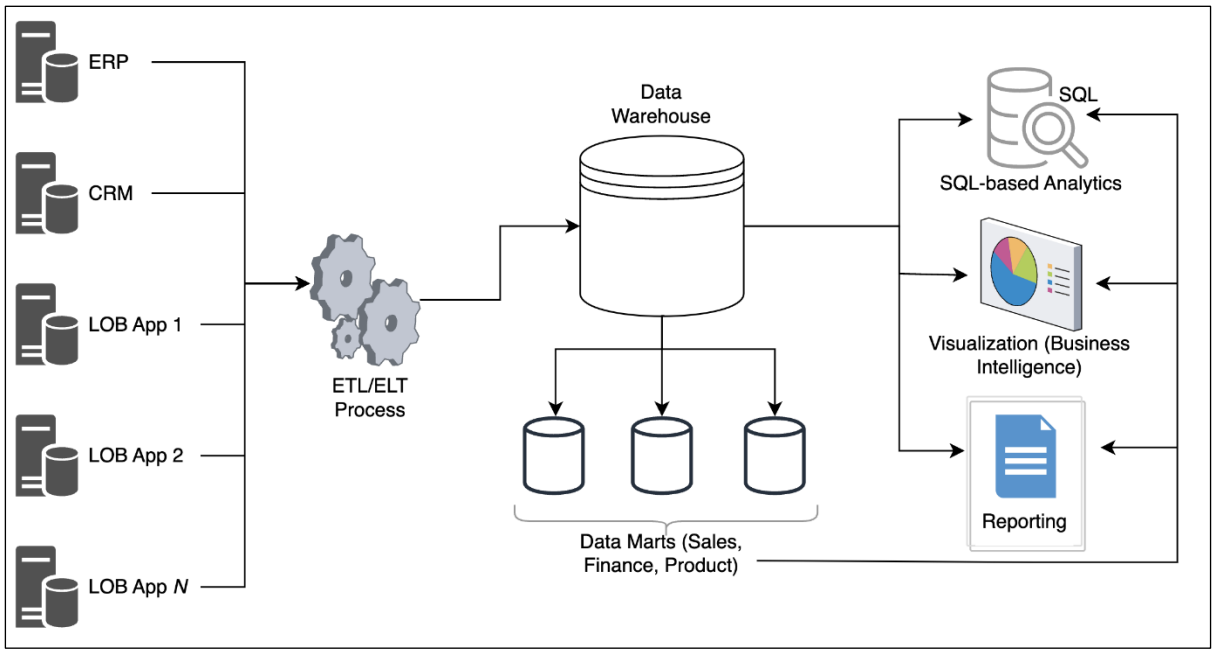
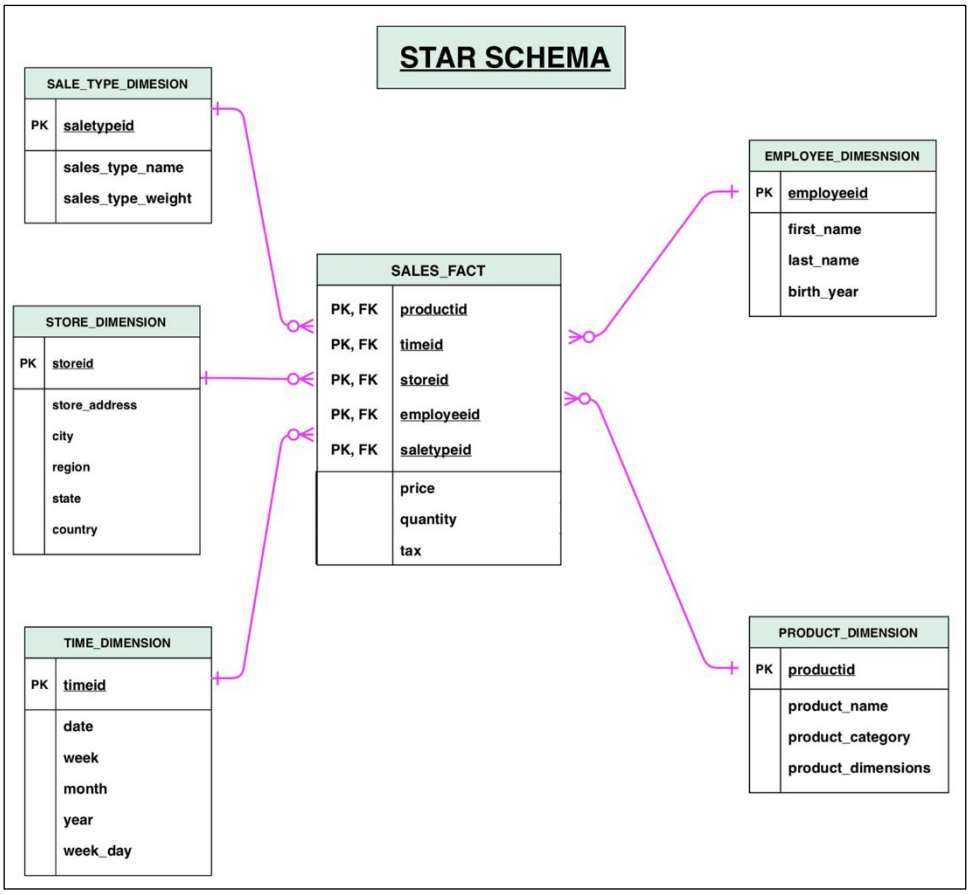
Chapter 1: An Introduction to Data Engineering

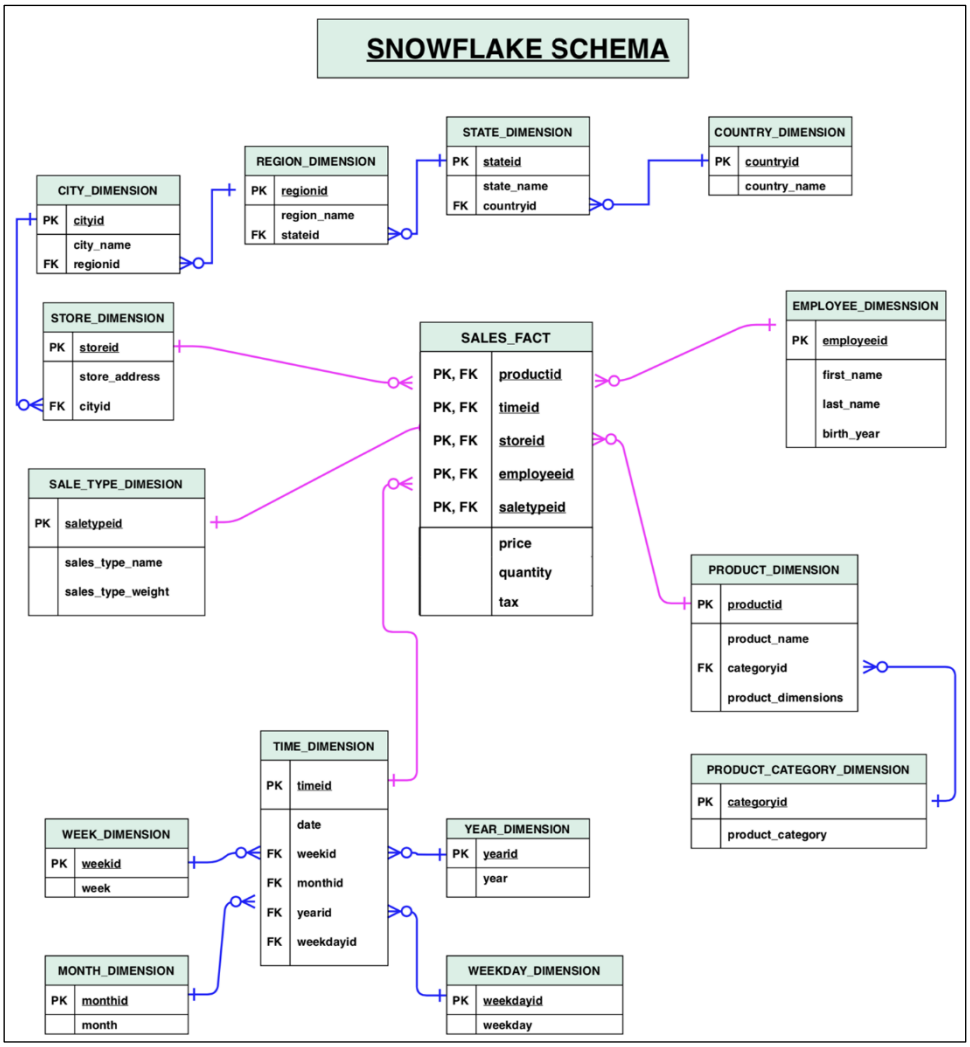


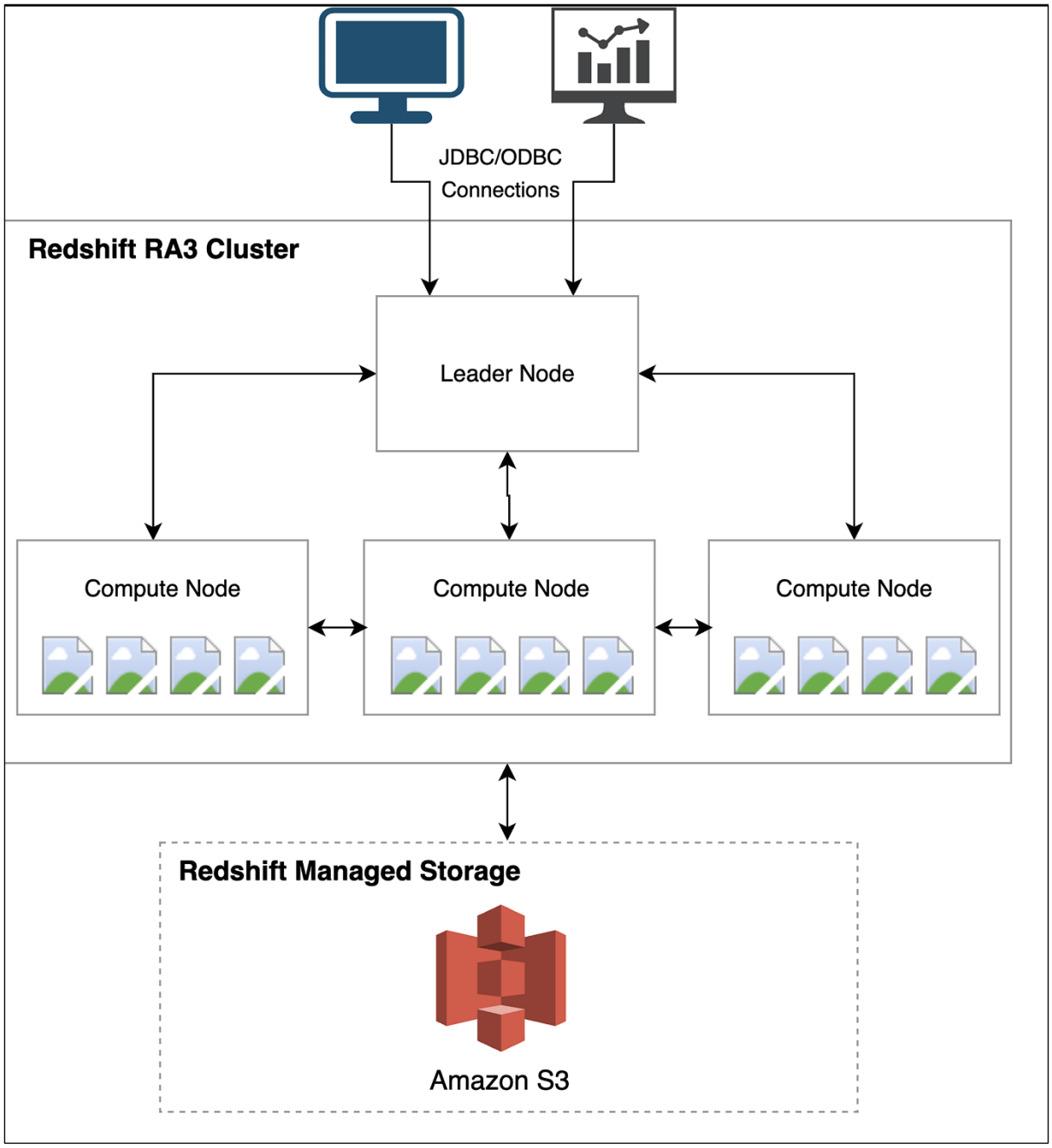


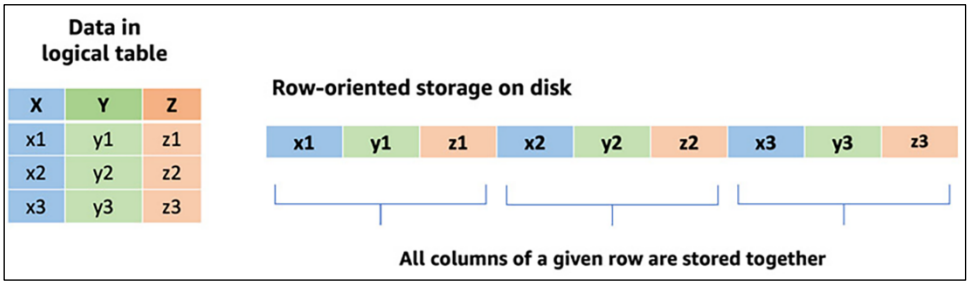
Chapter 2: Data Management Architectures for Analytics

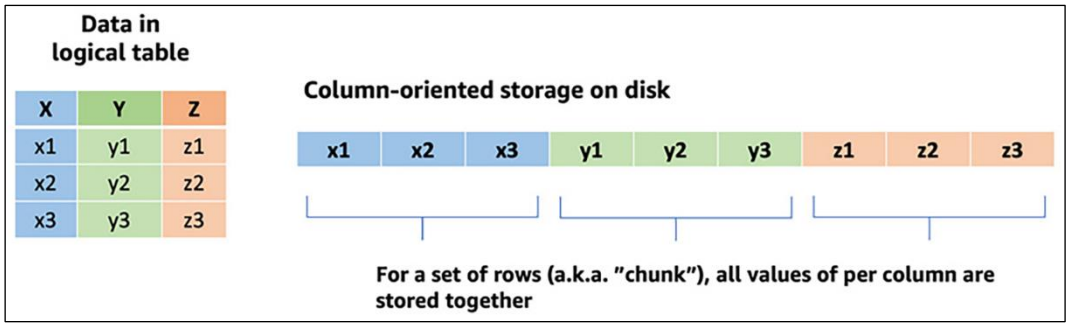


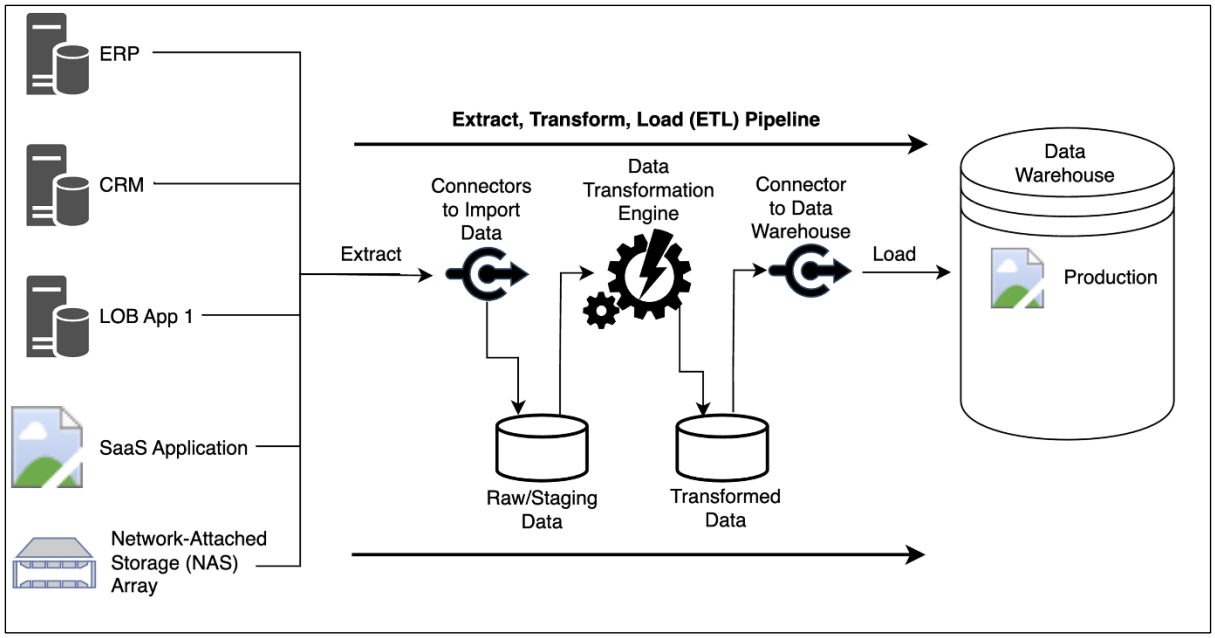


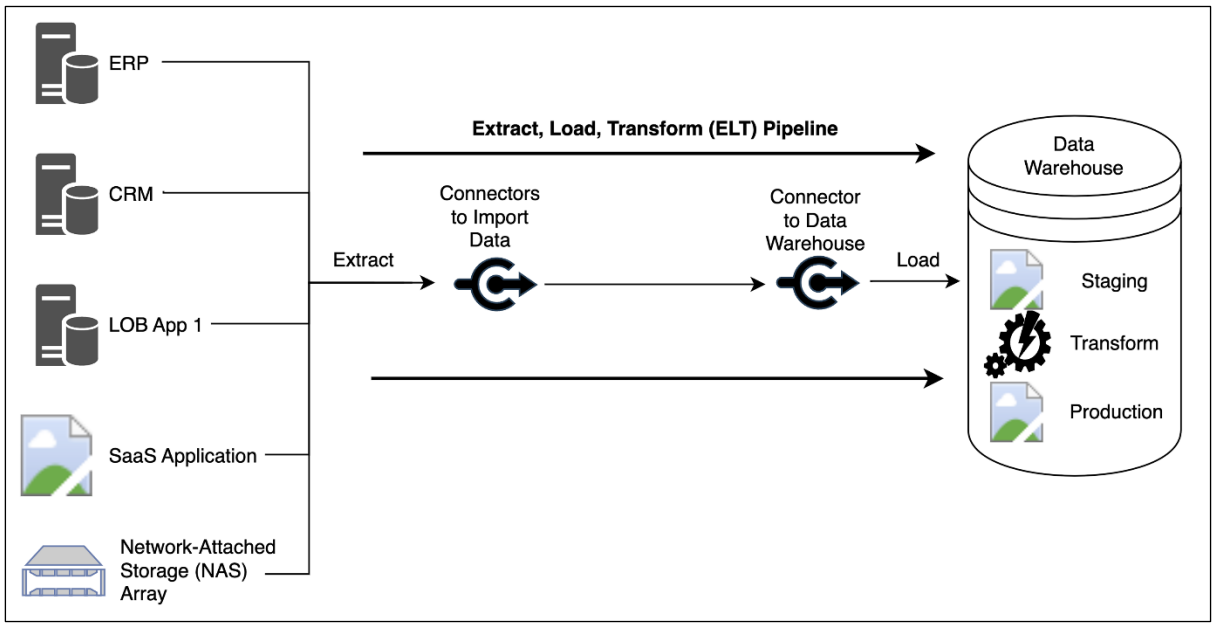

SNOWFLAKE SCHEMA
CITY_DIMENSION
PK cityid
city_name
FK regionid
REGION_DIMENSION
PK regionid
region_name
FK stateid
STATE_DIMENSION
PK stateid
state_name
FK countryid
COUNTRY_DIMENSION
PK countryid
country_name
STORE_DIMENSION
PK storeid
store_address
FK cityid
SALES_FACT
PK, FK productid
PK, FK timeid
PK, FK storeid
PK, FK employeeid
PK, FK saletypeid
price
quantity
tax
EMPLOYEE_DIMESNSION
PK employeeid
first_name
last_name
birth_year
SALE_TYPE_DIMESION
PK saletypeid
sales_type_name
sales_type_weight
PRODUCT_DIMENSION
PK productid
product_name
FK categoryid
product_dimensions
PRODUCT_CATEGORY_DIMENSION
PK categoryid
product_category
TIME_DIMENSION
PK timeid
date
FK weekid
FK monthid
FK yearid
FK weekdayid
WEEK_DIMENSION
PK weekid
week
YEAR_DIMENSION
PK yearid
year
MONTH_DIMENSION
PK monthid
month
WEEKDAY_DIMENSION
PK weekdayid
weekday

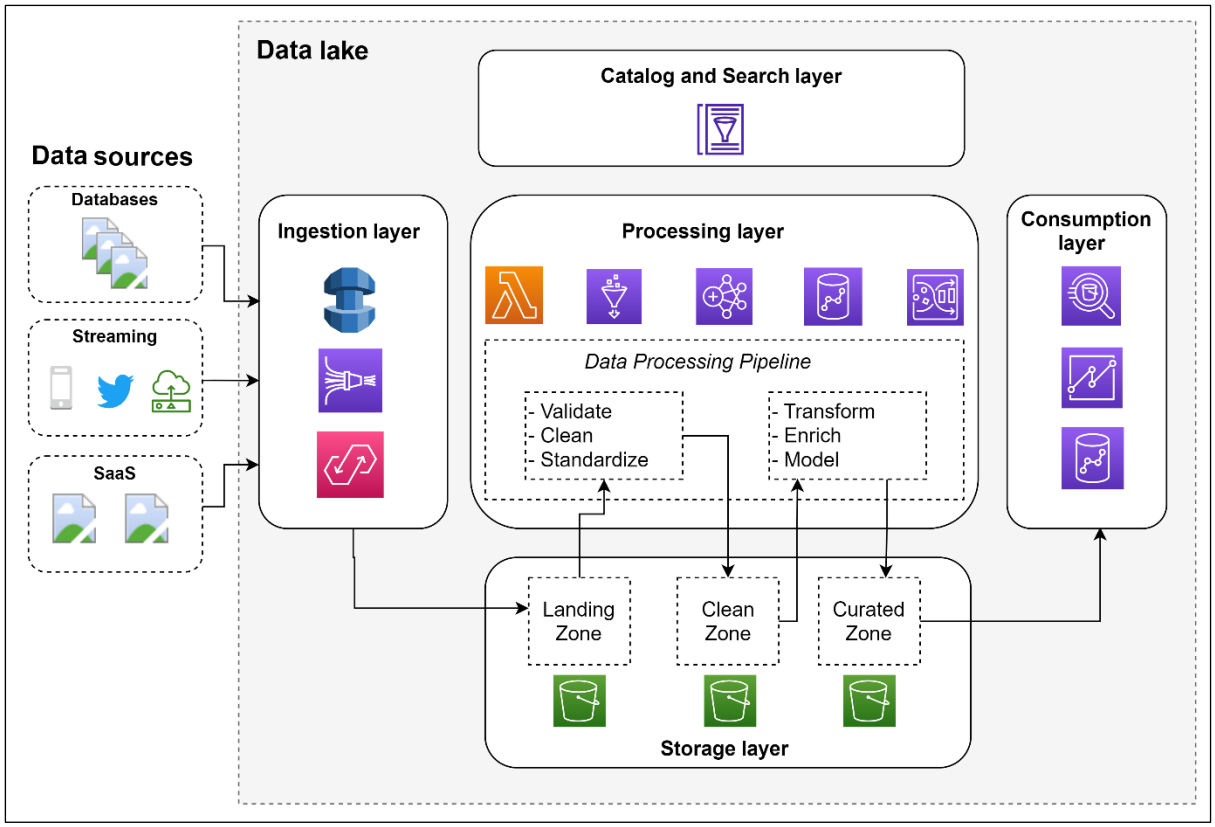

JDBC/ODBC
Connections
Redshift RA3 Cluster
Leader Node
Compute Node
Compute Node
Compute Node
Redshift Managed Storage
Amazon S3

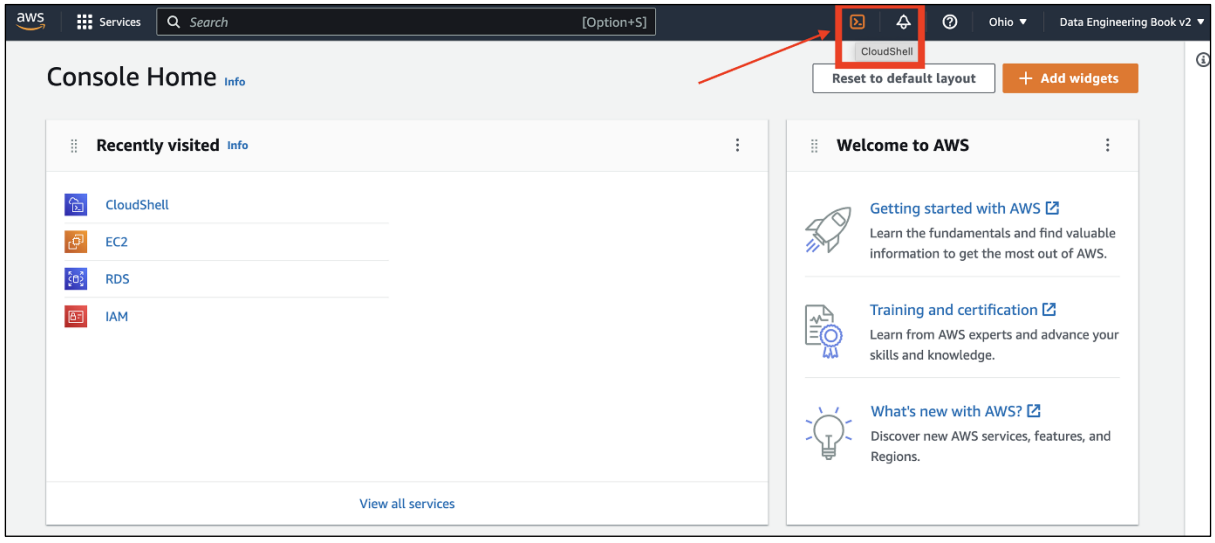

Data in
logical table
X
Y
Z
x1
y1
z1
x2
y2
z2
x3
y3
z3
Row-oriented storage on disk
x1
y1
z1
x2
y2
z2
x3
y3
z3
All columns of a given row are stored together

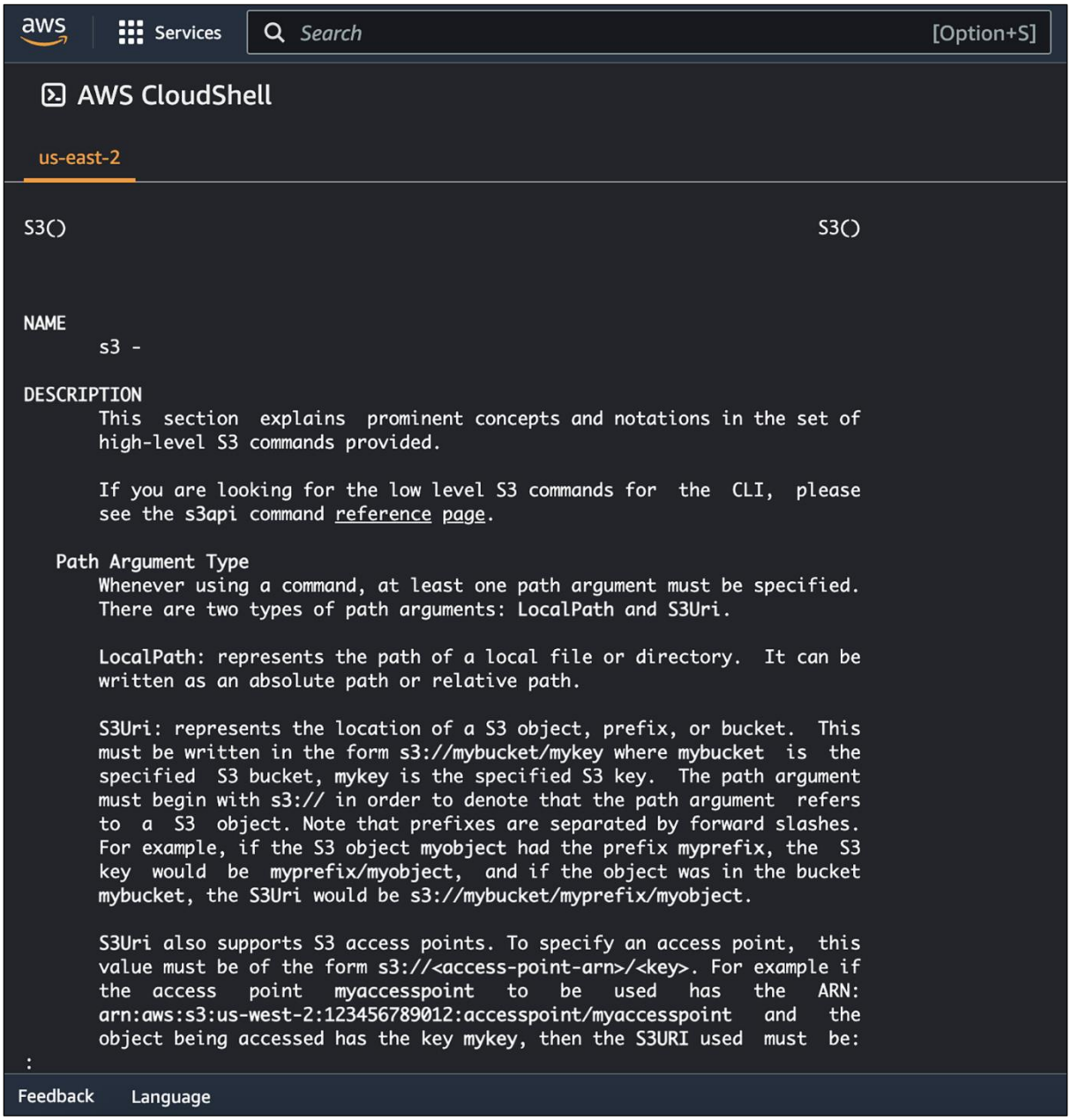

Data in logical table
X
Y
Z
x1
y1
z1
x2
y2
z2
x3
y3
z3
Column-oriented storage on disk
x1
x2
x3
y1
y2
y3
z1
z2
z3
For a set of rows (a.k.a. "chunk"), all values of per column are stored together


ERP
CRM
LOB App 1
SaaS Application
Network-Attached Storage (NAS) Array
Extract, Transform, Load (ETL) Pipeline
Extract
Connectors to Import Data
Data Transformation Engine
Connector to Data Warehouse
Load
Raw/Staging Data
Transformed Data
Data Warehouse
Production

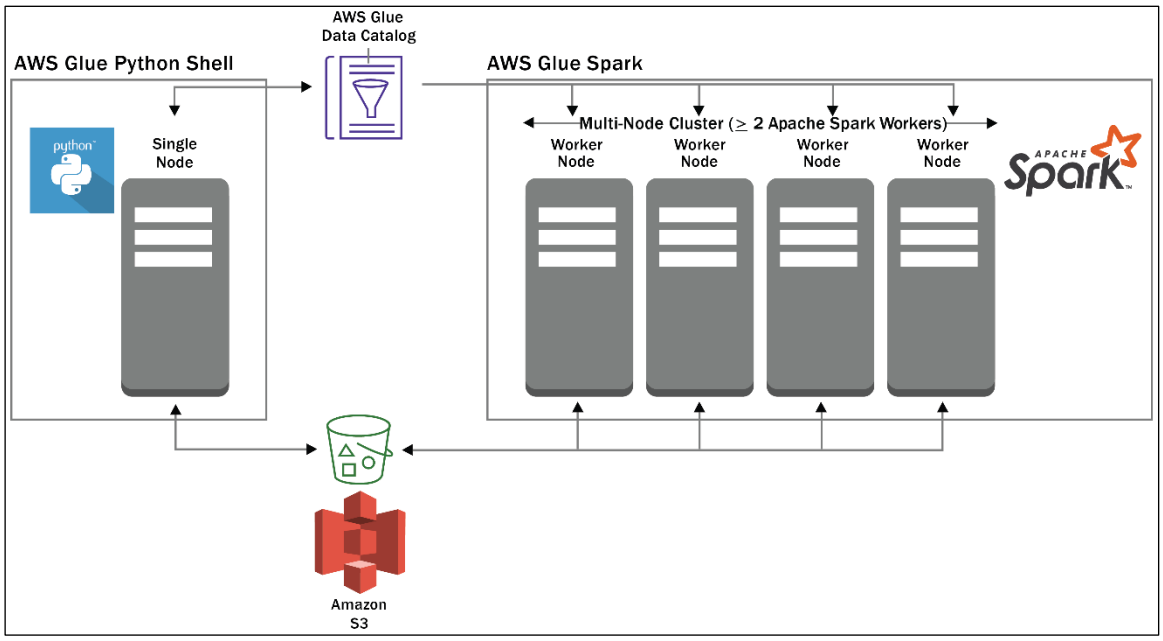

ERP
CRM
LOB App 1
SaaS Application
Network-Attached Storage (NAS) Array
Extract, Load, Transform (ELT) Pipeline
Extract
Connectors to Import Data
Connector to Data Warehouse
Load
Data Warehouse
Staging
Transform
Production

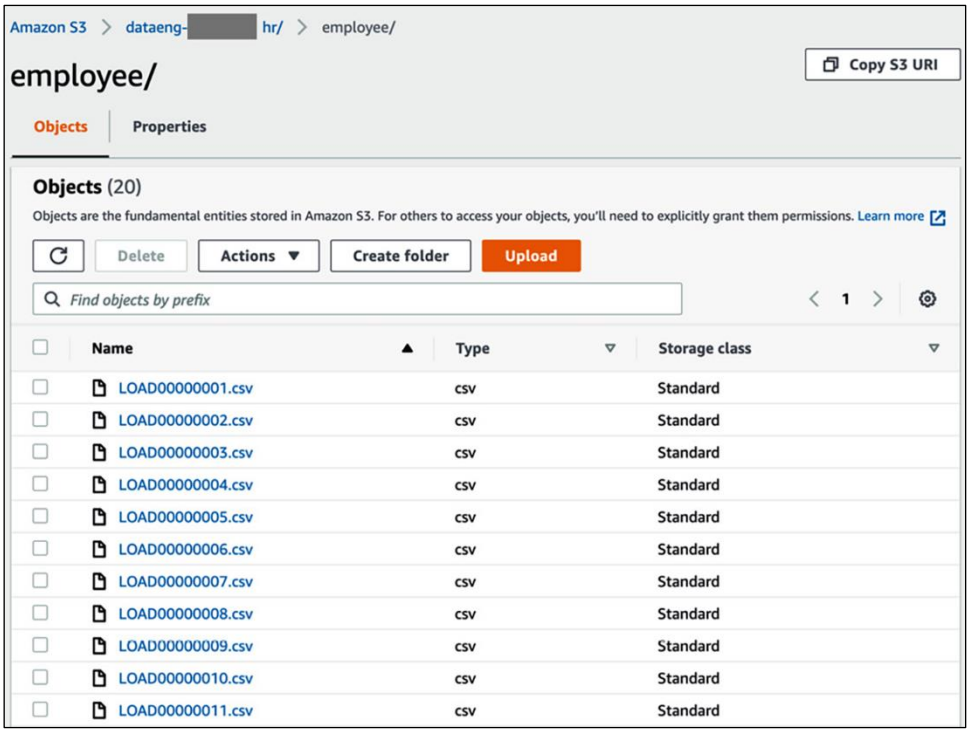

Data lake
Catalog and Search layer
Data sources
Databases
Streaming
SaaS
Ingestion layer
Processing layer
Data Processing Pipeline
- Validate
- Clean
- Standardize
- Transform
- Enrich
- Model
Consumption layer
Landing Zone
Clean Zone
Curated Zone
Storage layer


aws
Services
Search
[Option+S]
CloudShell
Ohio
Data Engineering Book v2
Console Home Info
Reset to default layout
Add widgets
Recently visited Info
CloudShell
EC2
RDS
IAM
View all services
Welcome to AWS
Getting started with AWS
Learn the fundamentals and find valuable information to get the most out of AWS.
Training and certification
Learn from AWS experts and advance your skills and knowledge.
What's new with AWS?
Discover new AWS services, features, and Regions.


aws
Services
Search
[Option+S]
AWS CloudShell
us-east-2
S3()
S3()
NAME
s3 -
DESCRIPTION
This section explains prominent concepts and notations in the set of high-level S3 commands provided.
If you are looking for the low level S3 commands for the CLI, please see the s3api command reference page.
Path Argument Type
Whenever using a command, at least one path argument must be specified. There are two types of path arguments: LocalPath and S3Uri.
LocalPath: represents the path of a local file or directory. It can be written as an absolute path or relative path.
S3Uri: represents the location of a S3 object, prefix, or bucket. This must be written in the form s3://mybucket/mykey where mybucket is the specified S3 bucket, mykey is the specified S3 key. The path argument must begin with s3:// in order to denote that the path argument refers to a S3 object. Note that prefixes are separated by forward slashes. For example, if the S3 object myobject had the prefix myprefix, the S3 key would be myprefix/myobject, and if the object was in the bucket mybucket, the S3Uri would be s3://mybucket/myprefix/myobject.
S3Uri also supports S3 access points. To specify an access point, this value must be of the form s3://<access-point-arn>/<key>. For example if the access point myaccesspoint to be used has the ARN: arn:aws:s3:us-west-2:123456789012:accesspoint/myaccesspoint and the object being accessed has the key mykey, then the S3URI used must be:
:
Feedback
Language

Chapter 3: The AWS Data Engineer's Toolkit



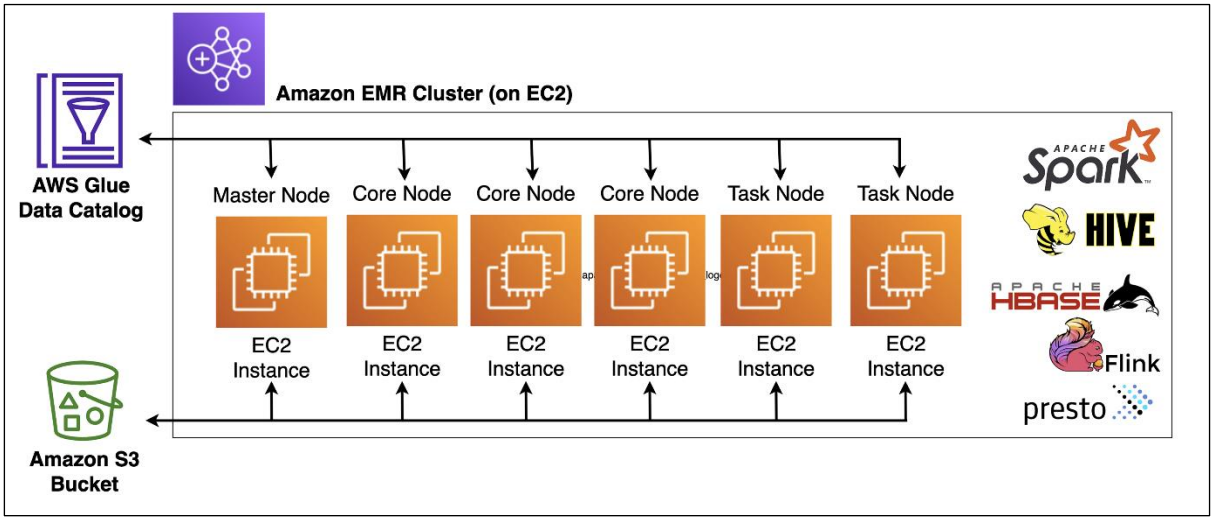


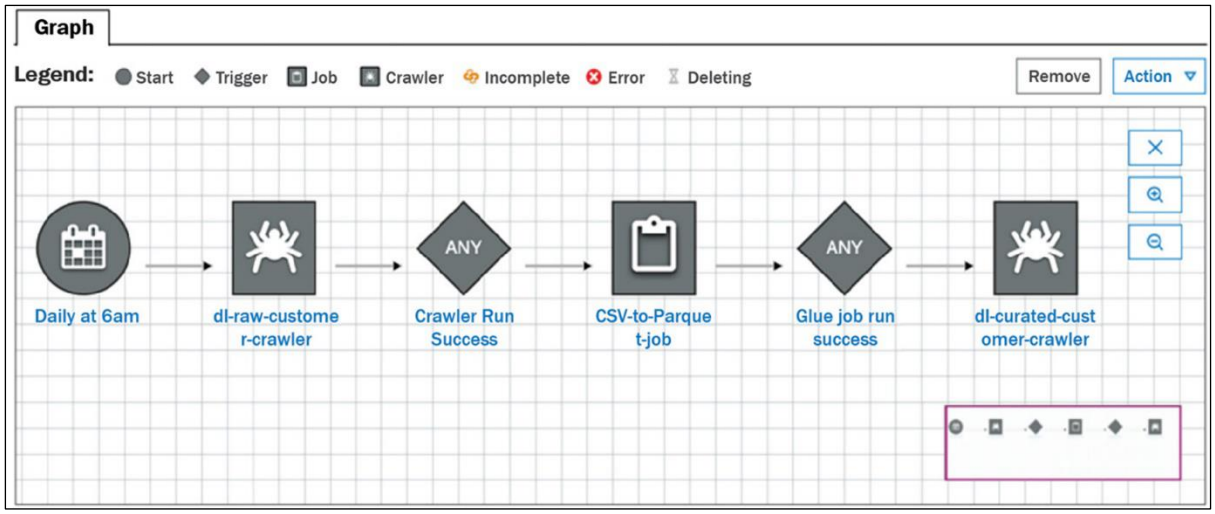
Name	employee
Description	
Database	hr
Classification	csv
Location	s3://dataeng-temp/hr/employee/
Connection	
Deprecated	No
Last updated	Wed Dec 09 21:46:50 GMT-500 2020
Input format	org.apache.hadoop.mapred.TextInputFormat
Output format	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde 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

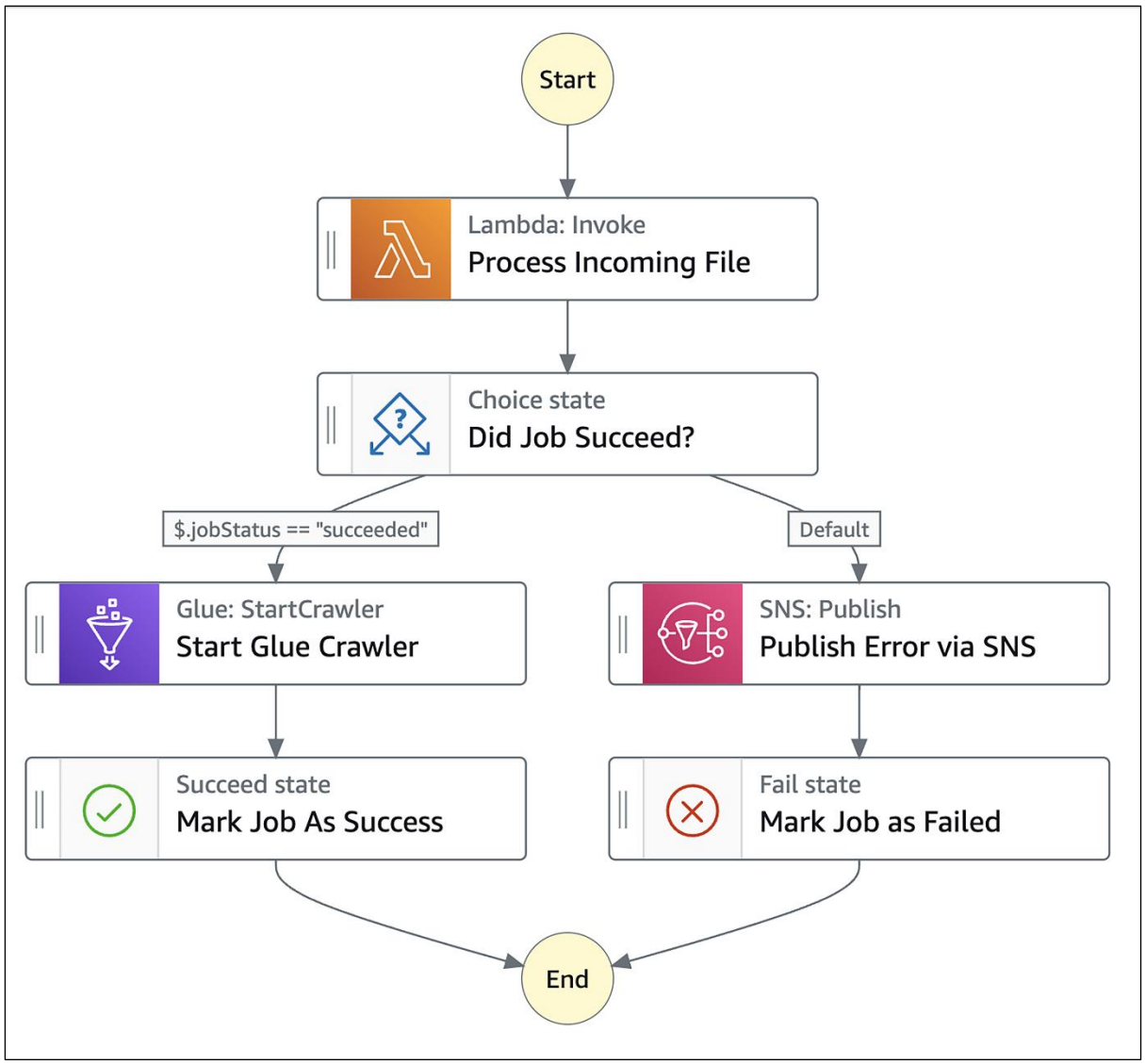
Schema

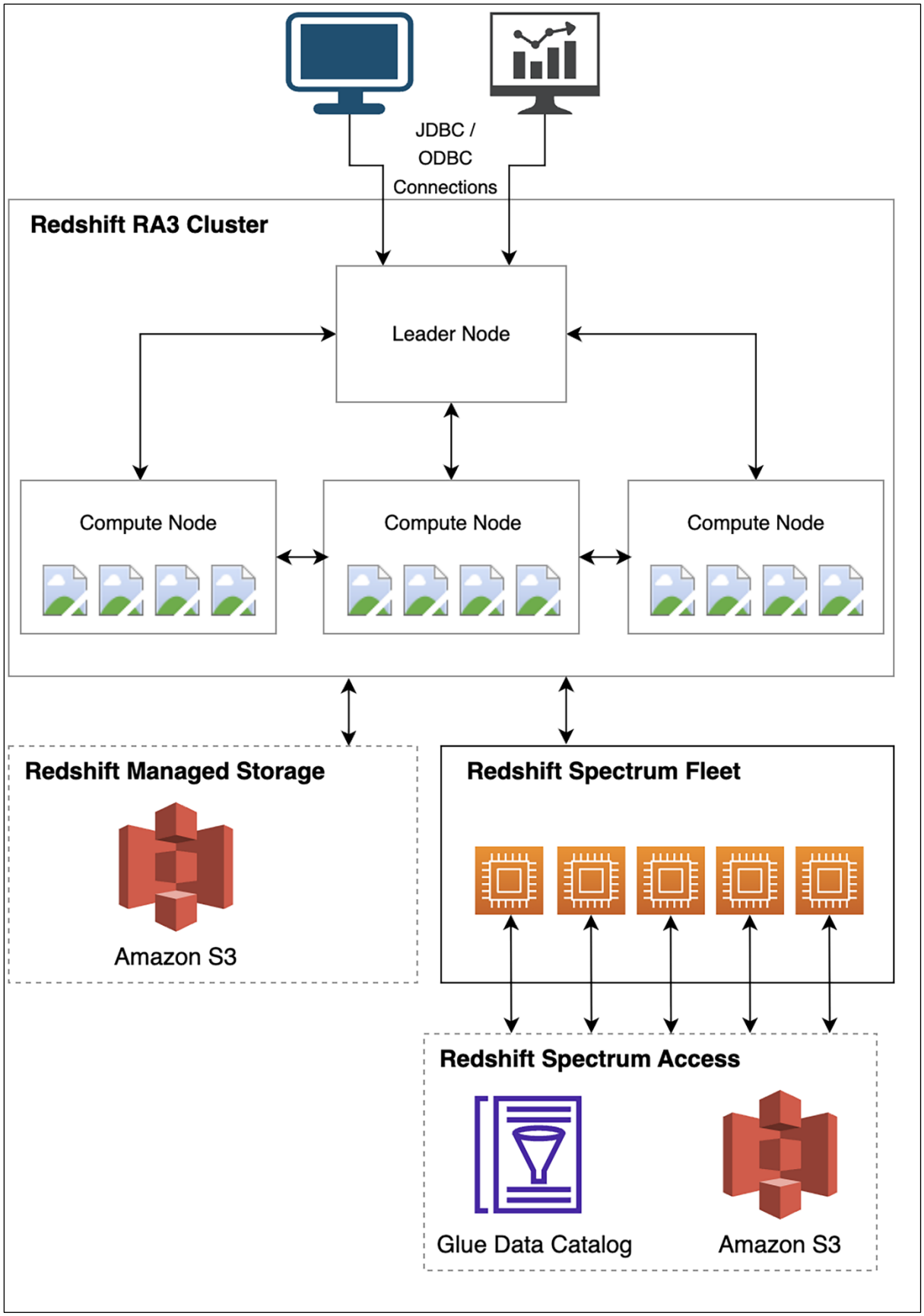
Showing: 1 -

	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		

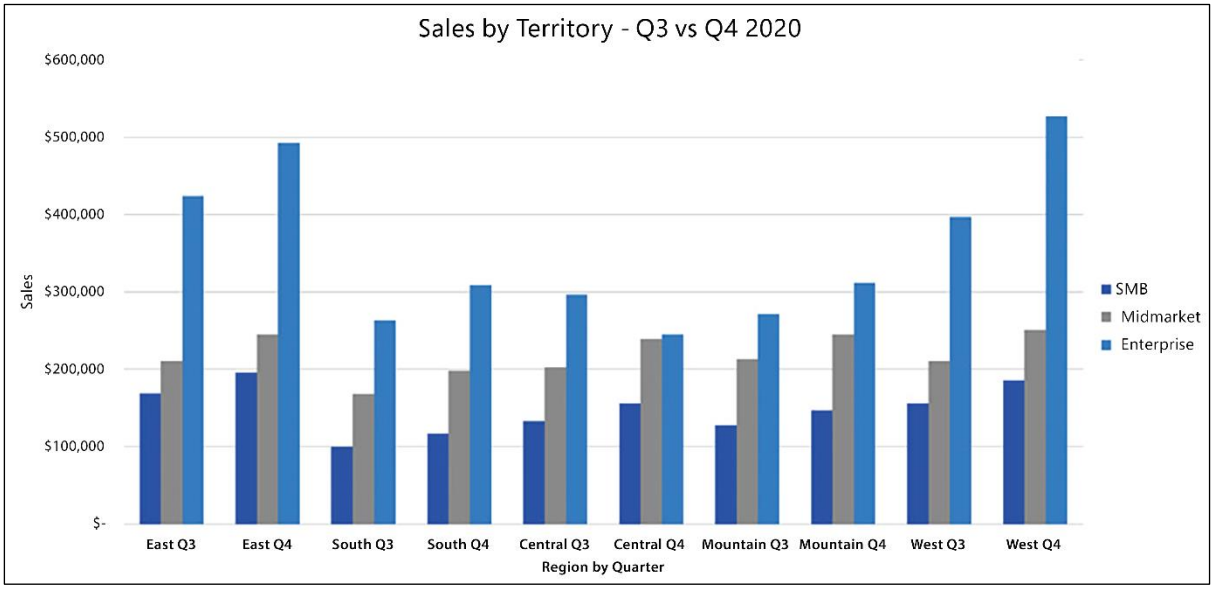


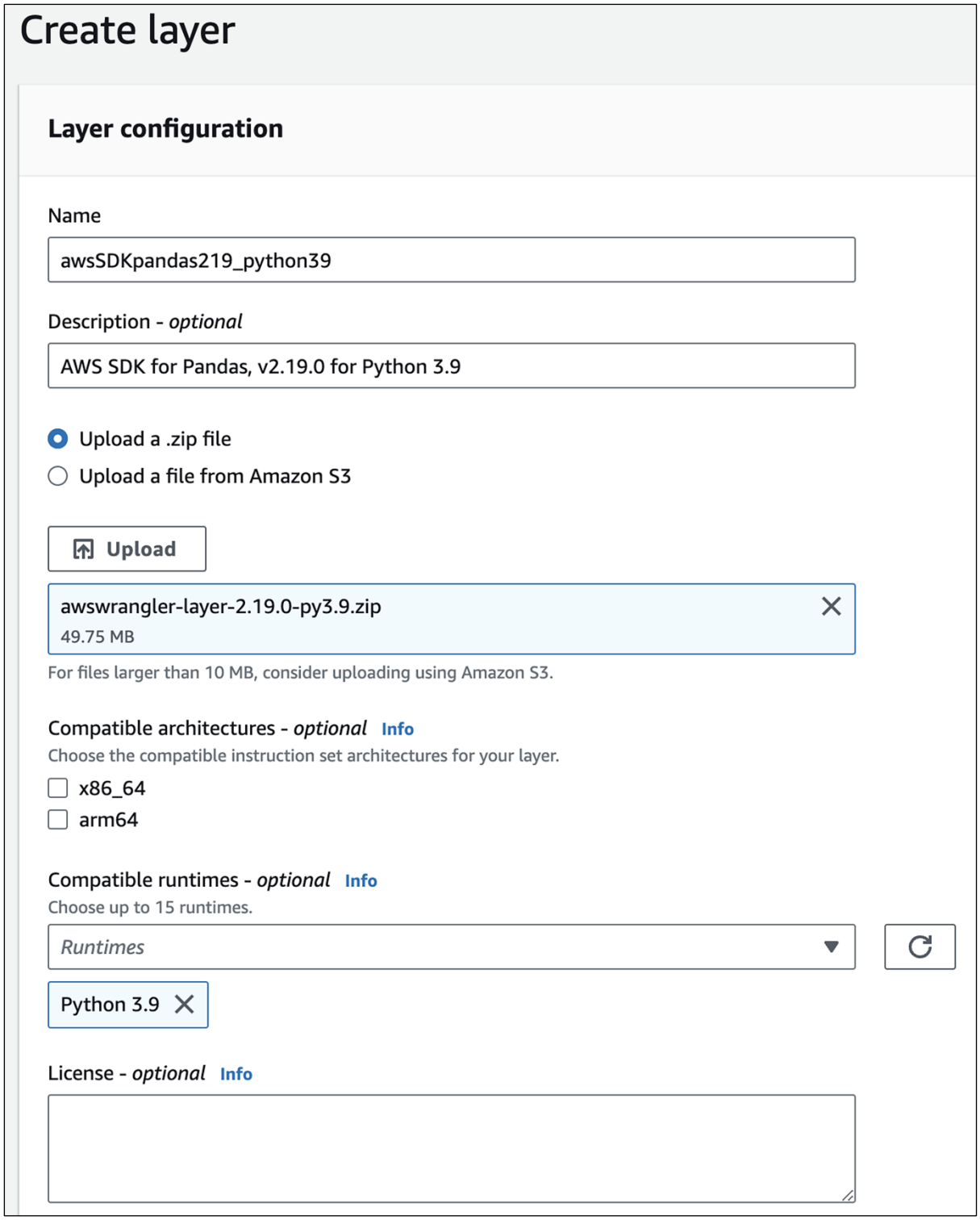

Graph
Legend:
Start
Trigger
Job
Crawler
Incomplete
Error
Deleting
Remove
Action
Daily at 6am
dl-raw-custome r-crawler
ANY
Crawler Run Success
CSV-to-Parque t-job
ANY
Glue job run success
dl-curated-cust omer-crawler


Start
Lambda: Invoke
Process Incoming File
Choice state
Did Job Succeed?
$.jobStatus == "succeeded"
Default
Glue: StartCrawler
Start Glue Crawler
SNS: Publish
Publish Error via SNS
Succeed state
Mark Job As Success
Fail state
Mark Job as Failed
End


JDBC /
ODBC
Connections
Redshift RA3 Cluster
Leader Node
Compute Node
Compute Node
Compute Node
Redshift Managed Storage
Amazon S3
Redshift Spectrum Fleet
Redshift Spectrum Access
Glue Data Catalog
Amazon S3

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
awsSDKpandas219_python39
Description - optional
AWS SDK for Pandas, v2.19.0 for Python 3.9
Upload a .zip file
Upload a file from Amazon S3
Upload
awswrangler-layer-2.19.0-py3.9.zip
49.75 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 3.9
License - optional Info

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. Note that the console code editor supports only Node.js, Python, and Ruby.

Python 3.9

Architecture Info

Choose the instruction set architecture you want for your function code.

- (•) x86_64
- ( ) arm64

Permissions Info

By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this default role later when adding triggers.

▼ 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 logs to Amazon CloudWatch Logs.

DataEngLambdaS3CWGlueRole

View the DataEngLambdaS3CWGlueRole role on the IAM console.

Lambda > Layers > Add layer

Add layer

Function runtime settings

Runtime	Architecture
Python 3.9	x86_64

Choose a layer

Layer source Info

Choose from layers with a compatible runtime and instruction set architecture or specify the Amazon Resource Name (ARN) of a layer version. You can also create a new layer.

- ( ) AWS layers — Choose a layer from a list of layers provided by AWS.
- (•) Custom layers — Choose a layer from a list of layers created by your AWS account or organization.
- ( ) Specify an ARN — Specify a layer by providing the ARN.

Custom layers

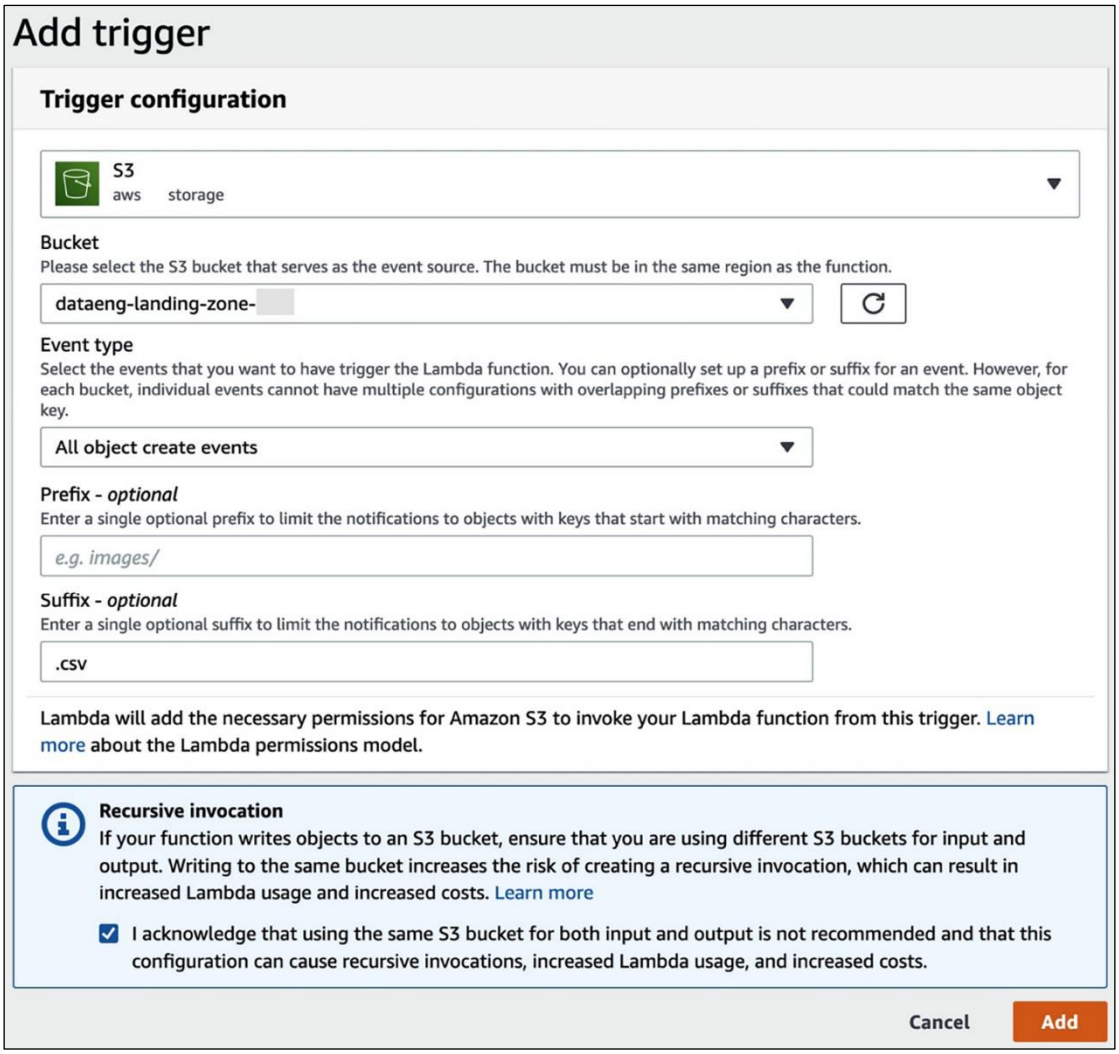
Layers created by your AWS account or organization that are compatible with your function's runtime.

awsSDKpandas219_python39

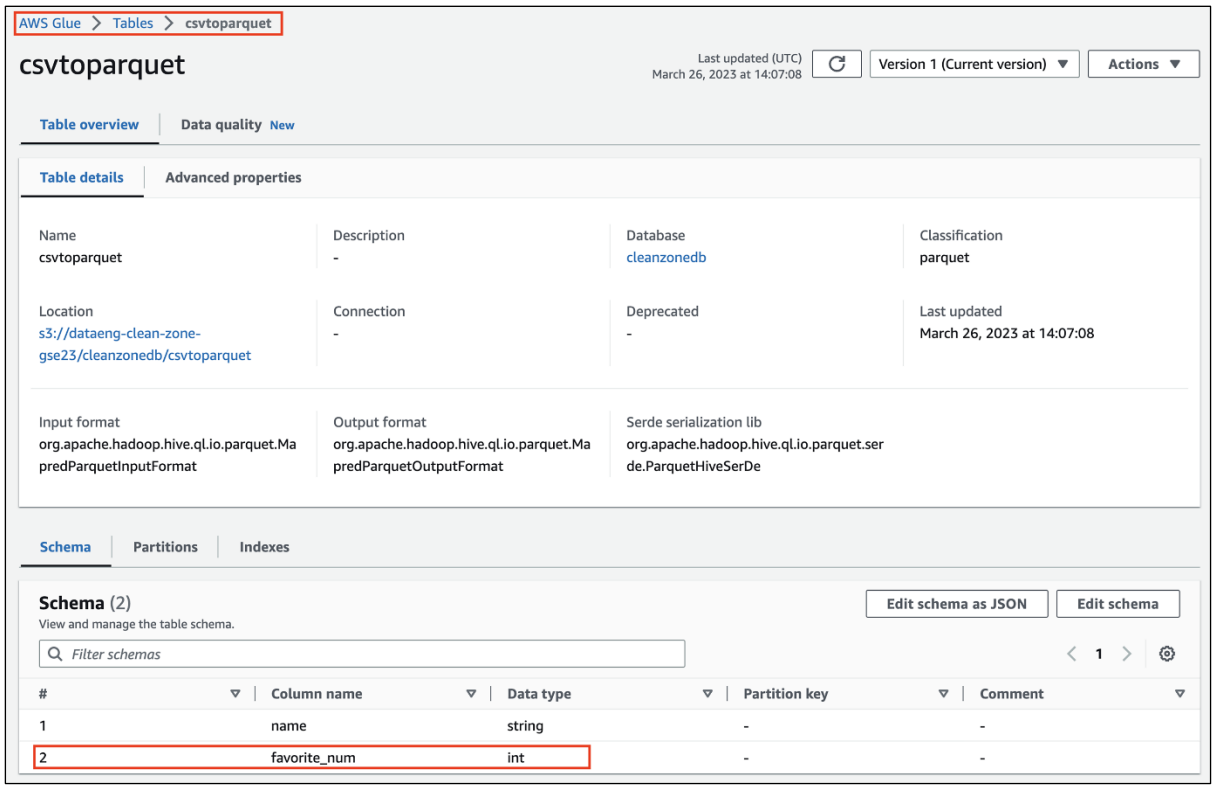
Version

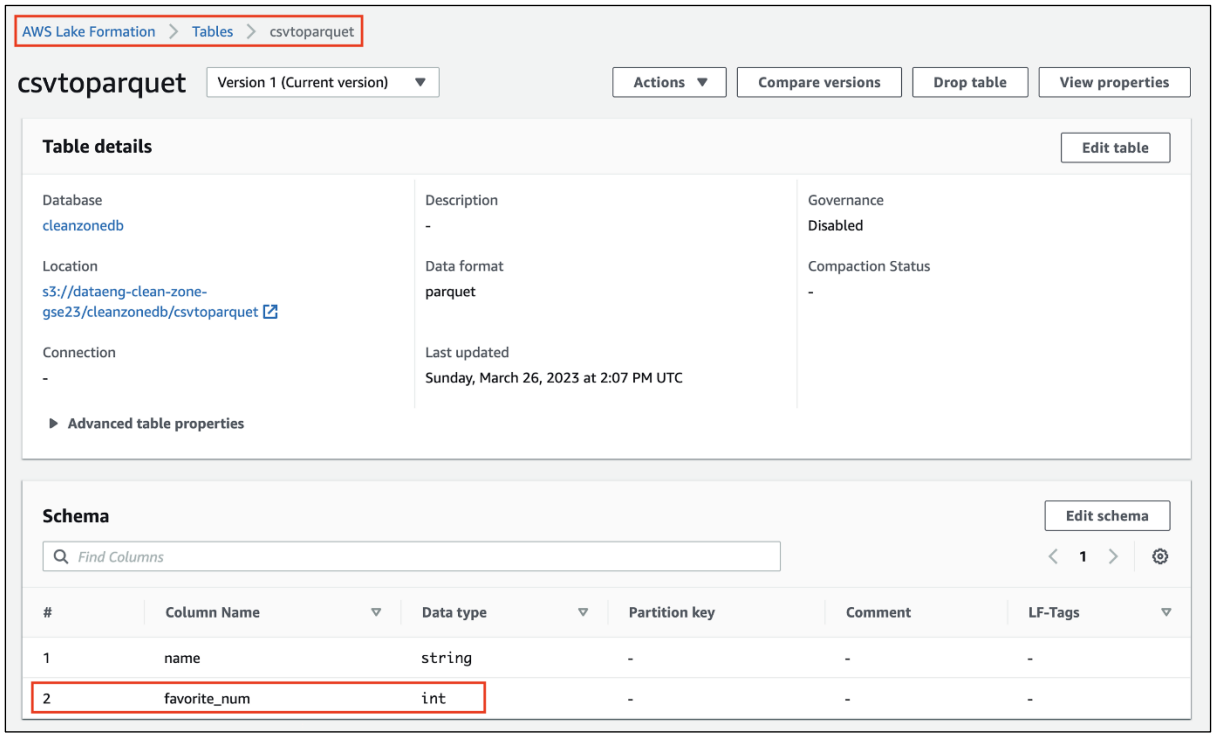
1

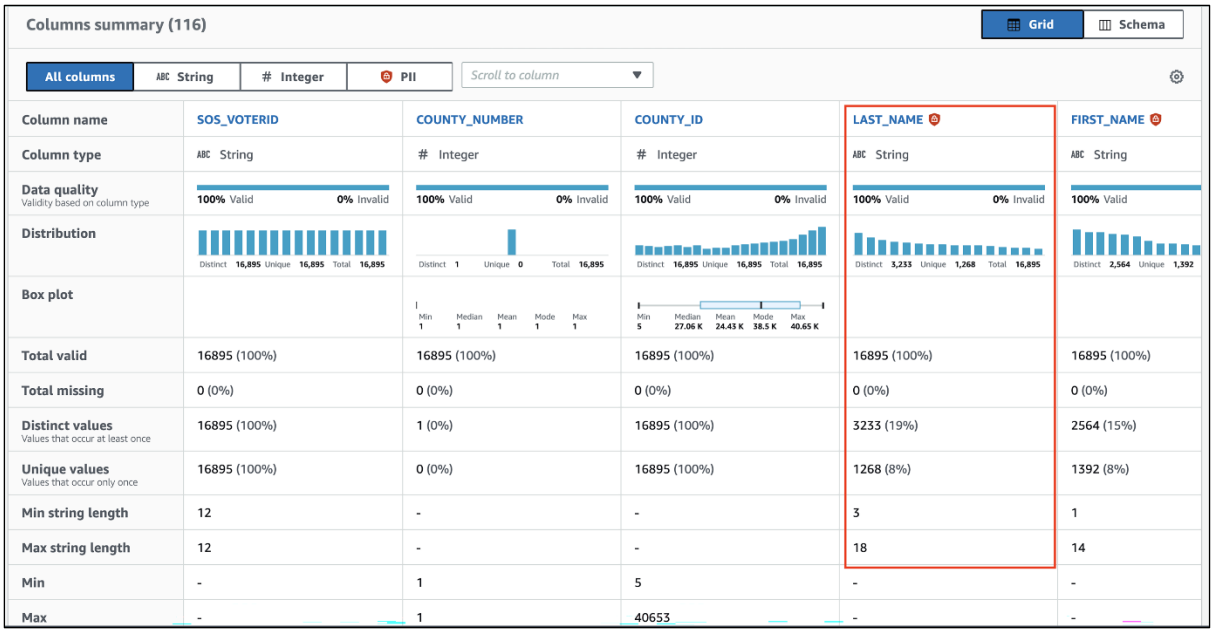
Cancel Add

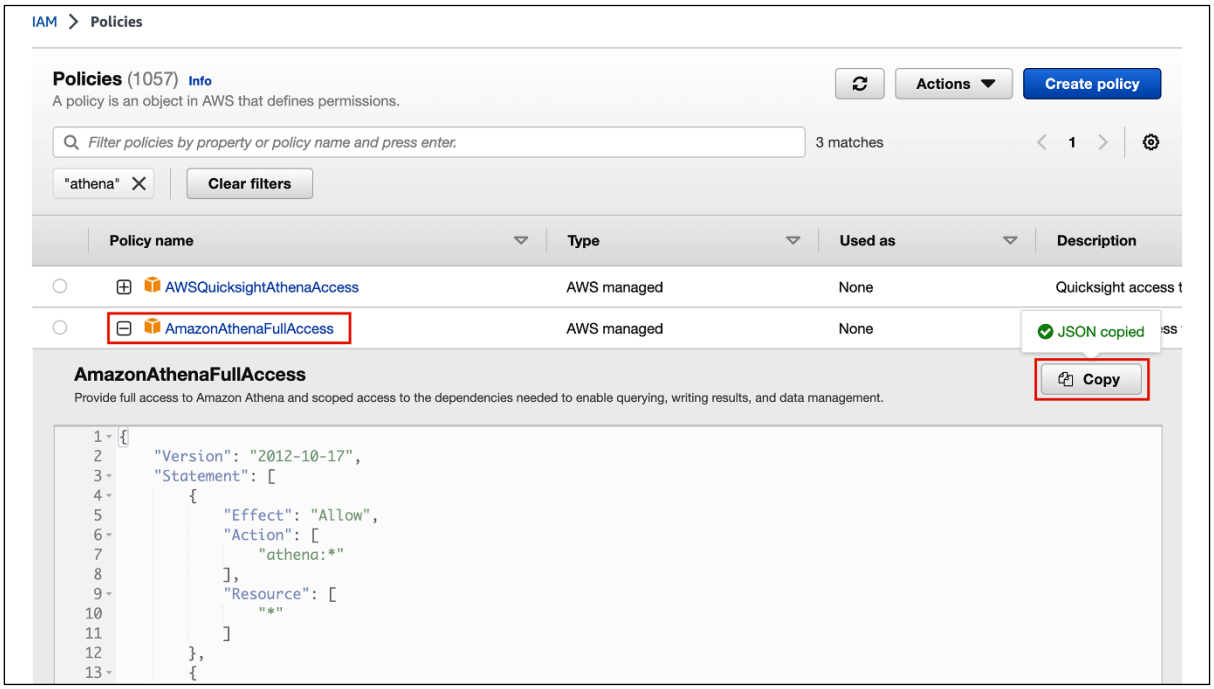

Add trigger
Trigger configuration
S3
aws storage
Bucket
Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.
dataeng-landing-zone-
Event type
Select the events that you want to have trigger the Lambda function. You can optionally set up a 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.
Cancel
Add

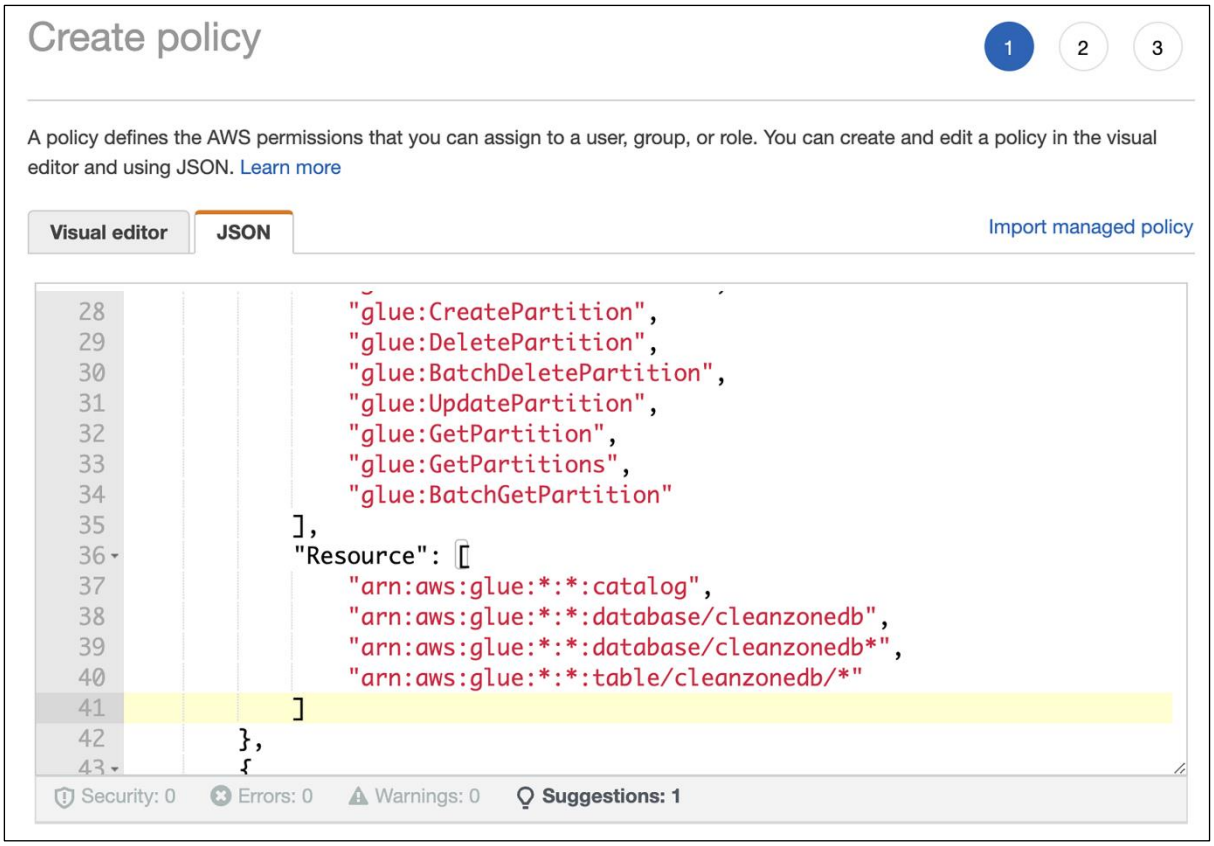
Chapter 4: Data Governance, Security, and Cataloging



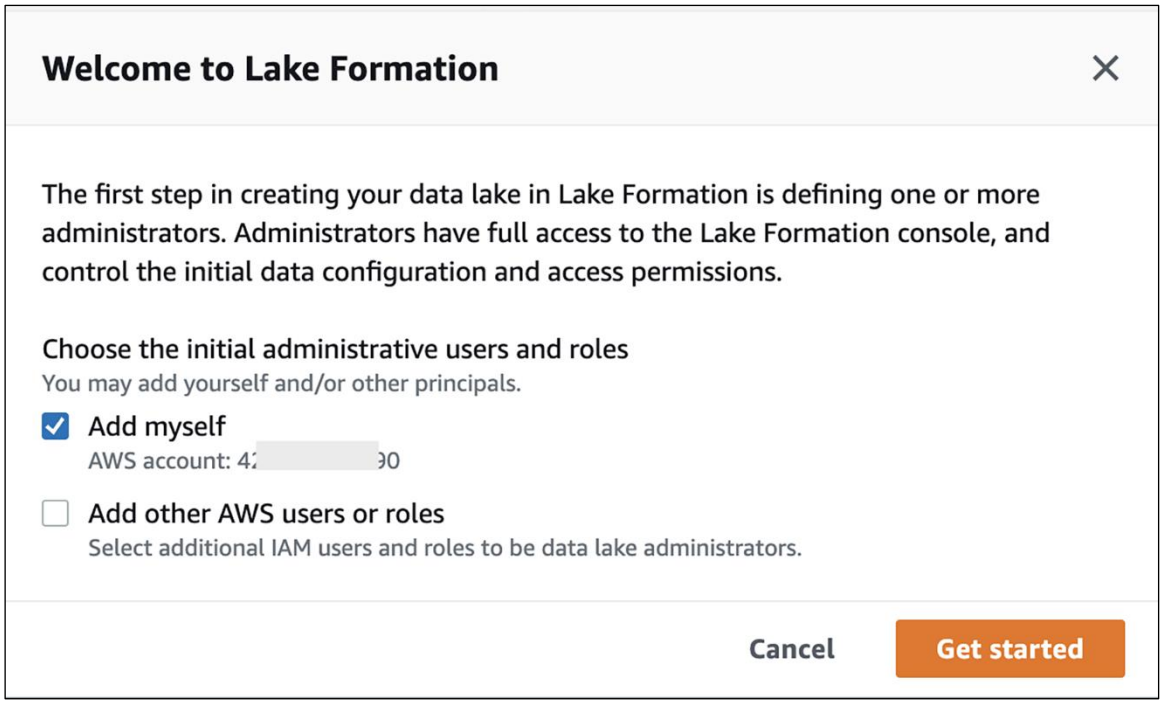


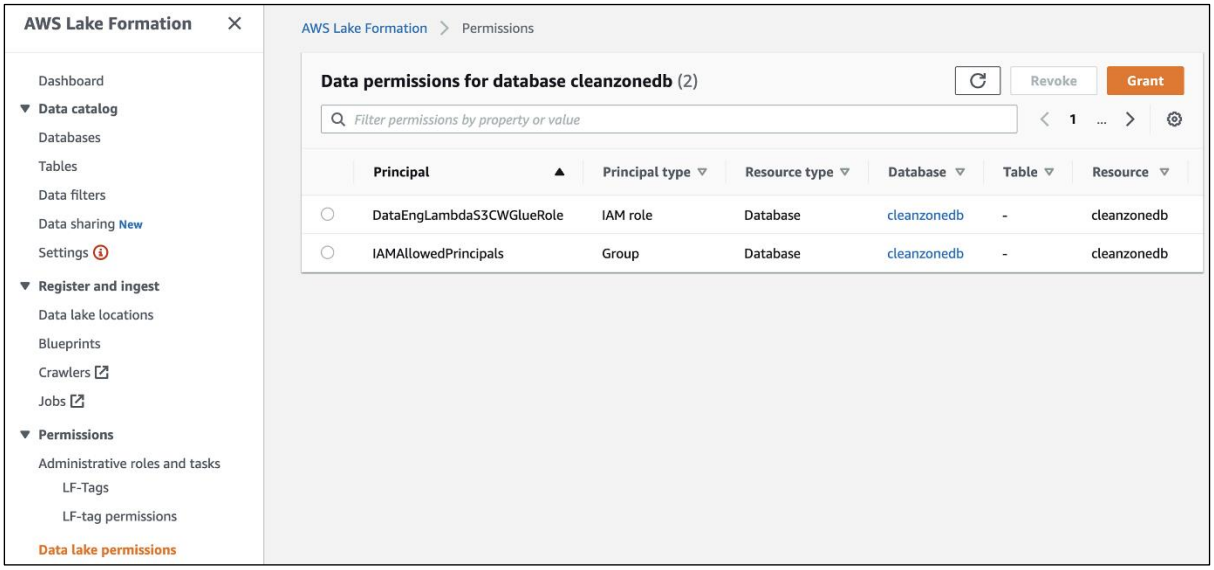

Columns summary (116)
Grid
Schema
All columns
String
Integer
PII
Scroll to column
Column name
SOS_VOTERID
COUNTY_NUMBER
COUNTY_ID
LAST_NAME
FIRST_NAME
Column type
String
Integer
Integer
String
String
Data quality
Validity based on column type
100% Valid 0% Invalid
100% Valid 0% Invalid
100% Valid 0% Invalid
100% Valid 0% Invalid
100% Valid
Distribution
Box plot
Total valid
16895 (100%)
16895 (100%)
16895 (100%)
16895 (100%)
16895 (100%)
Total missing
0 (0%)
0 (0%)
0 (0%)
0 (0%)
0 (0%)
Distinct values
Values that occur at least once
16895 (100%)
1 (0%)
16895 (100%)
3233 (19%)
2564 (15%)
Unique values
Values that occur only once
16895 (100%)
0 (0%)
16895 (100%)
1268 (8%)
1392 (8%)
Min string length
12
-
-
3
1
Max string length
12
-
-
18
14
Min
-
1
5
-
-
Max
-
1
40653
-
-

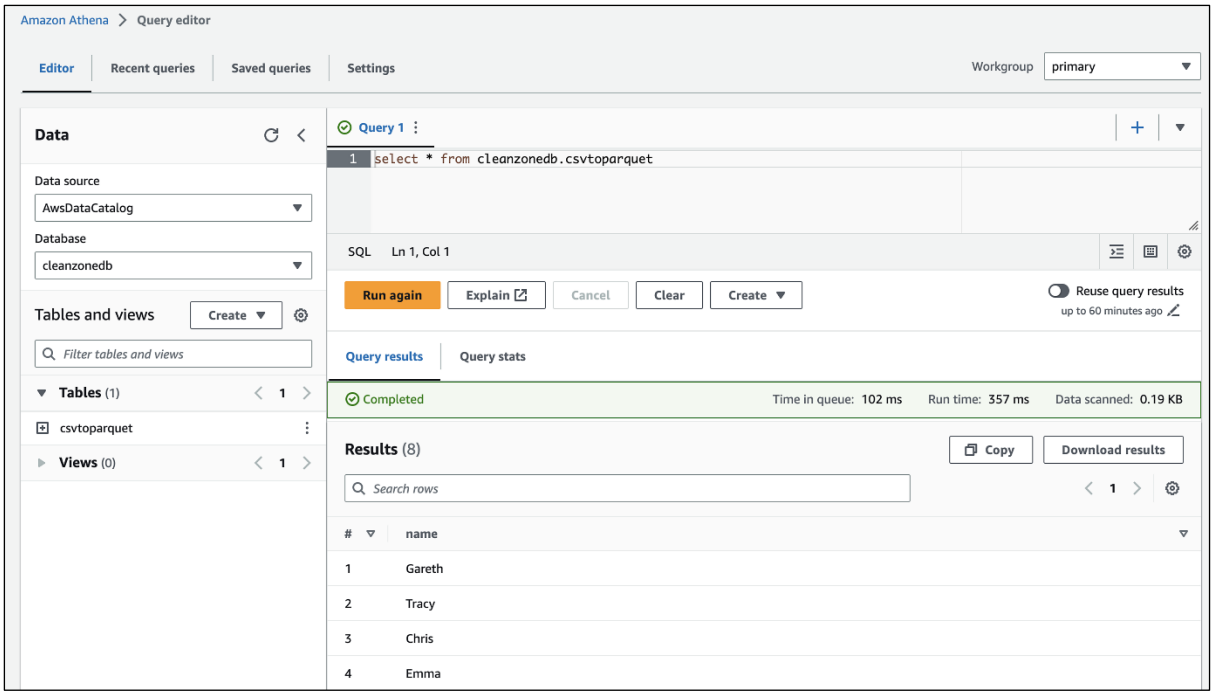

IAM
Policies
Policies (1057) Info
A policy is an object in AWS that defines permissions.
Actions
Create policy
Filter policies by property or policy name and press enter.
3 matches
1
"athena"
Clear filters
Policy name
Type
Used as
Description
AWSQuicksightAthenaAccess
AWS managed
None
Quicksight access t
AmazonAthenaFullAccess
AWS managed
None
JSON copied
Copy
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": [
"*"
]
},
{


Create policy
1
2
3
A policy defines the AWS permissions that you can assign to a user, group, or role. You can create and edit a policy in the visual editor and using JSON. Learn more
Visual editor
JSON
Import managed policy
"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/*"
]
},
{
Security: 0
Errors: 0
Warnings: 0
Suggestions: 1

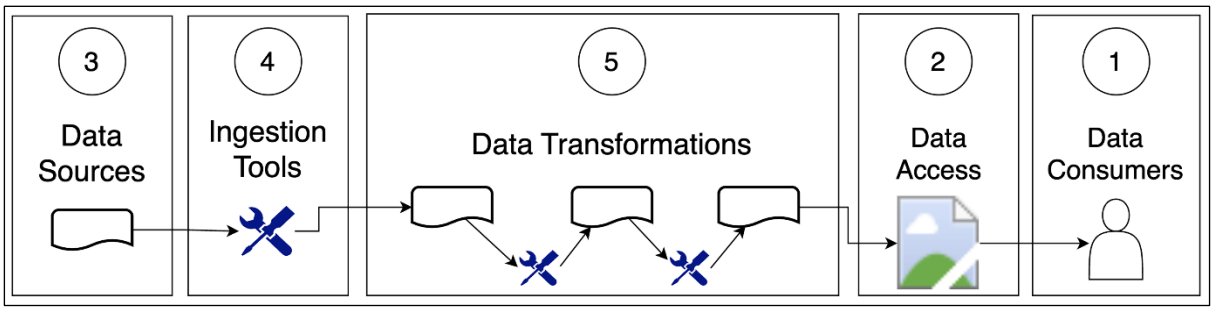

Visual editor
JSON
Import managed policy
"arn:aws:glue:*:*:table/cleanzonedb/*"
]
},
{
"Effect": "Allow",
"Action": [
"s3:GetBucketLocation",
"s3:GetObject",
"s3:ListBucket",
"s3:ListBucketMultipartUploads",
"s3:ListMultipartUploadParts",
"s3:AbortMultipartUpload",
"s3:PutObject"
],
"Resource": [
"arn:aws:s3:::dataeng-clean-zone-gse23/*"
]
},
{
"Effect": "Allow",
"Action": [
"s3:GetBucketLocation",
Security: 0
Errors: 0
Warnings: 0
Suggestions: 1

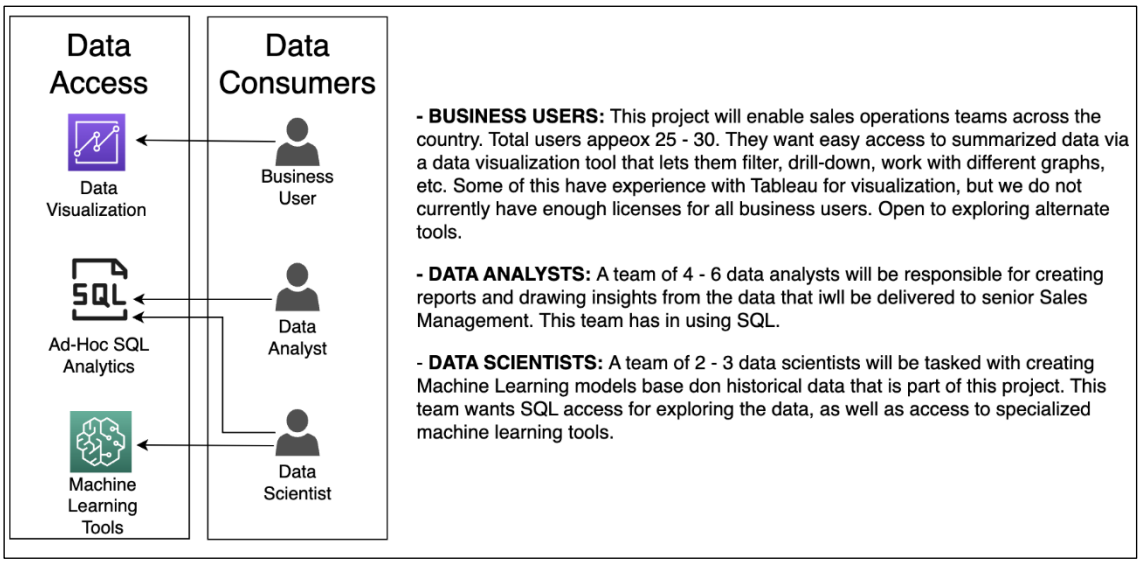

Welcome to Lake Formation
The first step in creating your data lake in Lake Formation is defining one or more administrators. Administrators have full access to the Lake Formation console, and control the initial data configuration and access permissions.
Choose the initial administrative users and roles
You may add yourself and/or other principals.
Add myself
AWS account: 42 90
Add other AWS users or roles
Select additional IAM users and roles to be data lake administrators.
Cancel
Get started


AWS Lake Formation
Dashboard
Data catalog
Databases
Tables
Data filters
Data sharing New
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
LF-Tags
LF-tag permissions
Data lake permissions
AWS Lake Formation > Permissions
Data permissions for database cleanzonedb (2)
Revoke
Grant
Filter permissions by property or value
1
Principal
Principal type
Resource type
Database
Table
Resource
DataEngLambdaS3CWGlueRole
IAM role
Database
cleanzonedb
-
cleanzonedb
IAMAllowedPrincipals
Group
Database
cleanzonedb
-
cleanzonedb


Amazon Athena
Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
primary
Data
Data source
AwsDataCatalog
Database
cleanzonedb
Tables and views
Create
Filter tables and views
Tables (1)
csvtoparquet
Views (0)
Query 1
select * from cleanzonedb.csvtoparquet
SQL
Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 102 ms
Run time: 357 ms
Data scanned: 0.19 KB
Results (8)
Copy
Download results
Search rows
name
1 Gareth
2 Tracy
3 Chris
4 Emma

Chapter 5: Architecting Data Engineering Pipelines

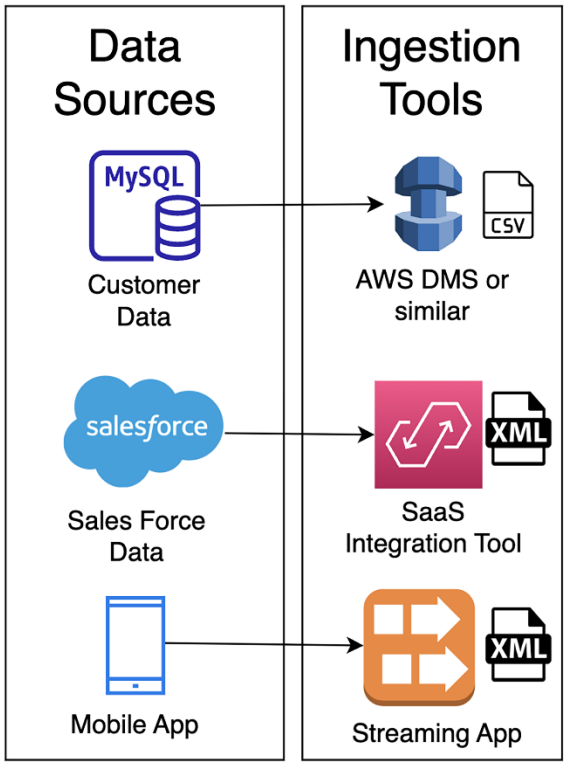




- BUSINESS USERS: This project will enable sales operations teams across the country. Total users appeox 25 - 30. They want easy access to summarized data via a data visualization tool that lets them filter, drill-down, work with different graphs, etc. Some of this have experience with Tableau for visualization,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

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

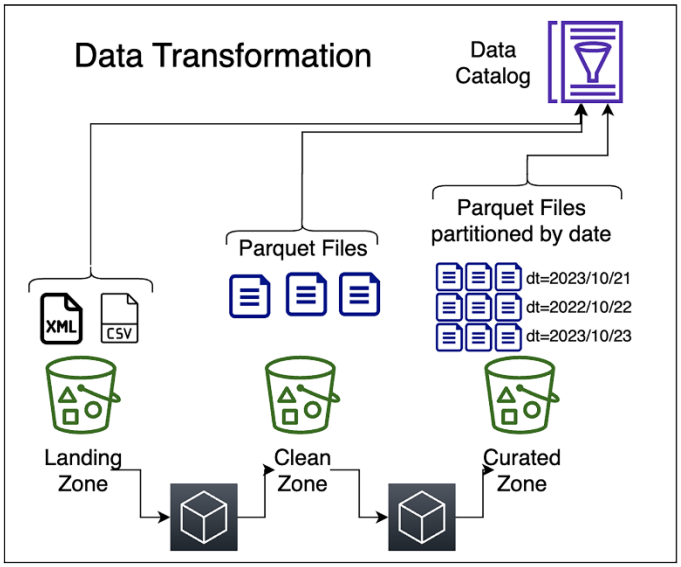
- **DATA SCIENTISTS:** A team of 2 - 3 data scientists will be tasked with creating Machine Learning models base don historical data that is part of this project. This team wants SQL access for exploring the data, as well as access to specialized machine learning tools.



- **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.

- **SALESFORCE DATA:** Company has Salesforce 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).



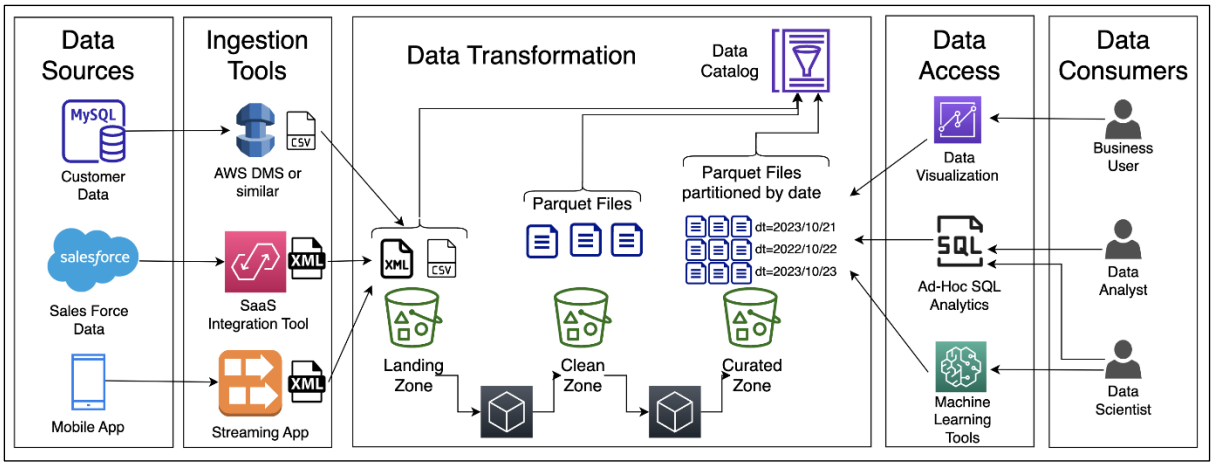
Data is ingested into the landing zone in raw format (CSV and XML)

- 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, teh 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.

- **SALESFORCE DATA:** Company has Salesforce 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, teh data is catalogged and meta-data is written to the data catalog.

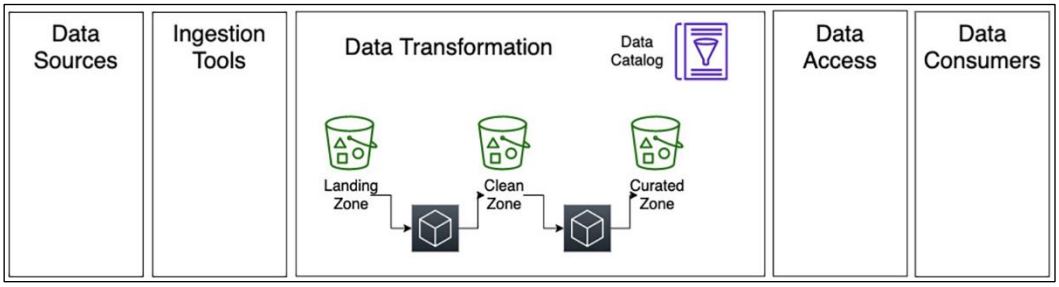
- The team does have previous experience with creating Spark ETL jobs written in PySpark.

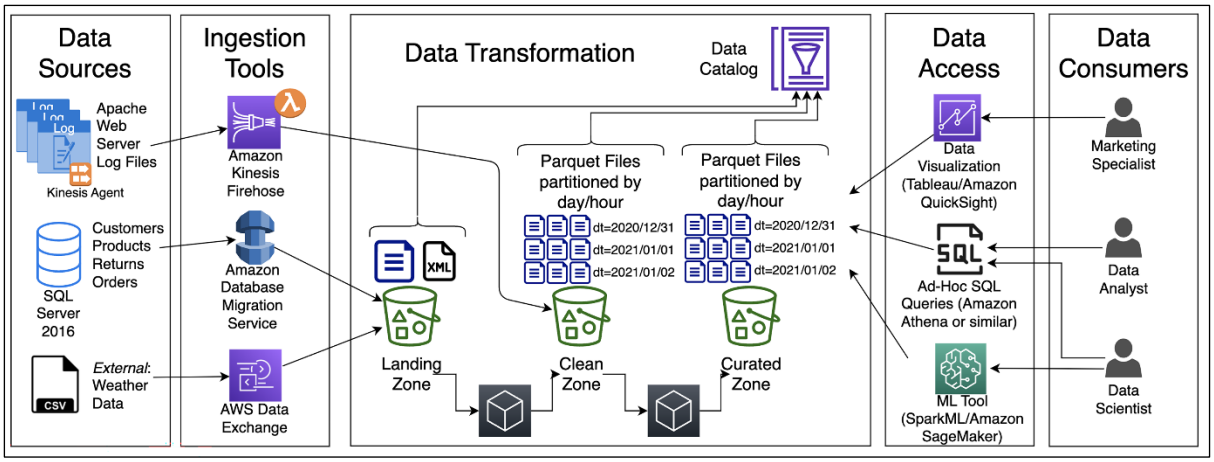
Data Consumers

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

- **DATA SCIENTISTS:** A team of 2 - 3 data scientists will be tasked with creating Machine Learning models base don 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 appeox 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 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.

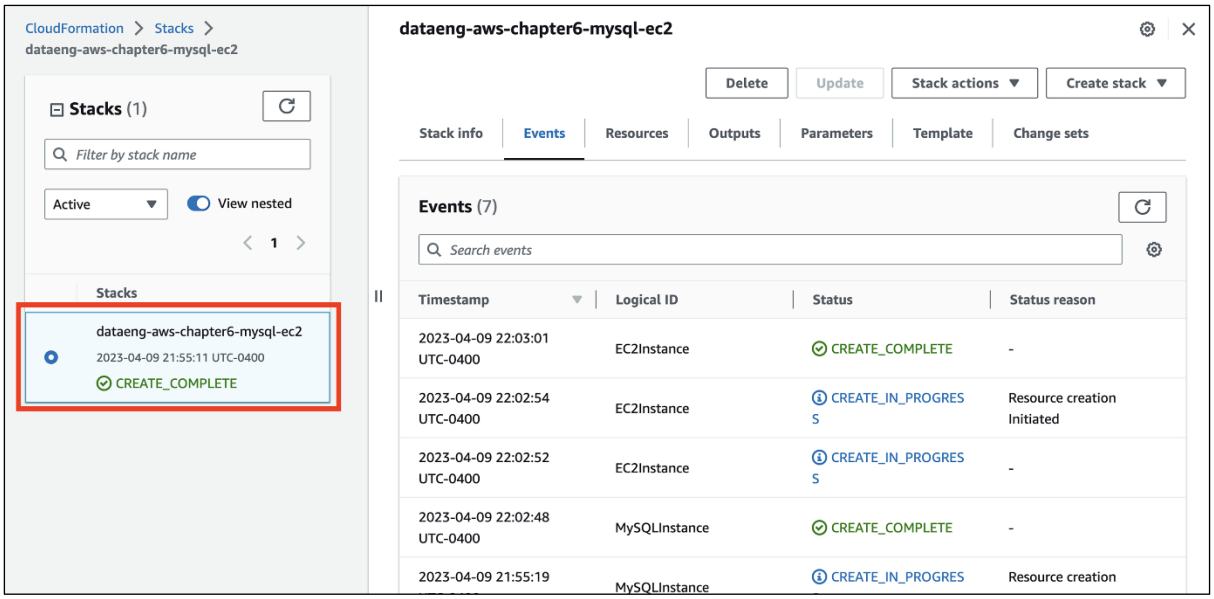
- **Data Scientists:** Need ad-hoc SQL access to databases, weather and web server log files. They currently use SparkML on-premises, but open to new cloud based tools that may make speed up delivery and collaboration for their machine learning products.

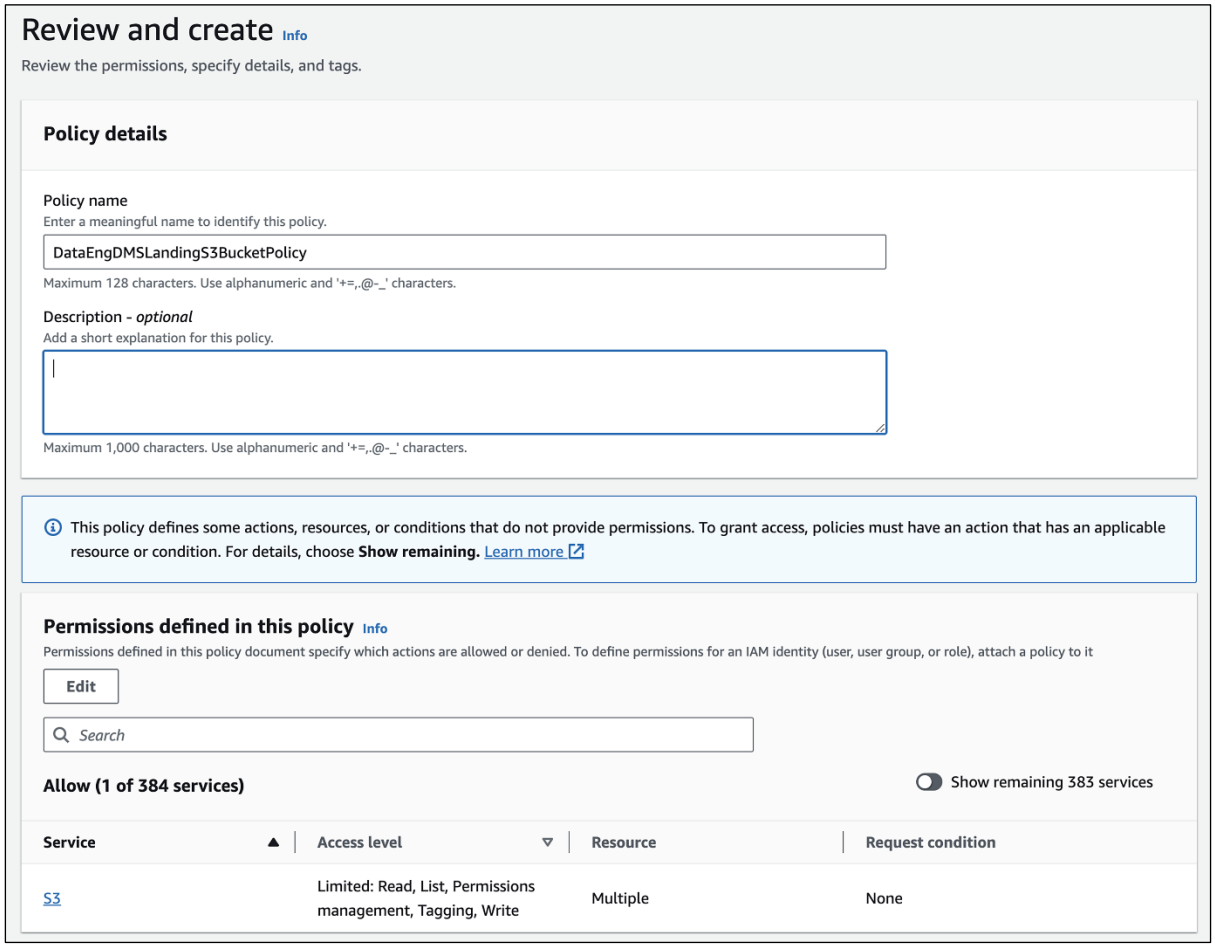
Chapter 6: Ingesting Batch and Streaming Data

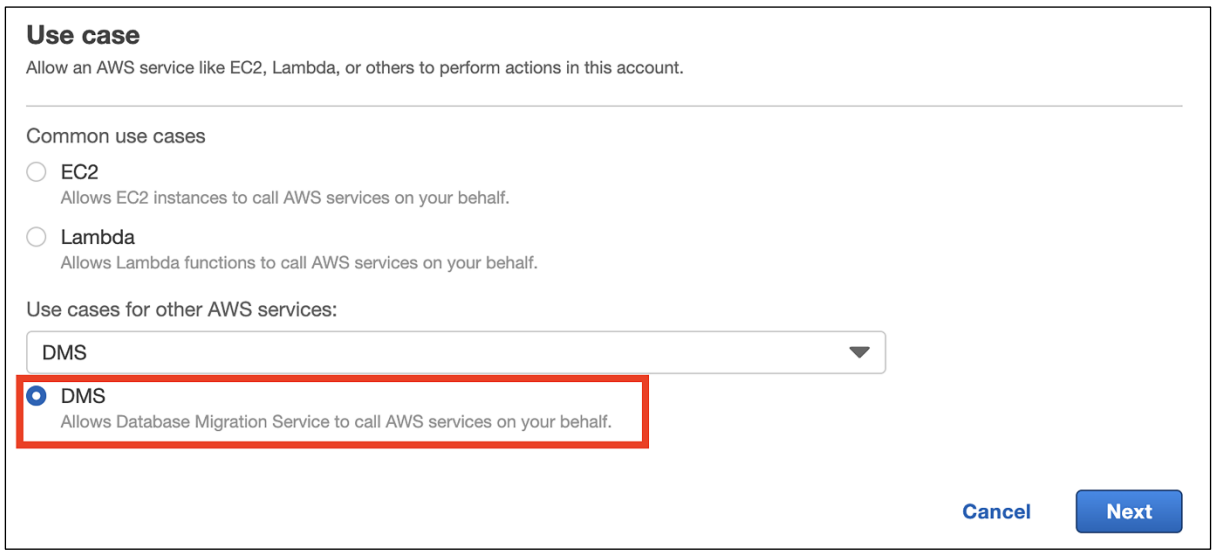
Food_Code	Display_Name	Portion_Display_Name	Total Calo
71411000	Potato skin with cheese & bacon	order (10 halves)	1667.4
24301010	Roasted duck	duck half	1283.52
21103120	Breaded fried steak (eat lean & fat)	large steak	1069.2
28141010	Fried chicken frozen meal	large meal (16 oz)	1024.92
27347100	Chicken 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

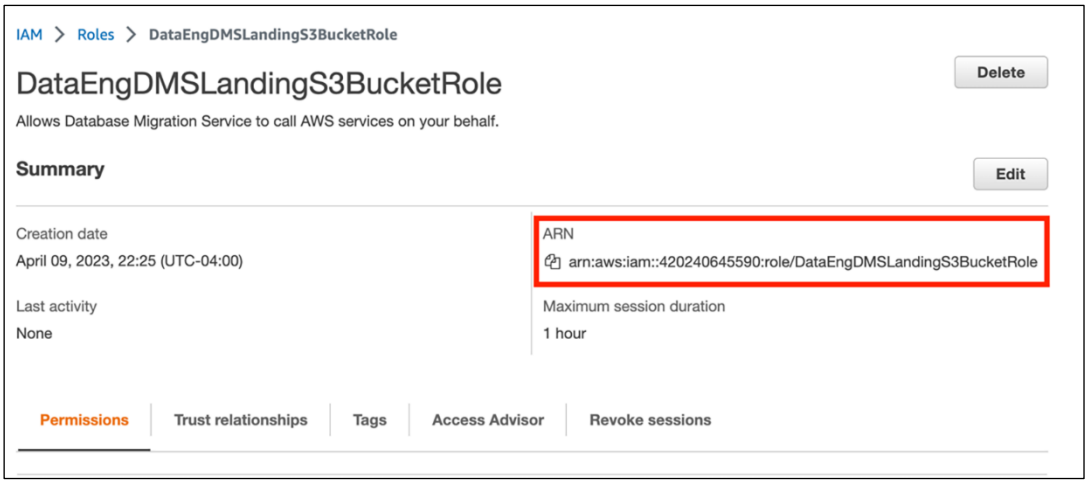
53 lines (45 sloc) | 1.67 KB Raw Blame

```
1   ---
2   AWSTemplateFormatVersion: 2010-09-09
3   Description: Chapter 6 - Data Engineering with AWS
4   Parameters:
5     DBPassword:
6       Type: String
7       NoEcho: true
8       Description: The database admin account password
9       MinLength: 8
10      AllowedPattern: ^[a-zA-Z0-9]*$
11      ConstraintDescription: Password must contain only alphanumeric characters.
12    LatestAmiId:
13      Type: 'AWS::SSM::Parameter::Value<AWS::EC2::Image::Id>'
14      Default: '/aws/service/ami-amazon-linux-latest/al2023-ami-kernel-6.1-x86_64'
15  Resources:
16    MySQLInstance:
```




Review and create Info
Review the permissions, specify details, and tags.
Policy details
Policy name
Enter a meaningful name to identify this policy.
DataEngDMSLandingS3BucketPolicy
Maximum 128 characters. Use alphanumeric and '+=,.@-_' characters.
Description - optional
Add a short explanation for this policy.
Maximum 1,000 characters. Use alphanumeric and '+=,.@-_' characters.
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
Permissions defined in this policy Info
Permissions defined in this policy document specify which actions are allowed or denied. To define permissions for an IAM identity (user, user group, or role), attach a policy to it
Edit
Search
Allow (1 of 384 services)
Show remaining 383 services
Service
Access level
Resource
Request condition
S3
Limited: Read, List, Permissions management, Tagging, Write
Multiple
None


Use case
Allow an AWS service like EC2, Lambda, or others to perform actions in this account.
Common use cases
EC2
Allows EC2 instances to call AWS services on your behalf.
Lambda
Allows Lambda functions to call AWS services on your behalf.
Use cases for other AWS services:
DMS
DMS
Allows Database Migration Service to call AWS services on your behalf.
Cancel
Next



Endpoint configuration

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

s3-landing-zone-sakila-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 6:role/DataEngDMSLandingS3BucketRole

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

dataeng-landing-zone

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	Remove

Add new setting

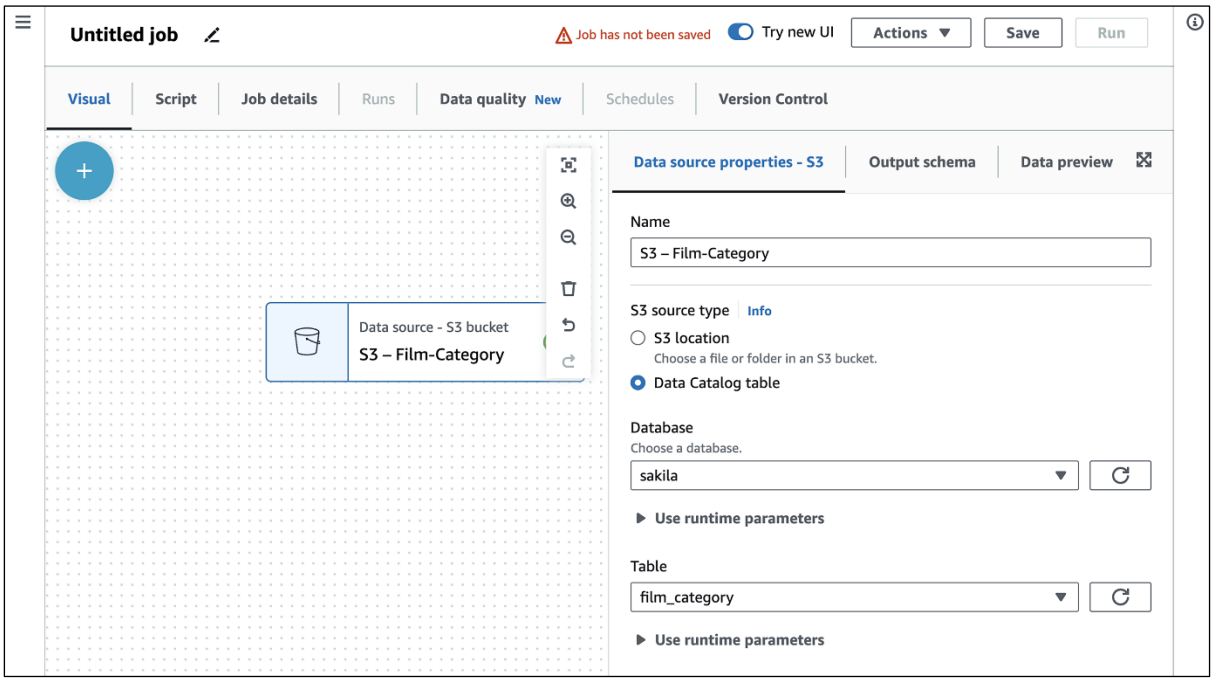
[ ] Use endpoint connection attributes

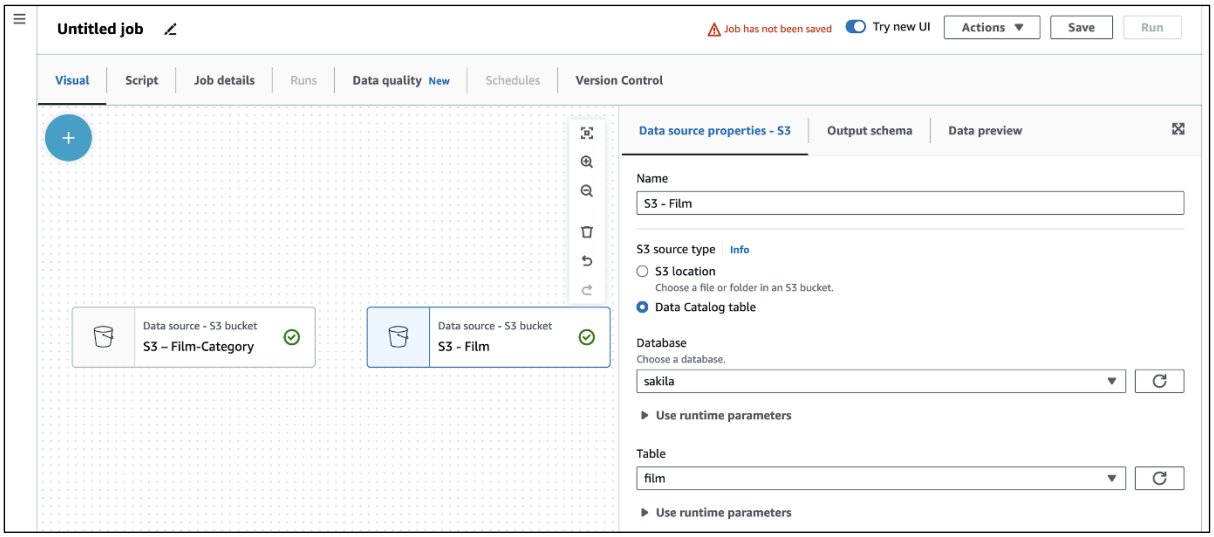
Chapter 7: Transforming Data to Optimize for Analytics

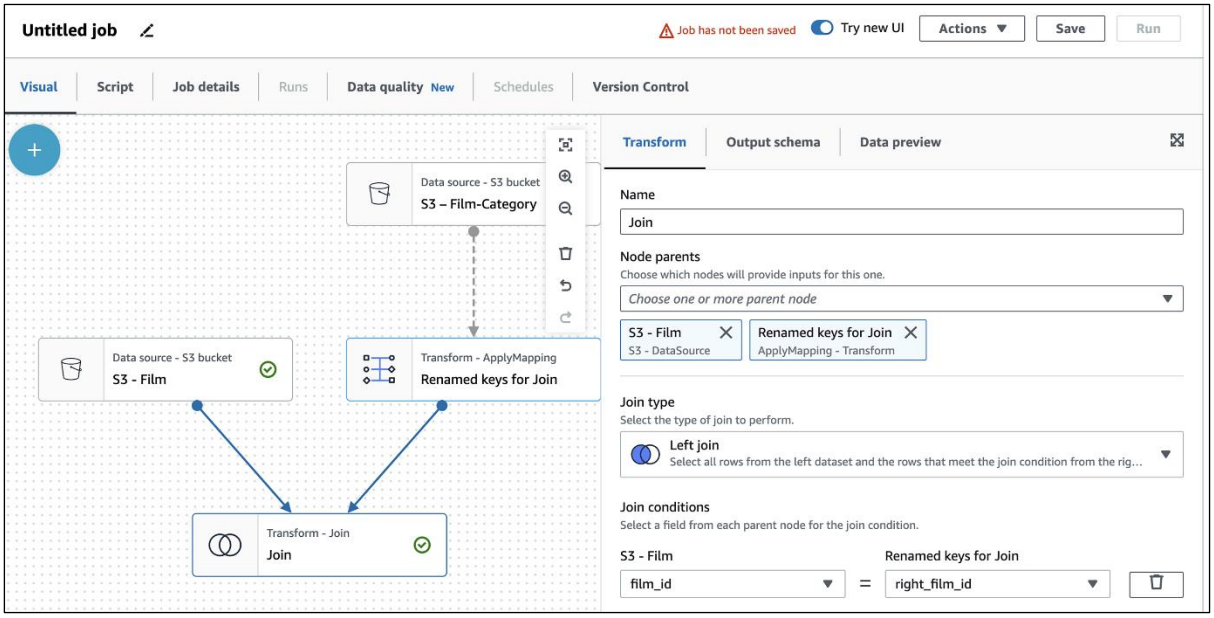
Customer_ID	Last_Name	First_Name	Address_Street	Address_City	Address_State	Phone_Number	Sales_Person_ID
1	Smith	Jonathan	123 Main Street	Springville	MA	555-943-1987	2
2	Mendez	Bruno	5449 South West Street	Jersey	PA	555-615-1609	3
3	Sachdeva	Viyoma	94 Midland Avenue	Oxford	NJ	555-664-0464	1

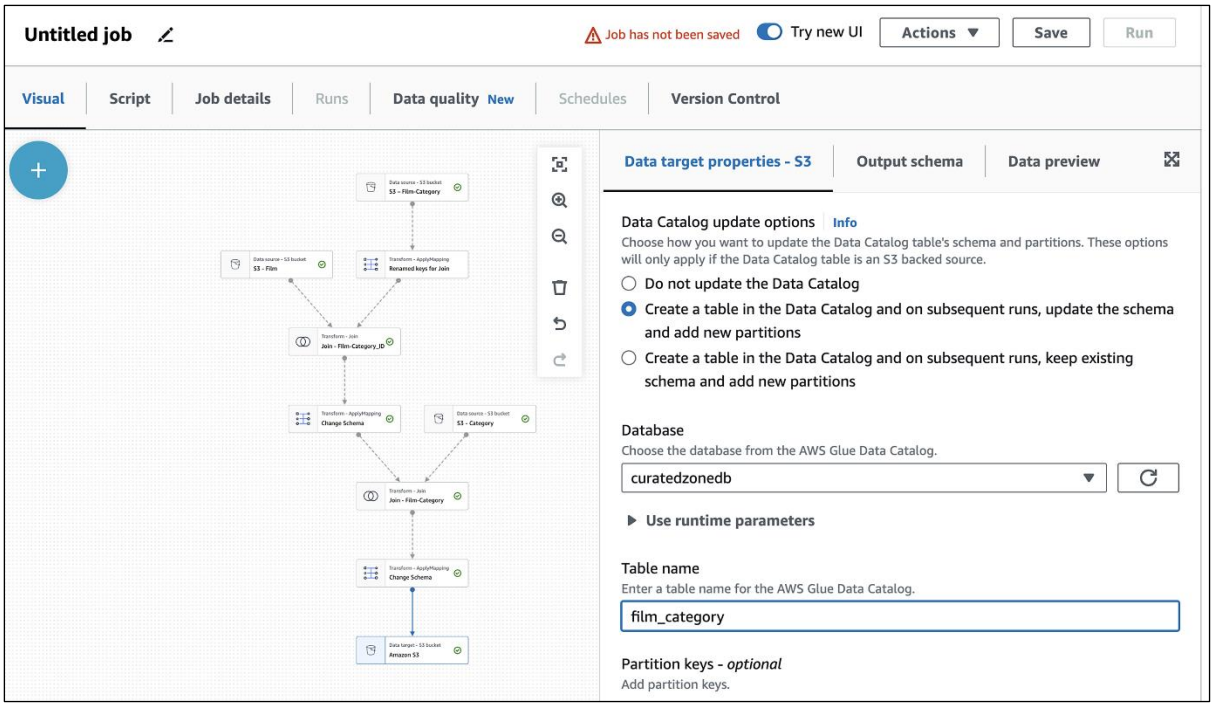
Sales_Person_ID	Last_Name	First_Name	Territory_Code
1	Taylor	Chris	95
2	Williams	Carmen	42
3	Kelly	Michael	23

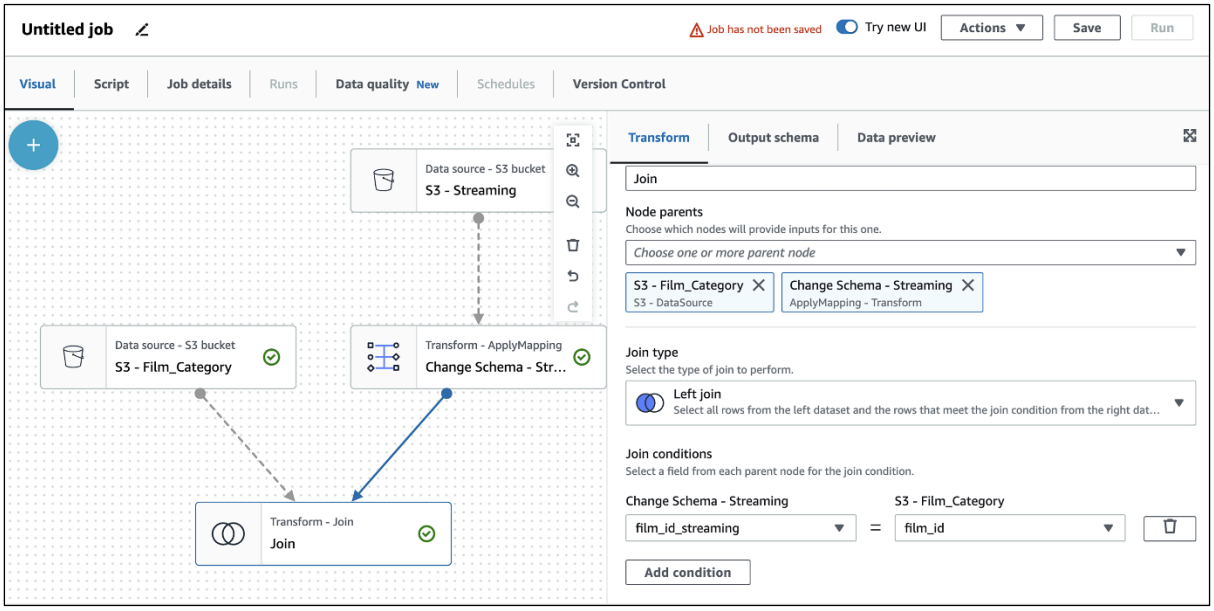
Customer_ID	Last_Name	First_Name	Address_Street	Address_City	Address_State	Phone_Number	Sales_Person_Last	Sales_Person_First
1	Smith	Jonathan	123 Main Street	Springville	MA	555-943-1987	Williams	Carmen
2	Mendez	Bruno	5449 South West Street	Jersey	PA	555-615-1609	Kelly	Michael
3	Sachdeva	Viyoma	94 Midland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

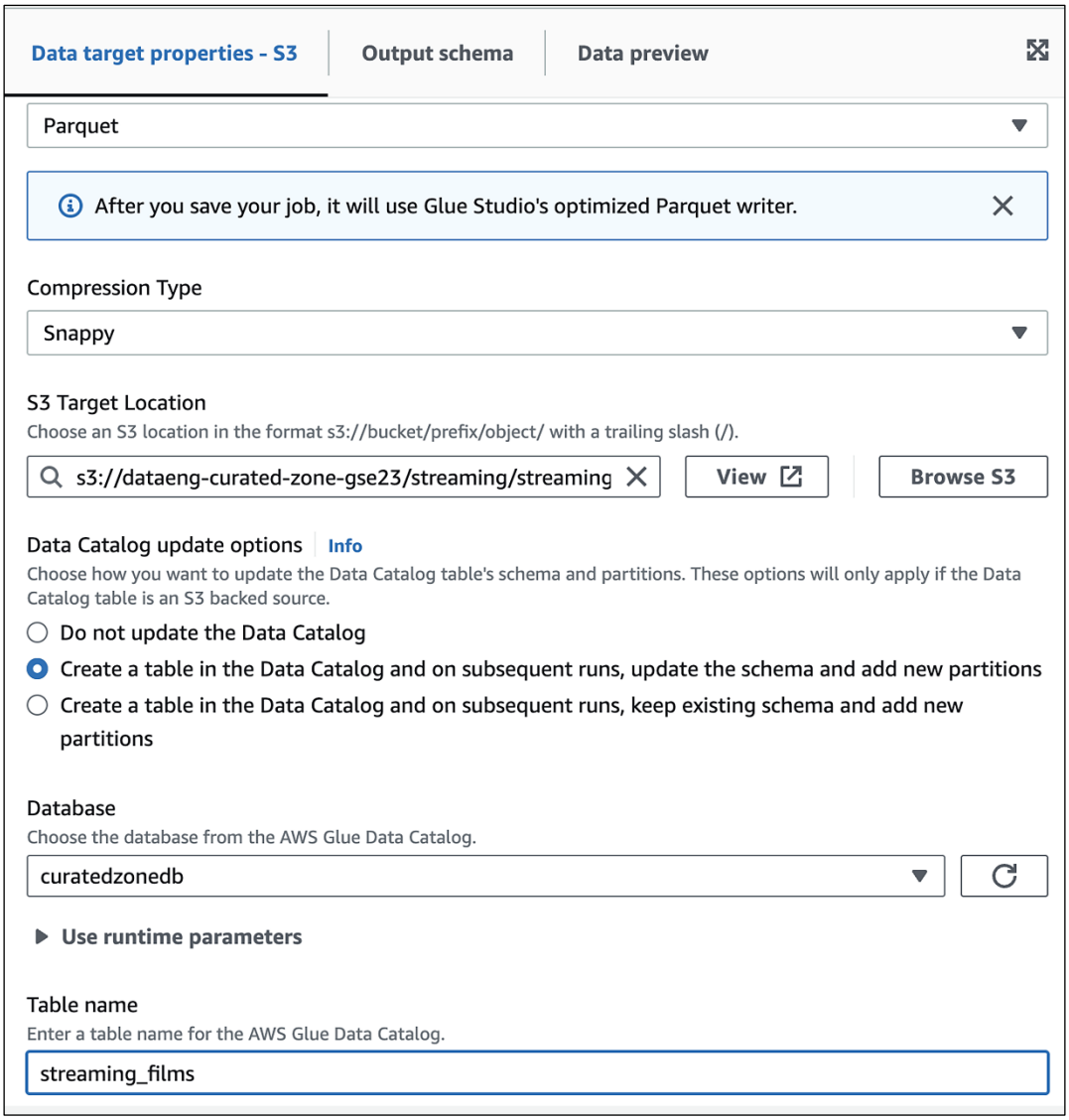



Untitled job
Job has not been saved
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Data source properties - S3
Output schema
Data preview
Name
S3 - Film
S3 source type Info
S3 location
Choose a file or folder in an S3 bucket.
Data Catalog table
Database
Choose a database.
sakila
Use runtime parameters
Table
film
Use runtime parameters
Data source - S3 bucket
S3 – Film-Category
Data source - S3 bucket
S3 - Film

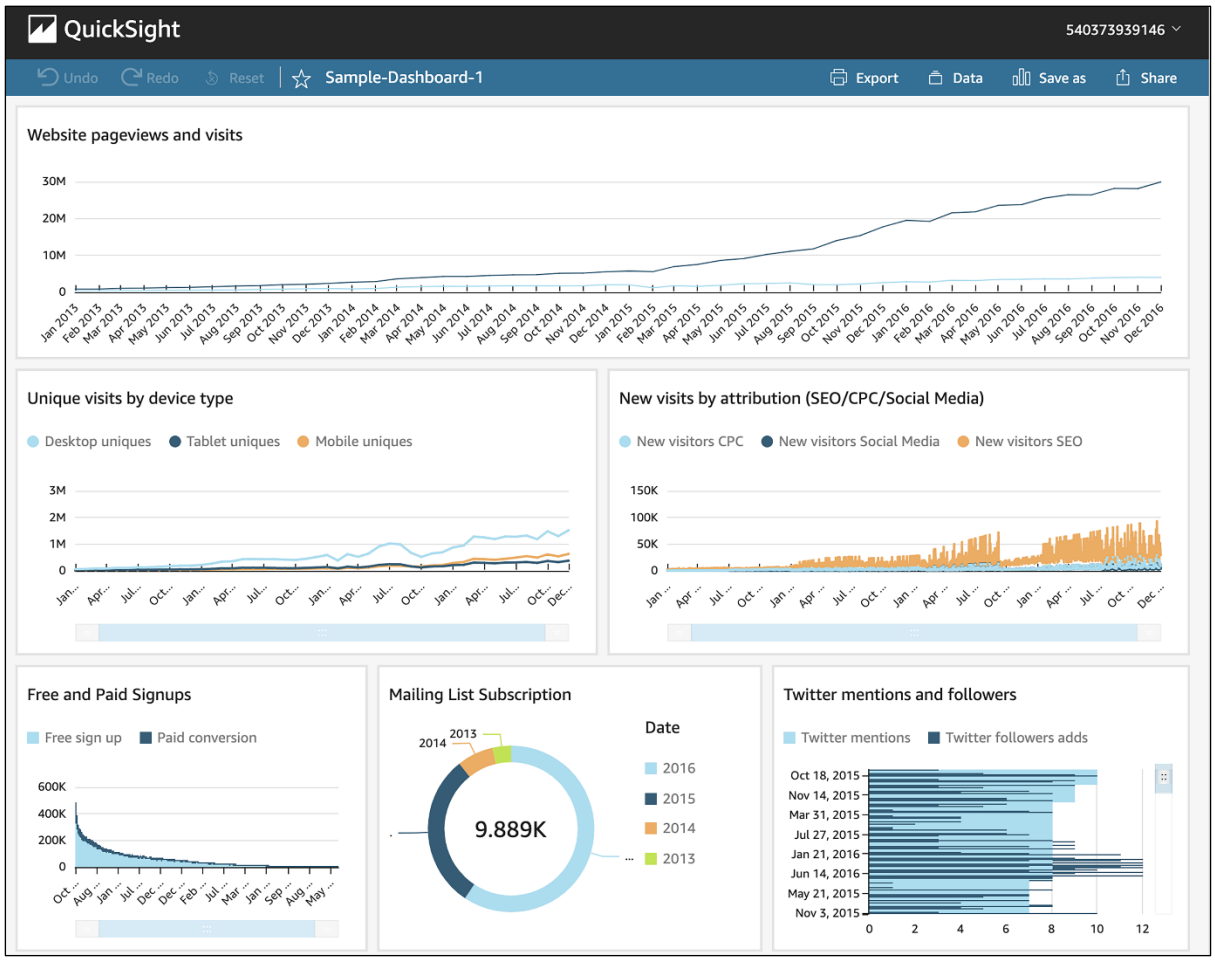

Untitled job
Job has not been saved
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Transform
Output schema
Data preview
Data source - S3 bucket
S3 – Film-Category
Data source - S3 bucket
S3 - Film
Transform - ApplyMapping
Renamed keys for Join
Transform - Join
Join
Name
Join
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
S3 - Film
S3 - DataSource
Renamed keys for Join
ApplyMapping - Transform
Join type
Select the type of join to perform.
Left join
Select all rows from the left dataset and the rows that meet the join condition from the rig...
Join conditions
Select a field from each parent node for the join condition.
S3 - Film
Renamed keys for Join
film_id
=
right_film_id

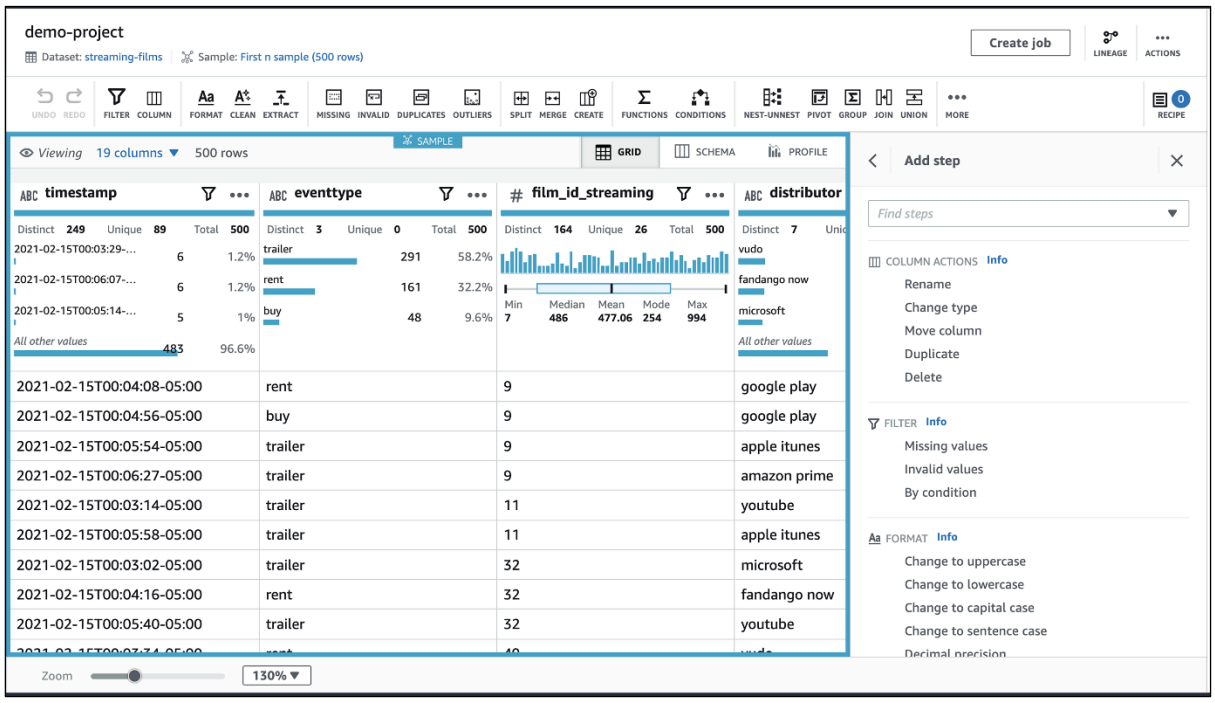

Untitled job
Job has not been saved
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Data target properties - S3
Output schema
Data preview
Data Catalog update options Info
Choose how you want to update the Data Catalog table's schema and partitions. 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
Use runtime parameters
Table name
Enter a table name for the AWS Glue Data Catalog.
film_category
Partition keys - optional
Add partition keys.

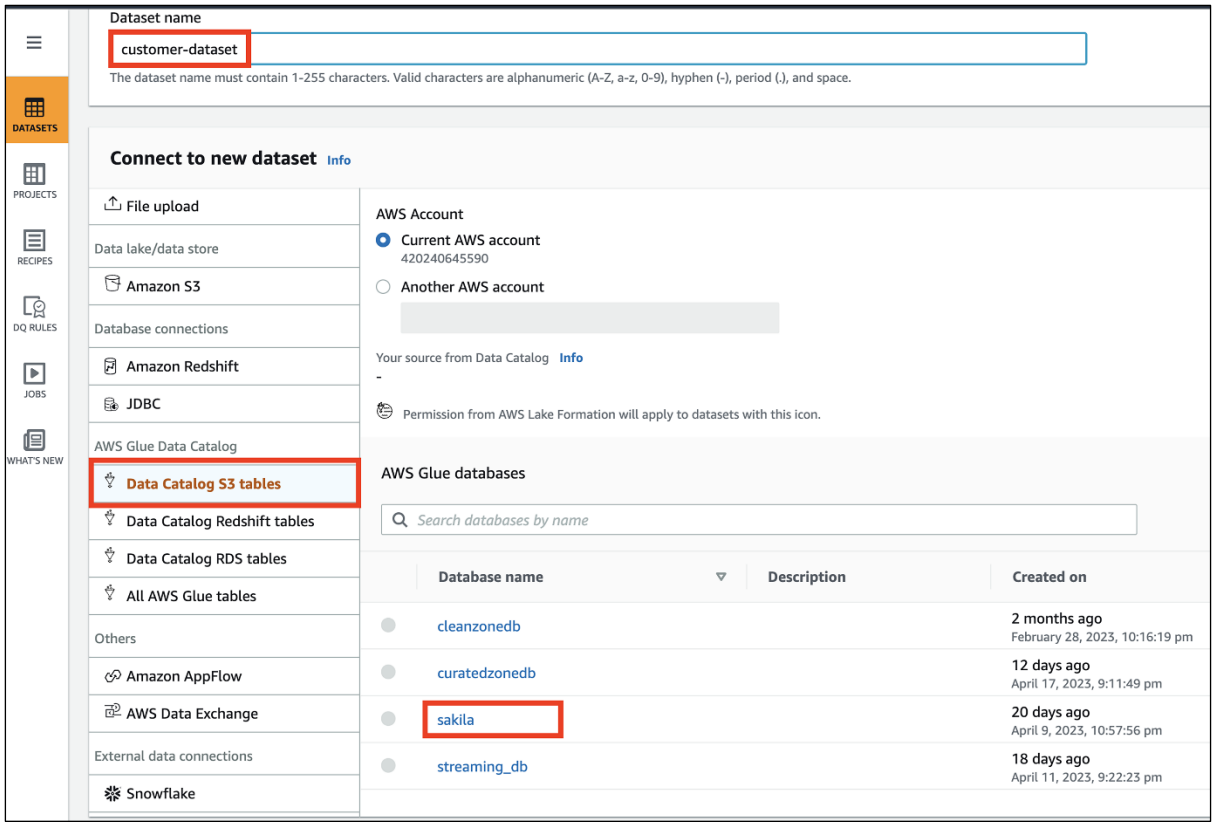

Untitled job
Job has not been saved
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Data source - S3 bucket
S3 - Streaming
Data source - S3 bucket
S3 - Film_Category
Transform - ApplyMapping
Change Schema - Str...
Transform - Join
Join
Transform
Output schema
Data preview
Join
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
S3 - Film_Category
S3 - DataSource
Change Schema - Streaming
ApplyMapping - Transform
Join type
Select the type of join to perform.
Left join
Select all rows from the left dataset and the rows that meet the join condition from the right dat...
Join conditions
Select a field from each parent node for the join condition.
Change Schema - Streaming
S3 - Film_Category
film_id_streaming
=
film_id
Add condition

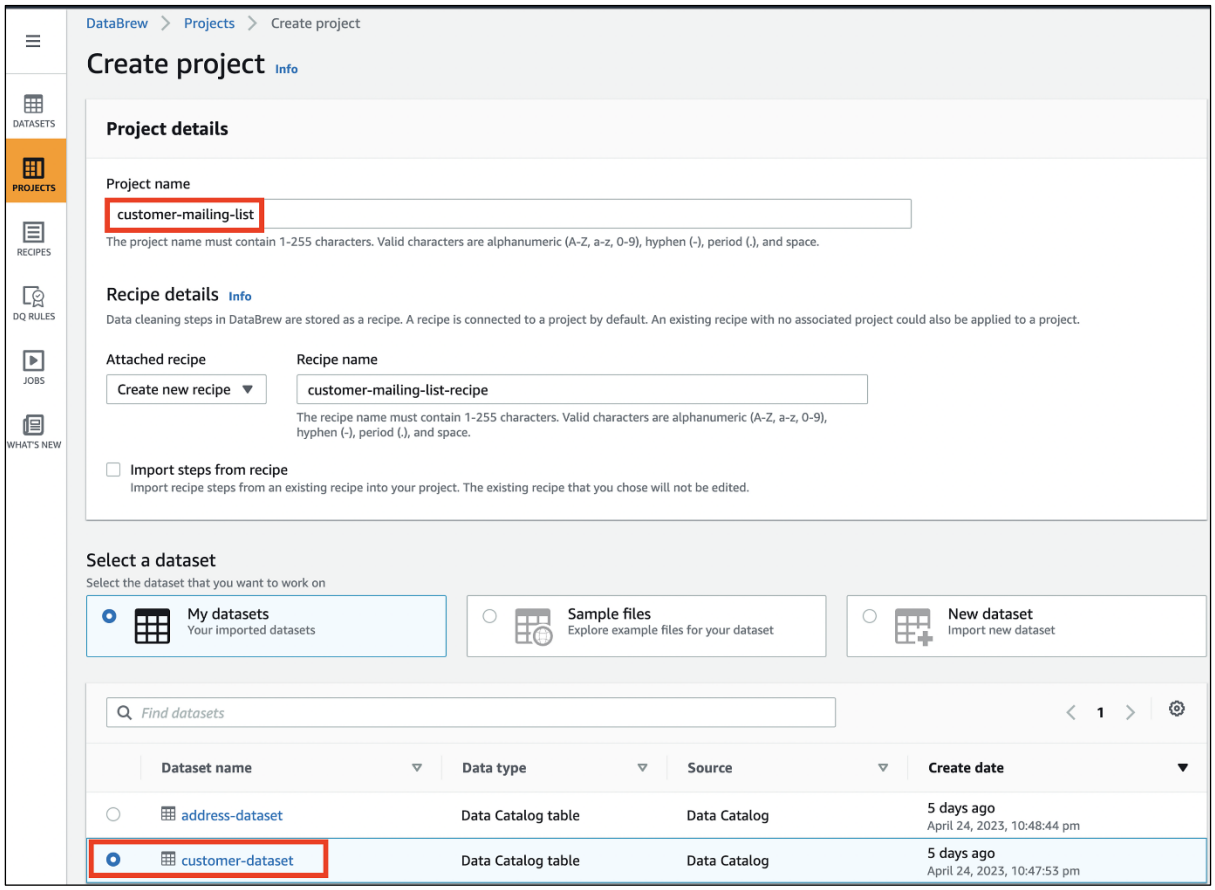

Data target properties - S3
Output schema
Data preview
Parquet
After you save your job, it will use Glue Studio's optimized Parquet writer.
Compression Type
Snappy
S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://dataeng-curated-zone-gse23/streaming/streaming
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
Use runtime parameters
Table name
Enter a table name for the AWS Glue Data Catalog.
streaming_films

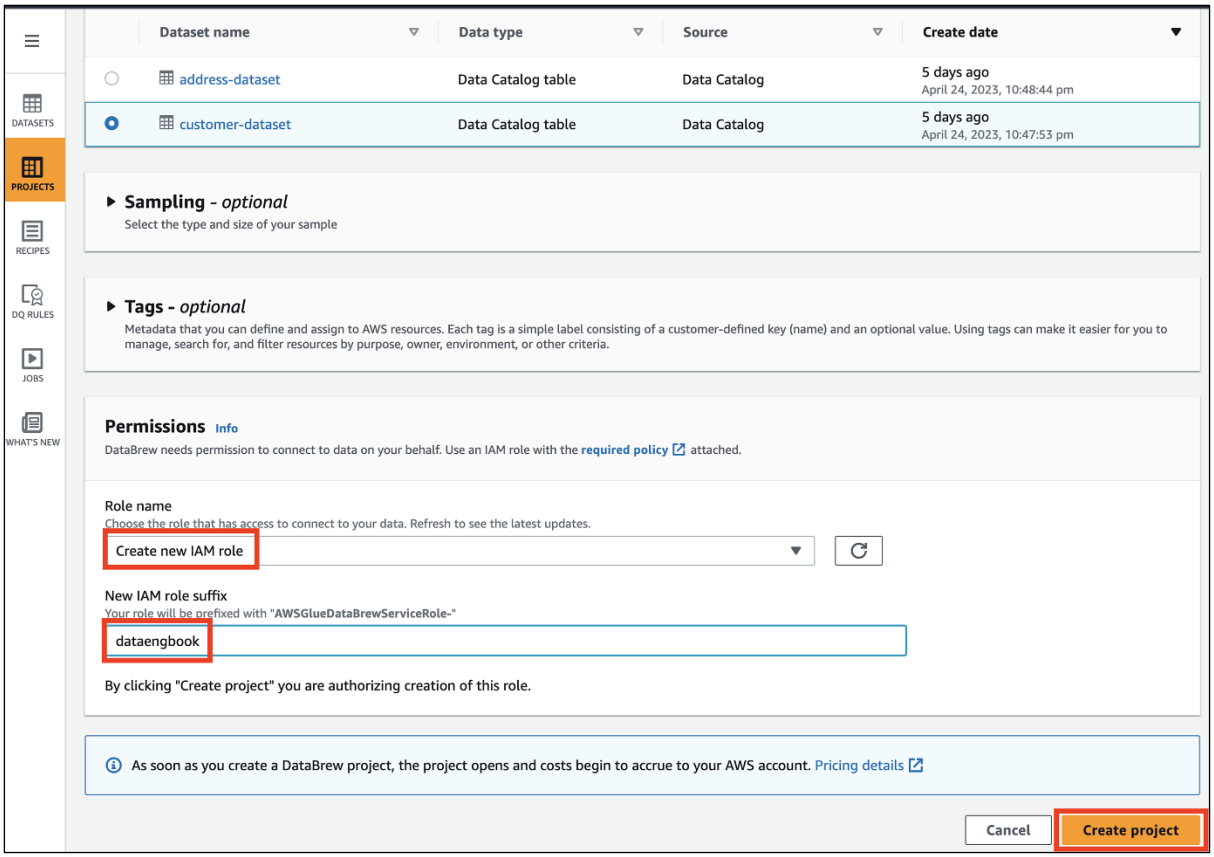
Chapter 8: Identifying and Enabling Data Consumers

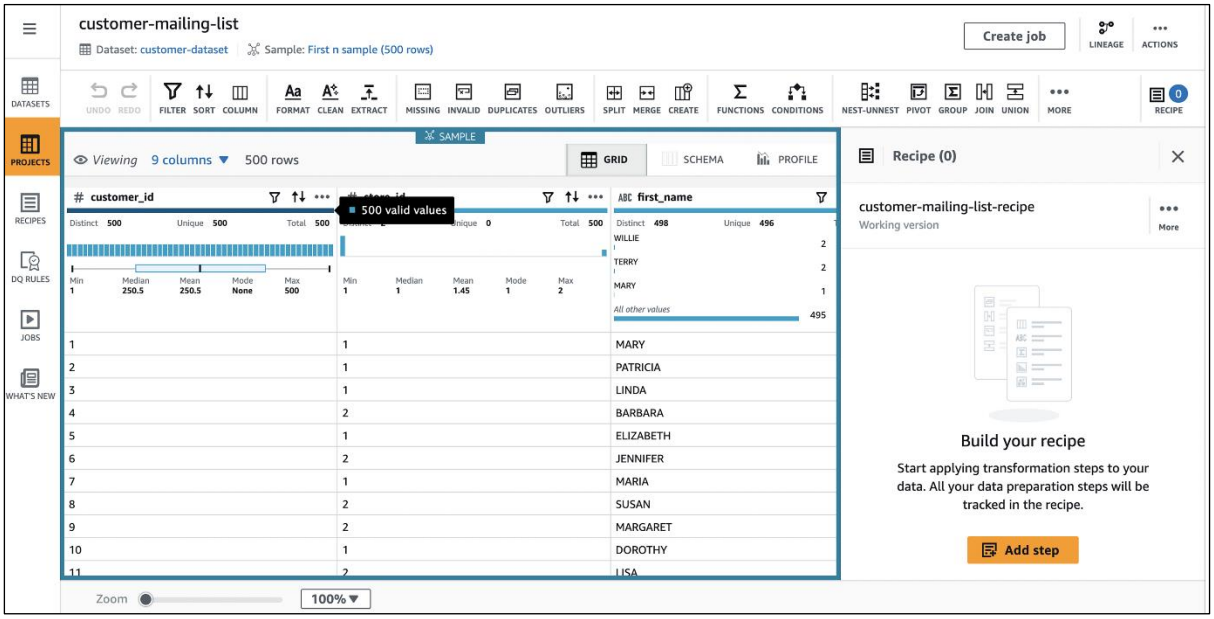


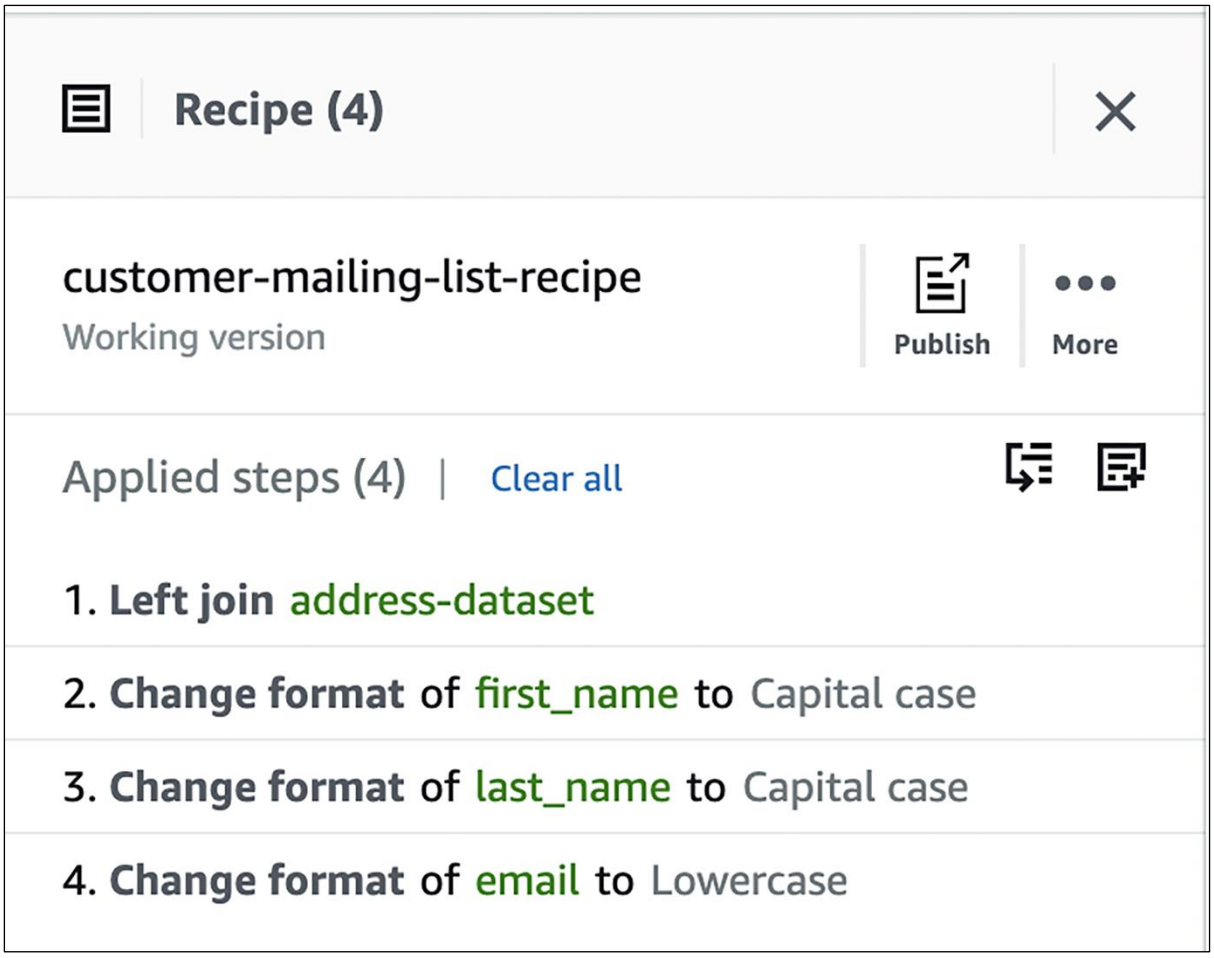


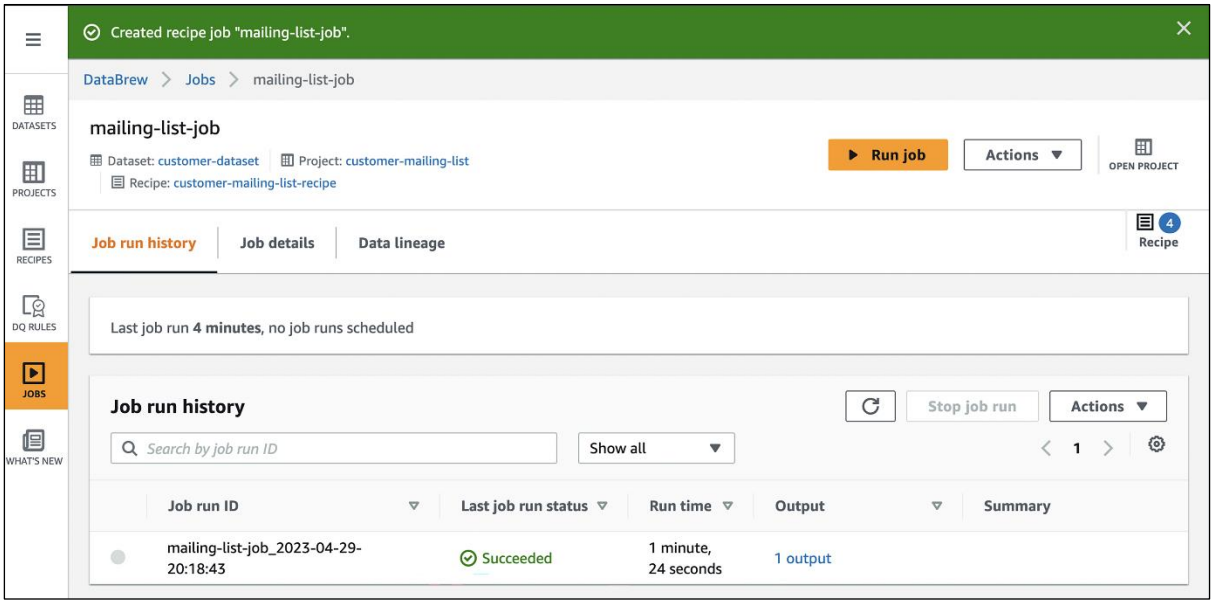

Dataset name
customer-dataset
The dataset name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
Connect to new dataset Info
File upload
Data lake/data store
Amazon S3
Database connections
Amazon Redshift
JDBC
AWS Glue Data Catalog
Data Catalog S3 tables
Data Catalog Redshift tables
Data Catalog RDS tables
All AWS Glue tables
Others
Amazon AppFlow
AWS Data Exchange
External data connections
Snowflake
AWS Account
Current AWS account
420240645590
Another AWS account
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
Created on
cleanzonedb
2 months ago
February 28, 2023, 10:16:19 pm
curatedzonedb
12 days ago
April 17, 2023, 9:11:49 pm
sakila
20 days ago
April 9, 2023, 10:57:56 pm
streaming_db
18 days ago
April 11, 2023, 9:22:23 pm


DataBrew > Projects > Create project
Create project Info
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
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
1
Dataset name
Data type
Source
Create date
address-dataset
Data Catalog table
Data Catalog
5 days ago
April 24, 2023, 10:48:44 pm
customer-dataset
Data Catalog table
Data Catalog
5 days ago
April 24, 2023, 10:47:53 pm

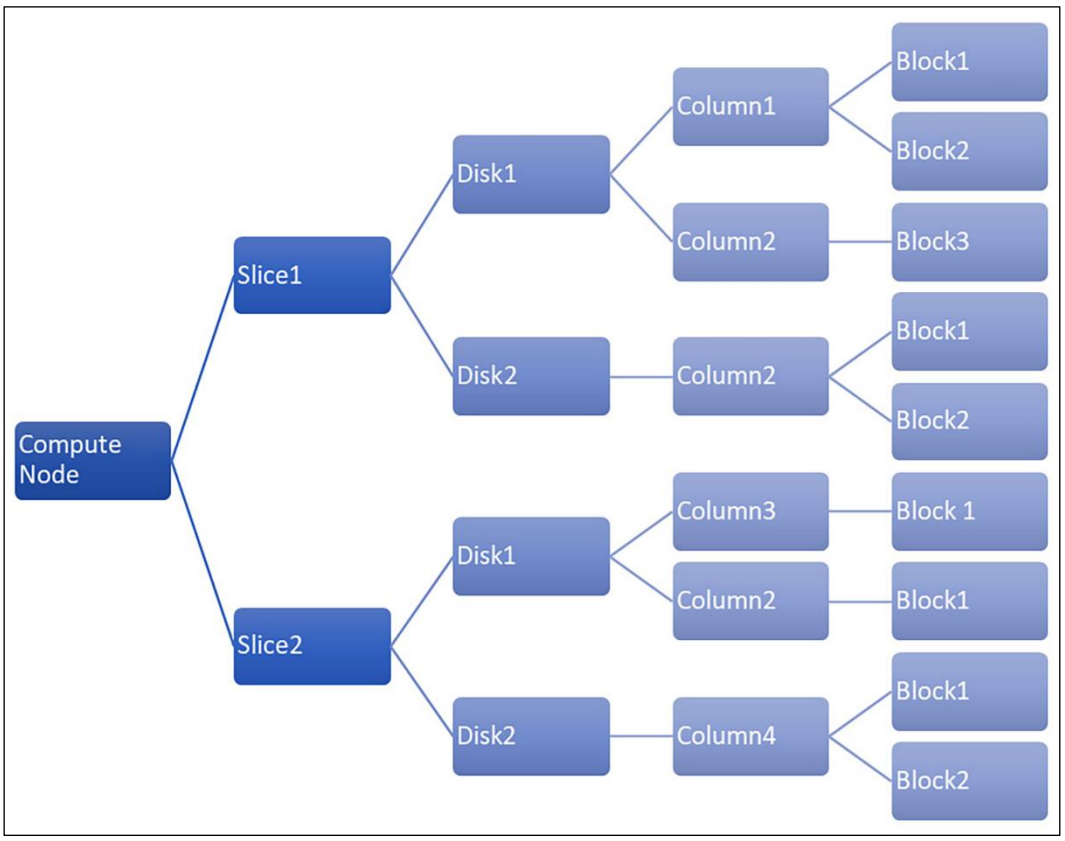

Dataset name
Data type
Source
Create date
address-dataset
Data Catalog table
Data Catalog
5 days ago
April 24, 2023, 10:48:44 pm
customer-dataset
Data Catalog table
Data Catalog
5 days ago
April 24, 2023, 10:47:53 pm
Sampling - optional
Select the type and size of your sample
Tags - optional
Metadata 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
Cancel
Create project
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW

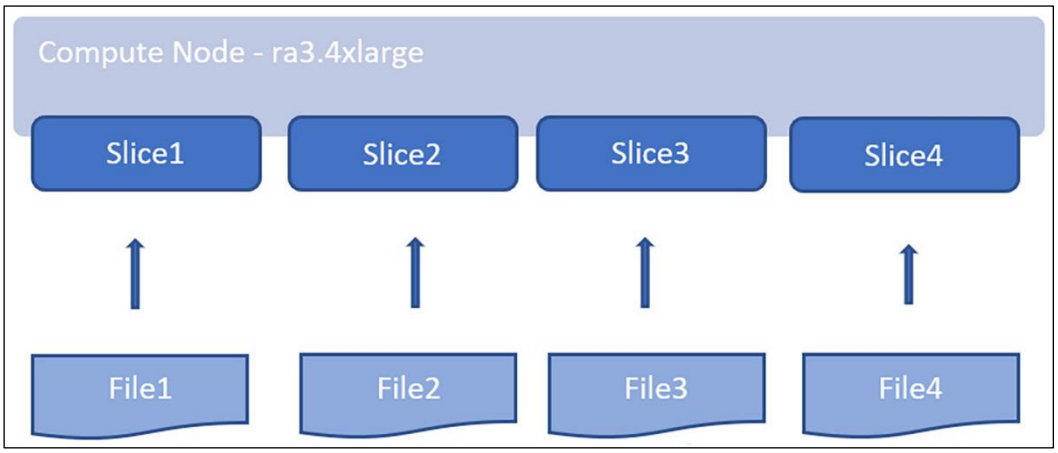

customer-mailing-list
Dataset: customer-dataset
Sample: First n sample (500 rows)
Create job
LINEAGE
ACTIONS
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
NEST-UNNEST
PIVOT
GROUP
JOIN
UNION
MORE
RECIPE
SAMPLE
Viewing 9 columns
500 rows
GRID
SCHEMA
PROFILE
customer_id
first_name
500 valid values
Recipe (0)
customer-mailing-list-recipe
Working version
More
WILLIE
TERRY
MARY
All other values
1 MARY
2 PATRICIA
3 LINDA
4 BARBARA
5 ELIZABETH
6 JENNIFER
7 MARIA
8 SUSAN
9 MARGARET
10 DOROTHY
Build your recipe
Start applying transformation steps to your data. All your data preparation steps will be tracked in the recipe.
Add step
Zoom
100%

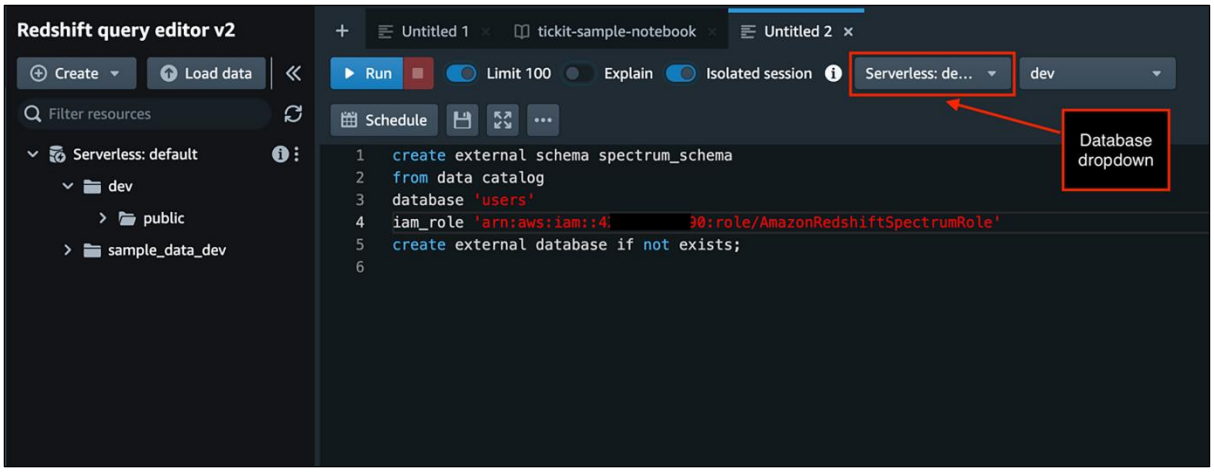

Recipe (4)
customer-mailing-list-recipe
Working version
Publish
More
Applied steps (4) | Clear all
1. Left join address-dataset
2. Change format of first_name to Capital case
3. Change format of last_name to Capital case
4. Change format of email to Lowercase


Created recipe job "mailing-list-job".
DataBrew > Jobs > mailing-list-job
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
mailing-list-job
Dataset: customer-dataset
Project: customer-mailing-list
Recipe: customer-mailing-list-recipe
Run job
Actions
OPEN PROJECT
Job run history
Job details
Data lineage
Recipe
Last job run 4 minutes, no job runs scheduled
Job run history
Stop job run
Actions
Search by job run ID
Show all
Job run ID
Last job run status
Run time
Output
Summary
mailing-list-job_2023-04-29-20:18:43
Succeeded
1 minute, 24 seconds
1 output

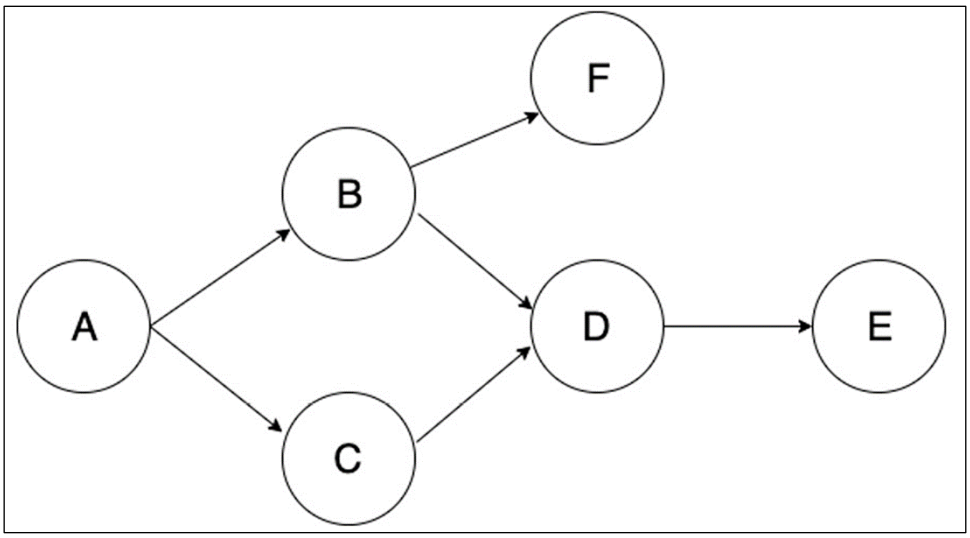
Chapter 9: A Deeper Dive into Data Marts and Amazon Redshift

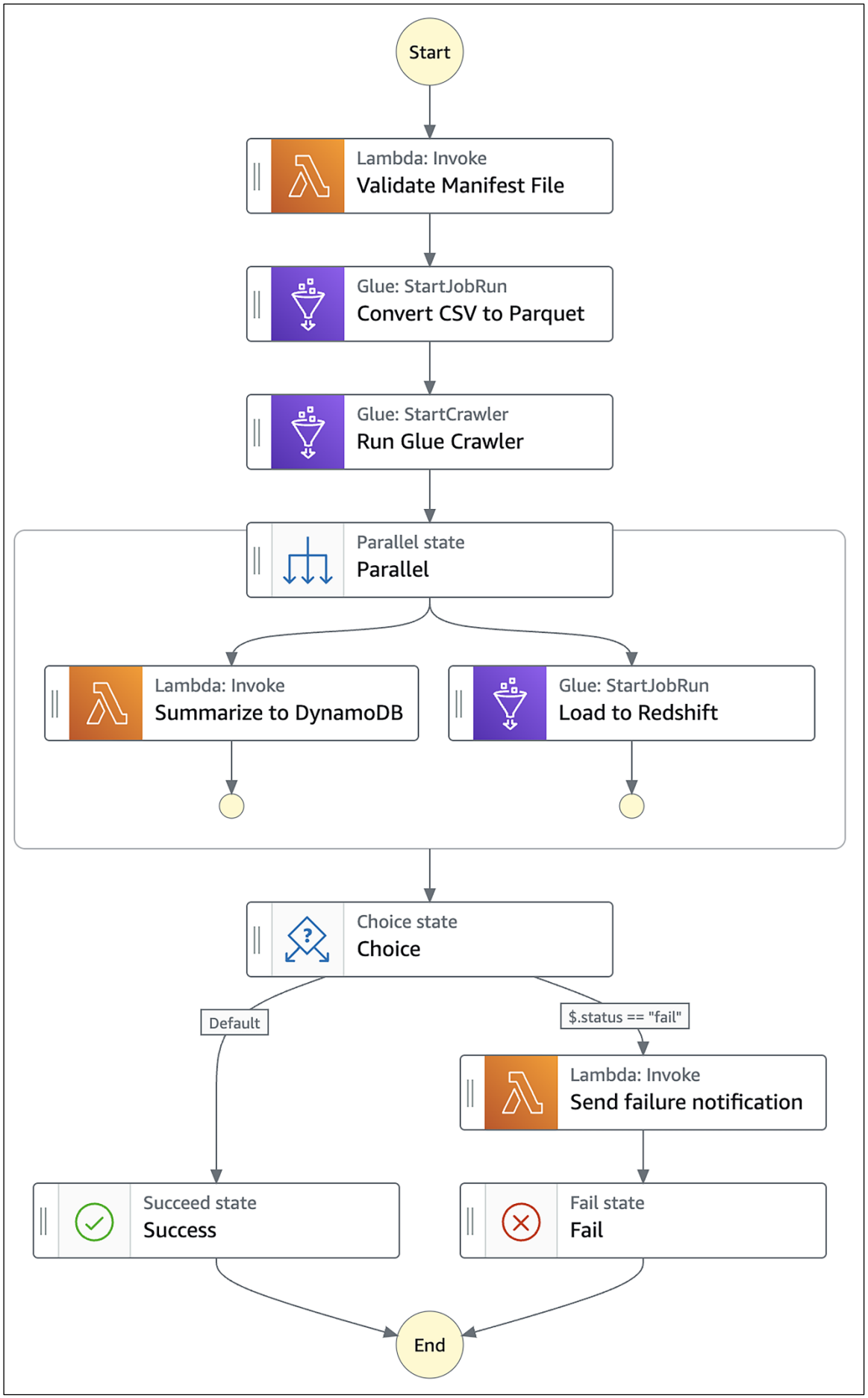




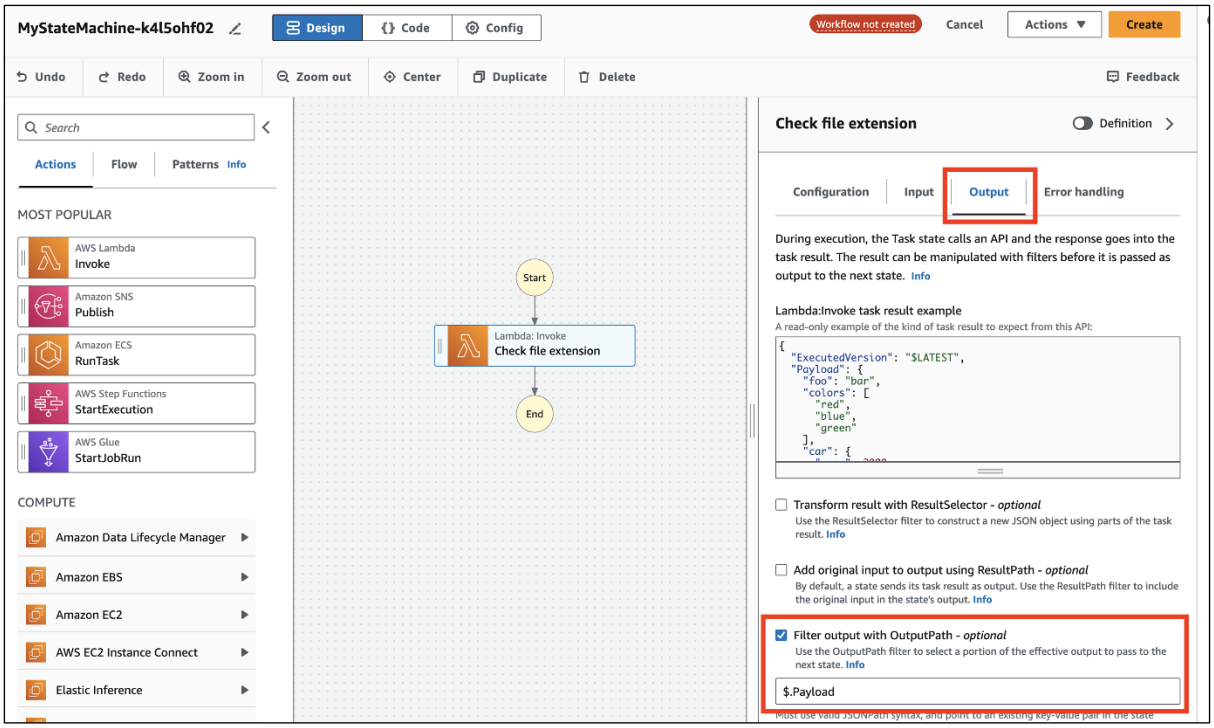

Redshift query editor v2
Create
Load data
Filter resources
Serverless: default
dev
public
sample_data_dev
Untitled 1
tickit-sample-notebook
Untitled 2
Run
Limit 100
Explain
Isolated session
Serverless: de...
dev
Schedule
Database dropdown
1 create external schema spectrum_schema
2 from data catalog
3 database 'users'
4 iam_role 'arn:aws:iam::4 90:role/AmazonRedshiftSpectrumRole'
5 create external database if not exists;
6

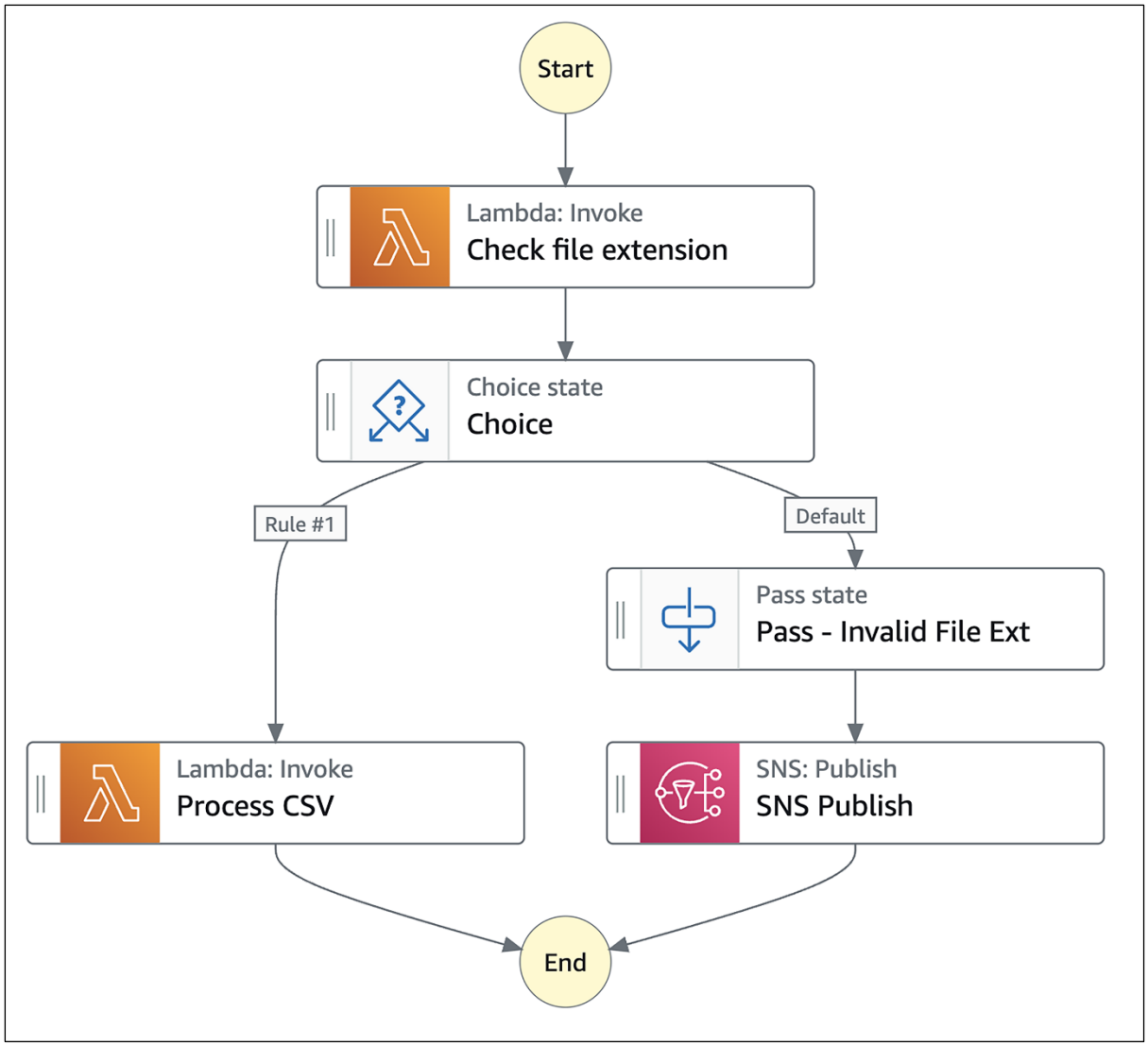
Chapter 10: Orchestrating the Data Pipeline

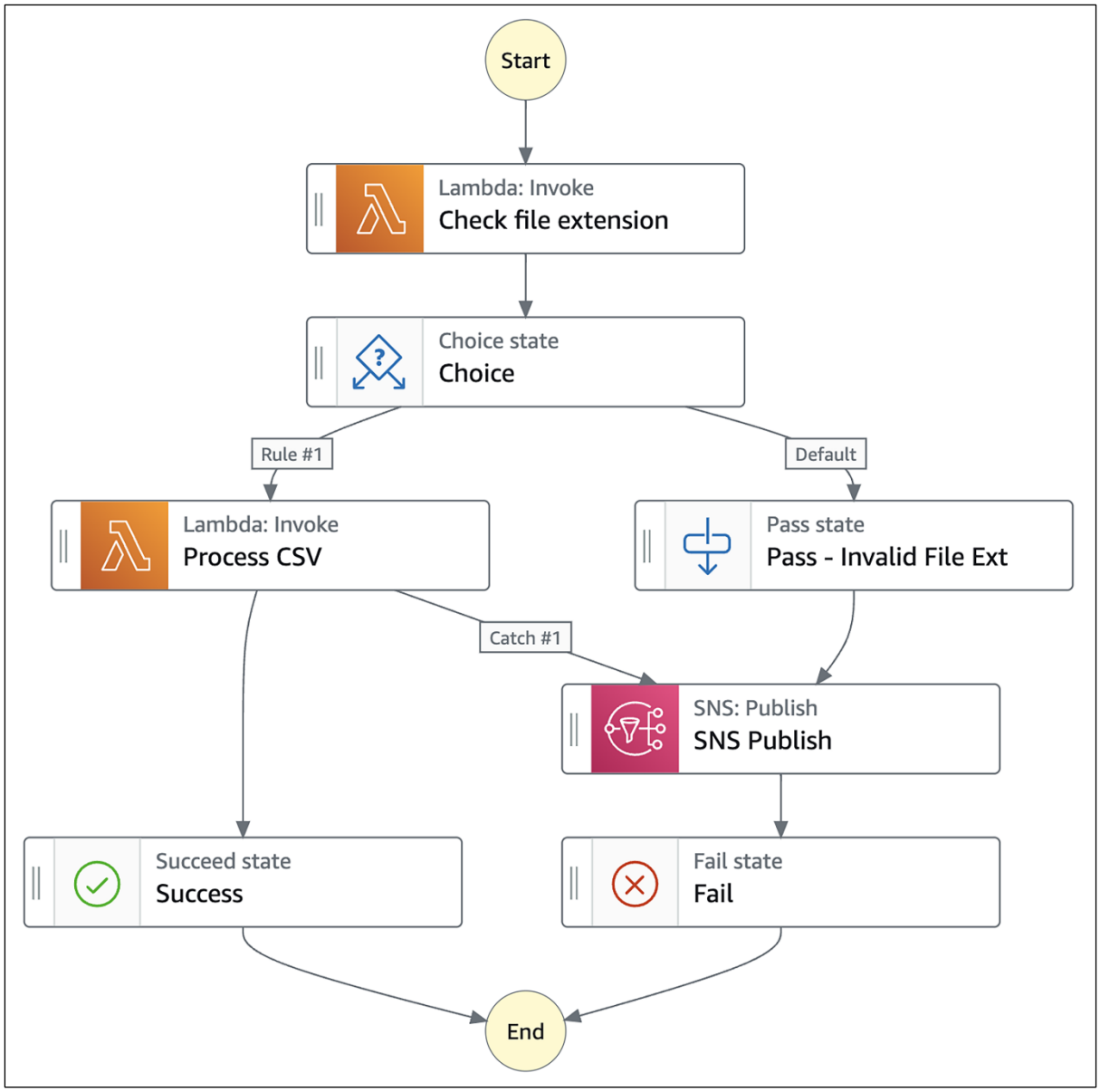


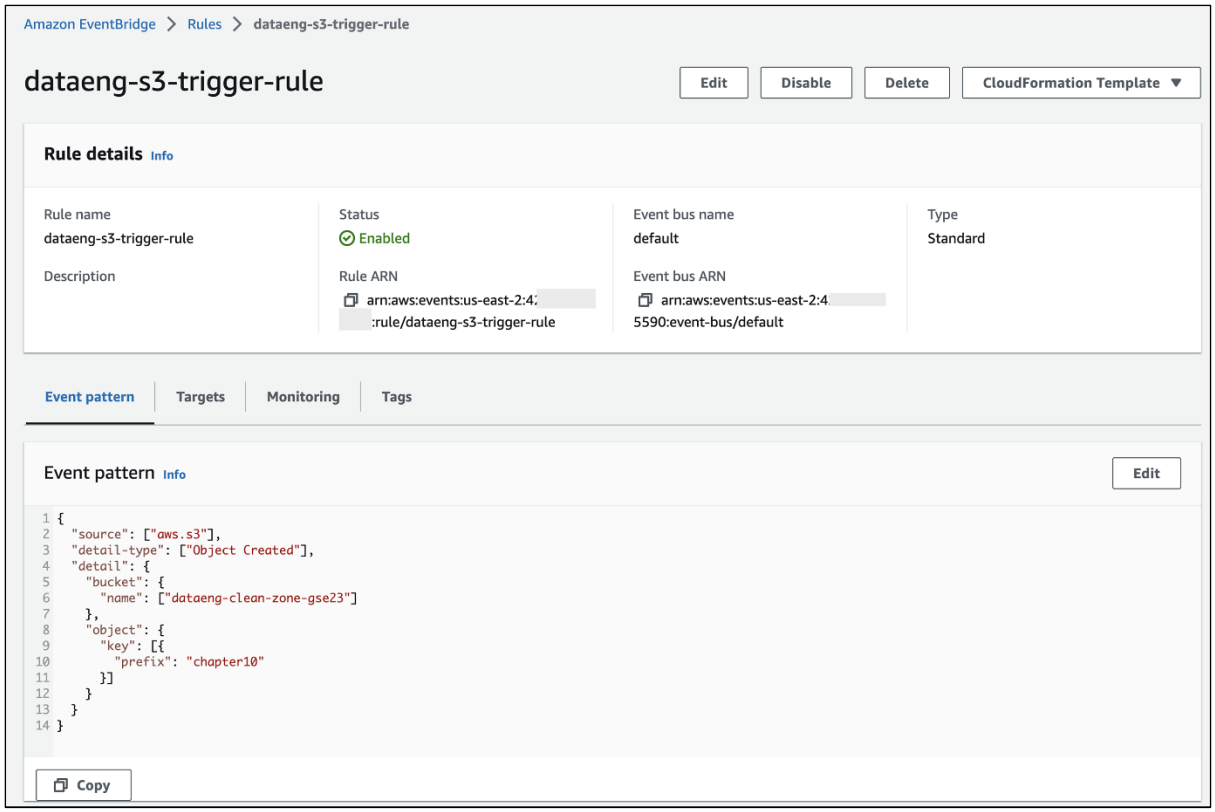

Start
Lambda: Invoke
Validate Manifest File
Glue: StartJobRun
Convert CSV to Parquet
Glue: StartCrawler
Run Glue Crawler
Parallel state
Parallel
Lambda: Invoke
Summarize to DynamoDB
Glue: StartJobRun
Load to Redshift
Choice state
Choice
Default
$.status == "fail"
Lambda: Invoke
Send failure notification
Succeed state
Success
Fail state
Fail
End

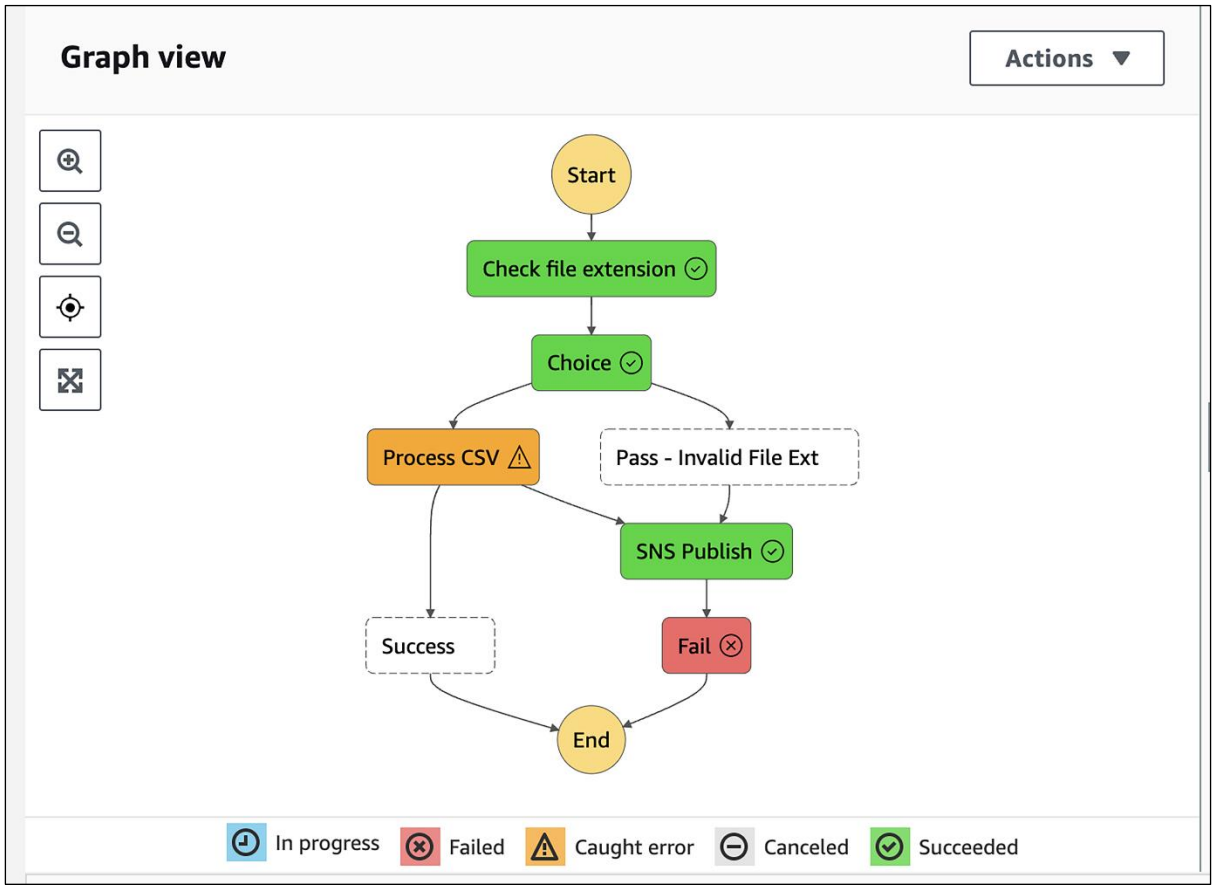
Criteria	AWS Step Functions	Amazon Managed Workflows for Apache Airflow (MWAA)
Short description	Serverless AWS native orchestration service	Managed AWS service for open source Apache Airflow
Graphical pipeline development	Yes	No
Graphical run visualization	Yes	Yes
Error and retry single step	Yes	Yes
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


MyStateMachine-k4l5ohf02
Design
Code
Config
Workflow not created
Cancel
Actions
Create
Undo
Redo
Zoom in
Zoom out
Center
Duplicate
Delete
Feedback
Search
Actions
Flow
Patterns Info
MOST POPULAR
AWS Lambda
Invoke
Amazon SNS
Publish
Amazon ECS
RunTask
AWS Step Functions
StartExecution
AWS Glue
StartJobRun
COMPUTE
Amazon Data Lifecycle Manager
Amazon EBS
Amazon EC2
AWS EC2 Instance Connect
Elastic Inference
Start
Lambda: Invoke
Check file extension
End
Check file extension
Definition
Configuration
Input
Output
Error handling
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
Use the ResultSelector filter to construct a new JSON object using parts of the task result. Info
Add original input to output using ResultPath - optional
By default, a state sends its task result as output. Use the ResultPath filter to include the original input in the state's output. Info
Filter output with OutputPath - optional
Use the OutputPath filter to select a portion of the effective output to pass to the next state. Info
$.Payload

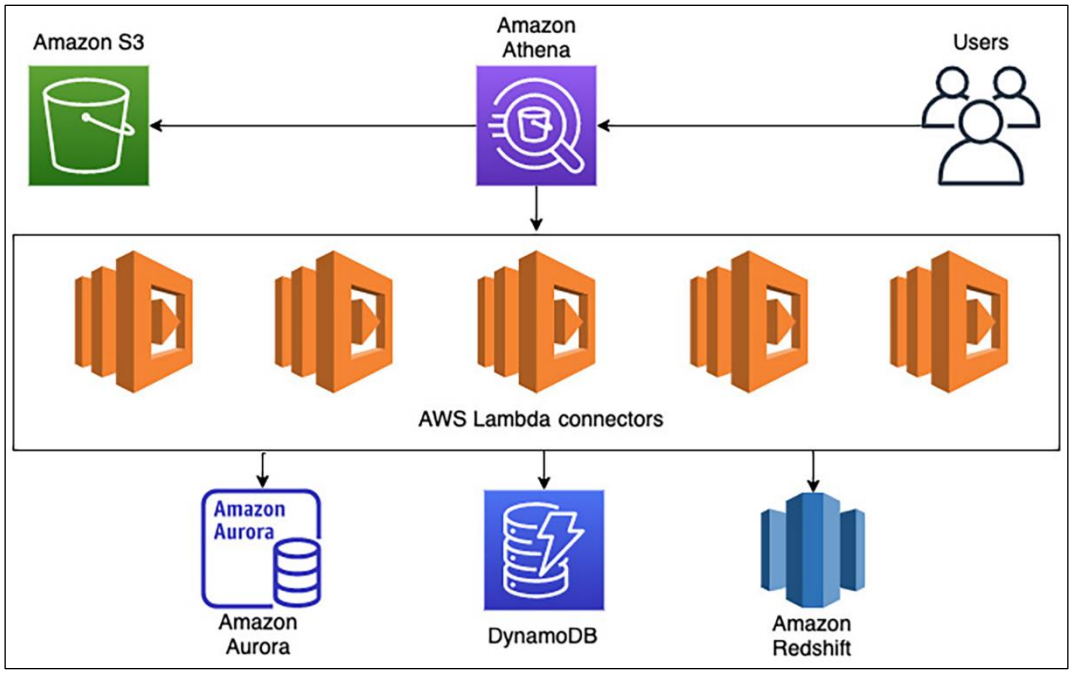

Start
Lambda: Invoke
Check file extension
Choice state
Choice
Rule #1
Default
Pass state
Pass - Invalid File Ext
Lambda: Invoke
Process CSV
SNS: Publish
SNS Publish
End

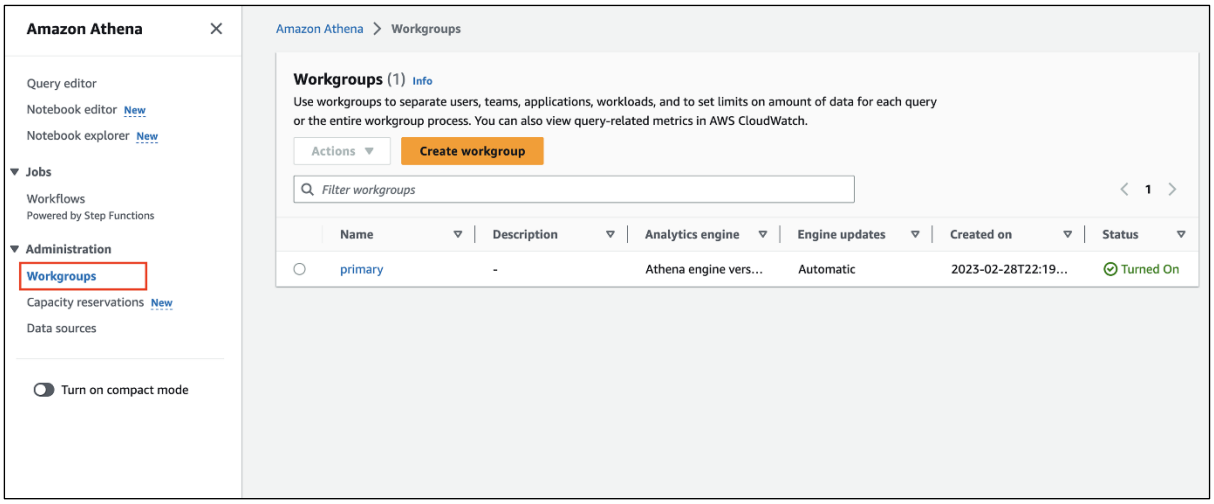

Start
Lambda: Invoke
Check file extension
Choice state
Choice
Rule #1
Default
Lambda: Invoke
Process CSV
Pass state
Pass - Invalid File Ext
Catch #1
SNS: Publish
SNS Publish
Succeed state
Success
Fail state
Fail
End

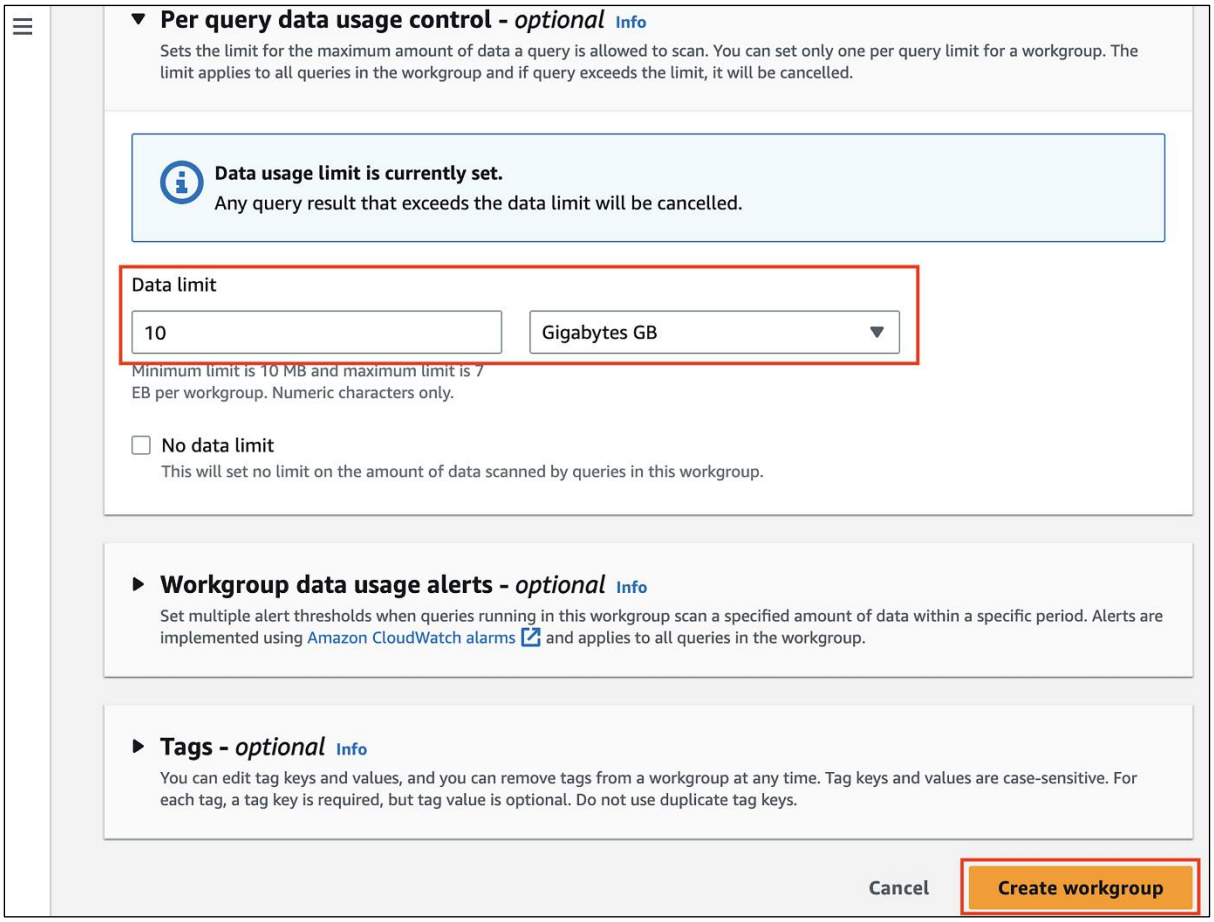

Amazon EventBridge > Rules > dataeng-s3-trigger-rule
dataeng-s3-trigger-rule
Edit
Disable
Delete
CloudFormation Template
Rule details Info
Rule name
dataeng-s3-trigger-rule
Status
Enabled
Event bus name
default
Type
Standard
Description
Rule ARN
arn:aws:events:us-east-2:4
:rule/dataeng-s3-trigger-rule
Event bus ARN
arn:aws:events:us-east-2:4
5590:event-bus/default
Event pattern
Targets
Monitoring
Tags
Event pattern Info
Edit
{
"source": ["aws.s3"],
"detail-type": ["Object Created"],
"detail": {
"bucket": {
"name": ["dataeng-clean-zone-gse23"]
},
"object": {
"key": [{
"prefix": "chapter10"
}]
}
}
}
Copy

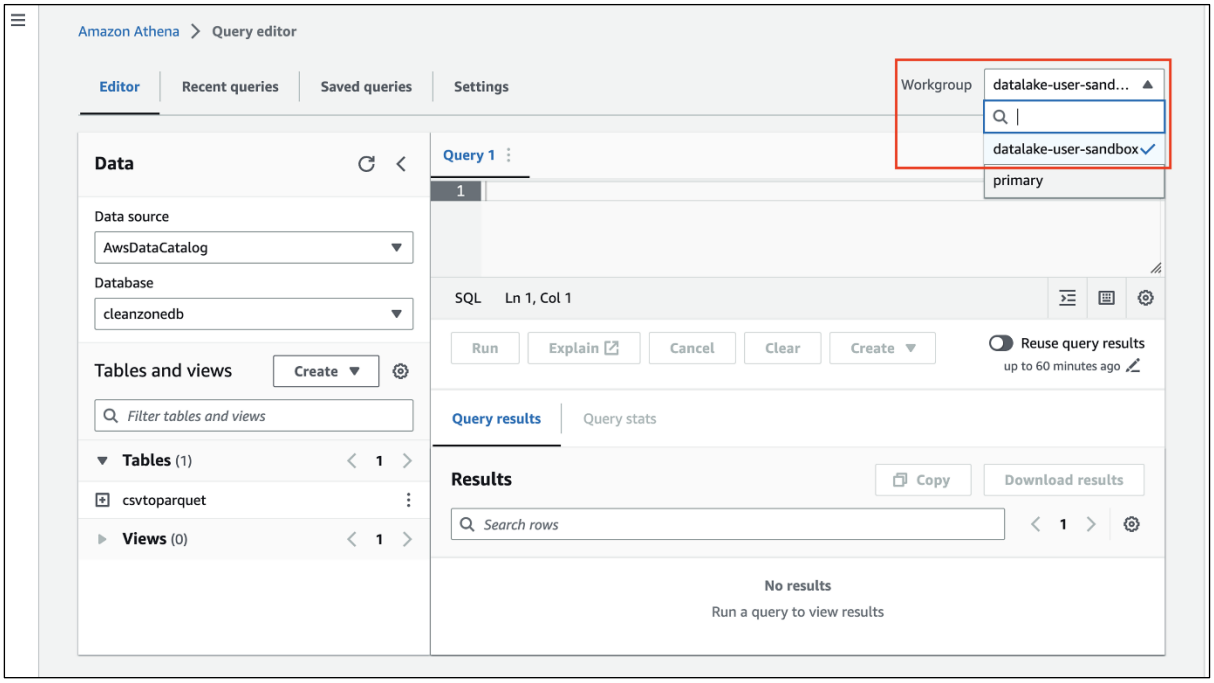

Graph view
Actions
Start
Check file extension
Choice
Process CSV
Pass - Invalid File Ext
SNS Publish
Success
Fail
End
In progress
Failed
Caught error
Canceled
Succeeded

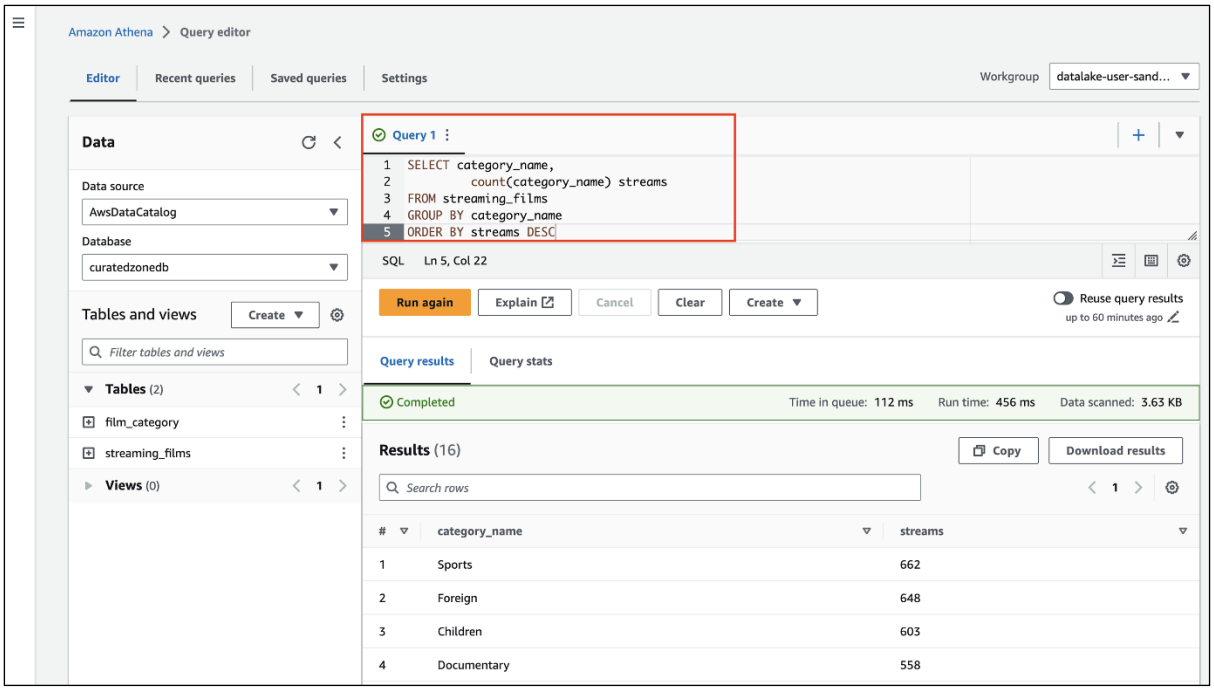
Chapter 11: Ad Hoc Queries with Amazon Athena

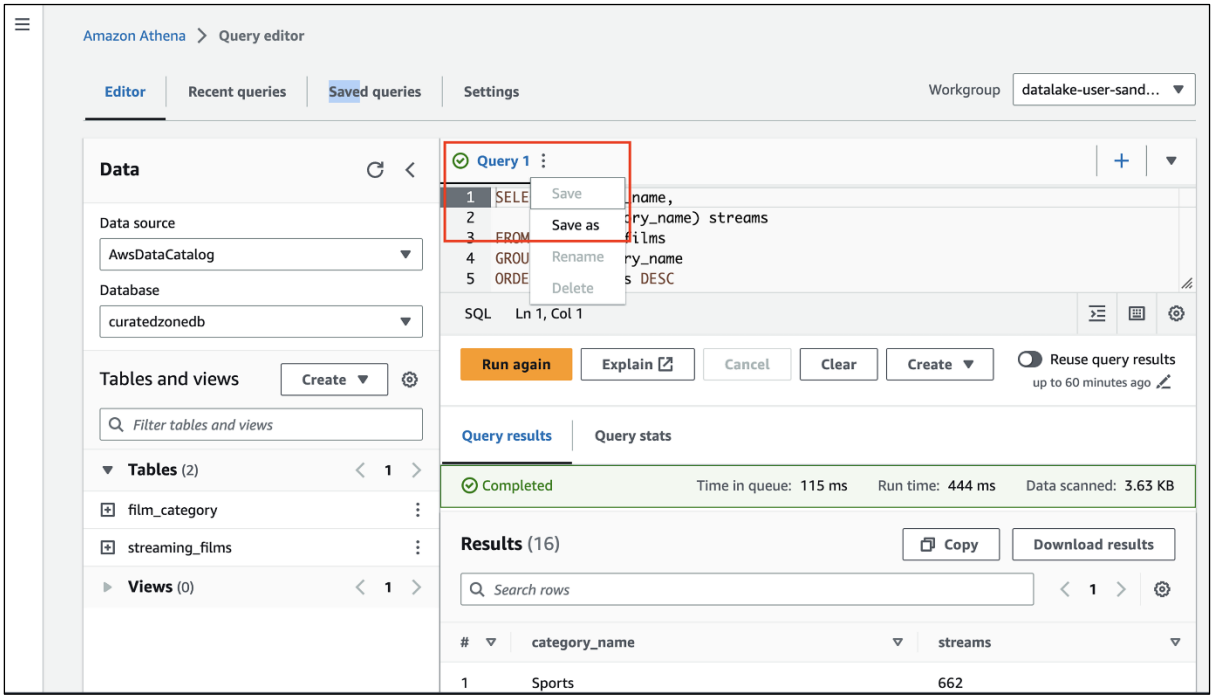


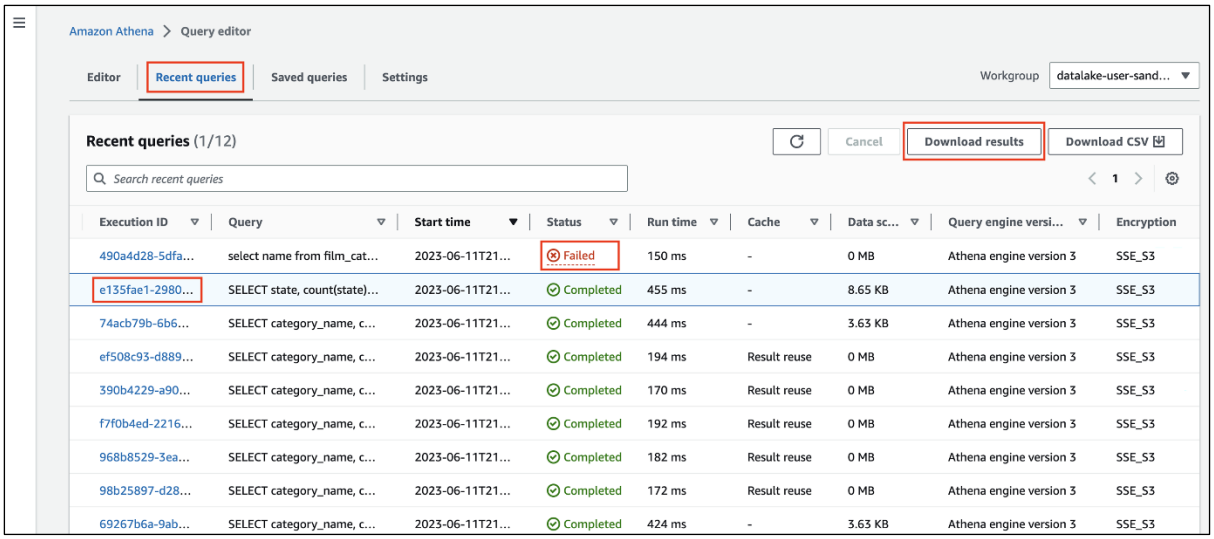



Per query data usage control - optional Info
Sets the limit for the maximum amount of data a query is allowed to scan. You can set only one per query limit for a workgroup. The limit applies to all queries in the workgroup and if query exceeds the limit, it will be cancelled.
Data usage limit is currently set.
Any query result that exceeds the data limit will be cancelled.
Data limit
10
Gigabytes GB
Minimum limit is 10 MB and maximum limit is 7 EB per workgroup. Numeric characters only.
No data limit
This will set no limit on the amount of data scanned by queries in this workgroup.
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 workgroup 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.
Cancel
Create workgroup

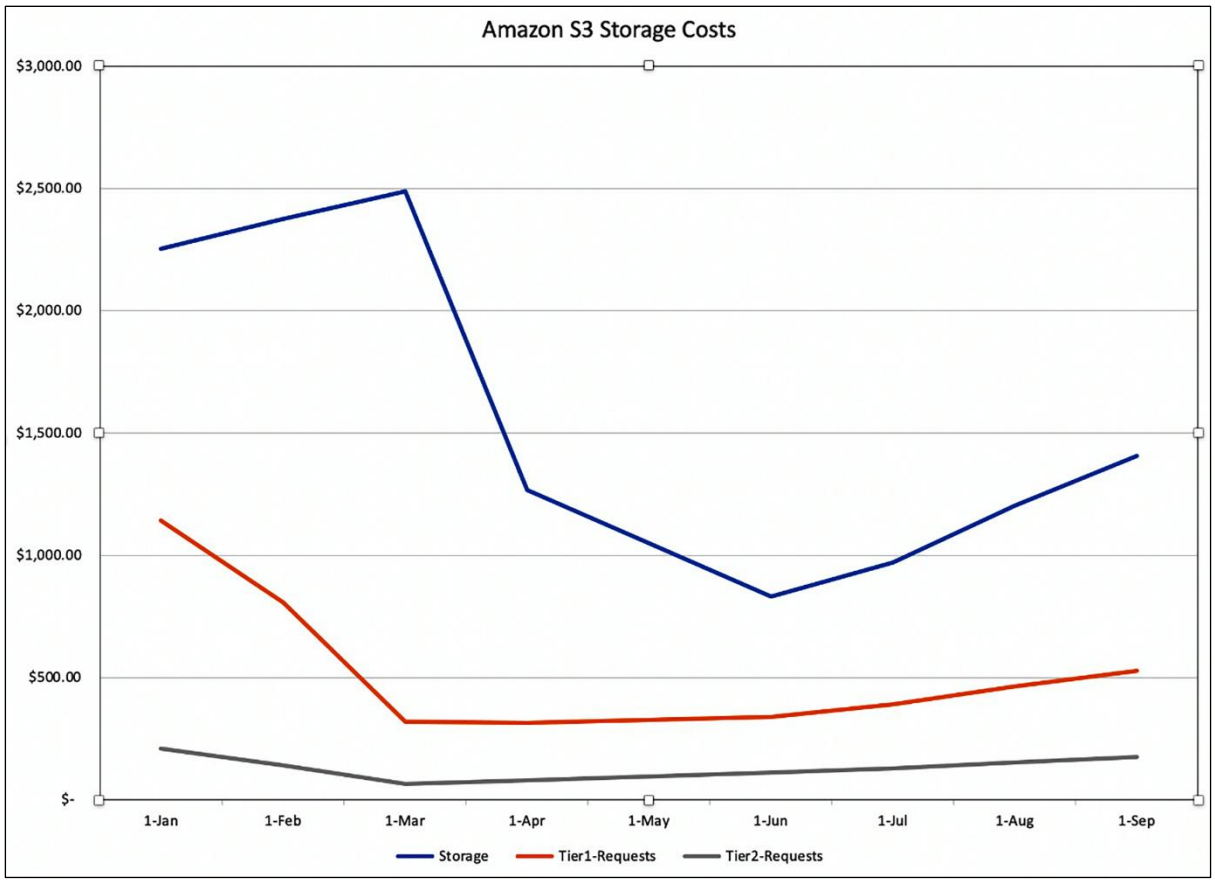

Amazon Athena
Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
datalake-user-sand...
datalake-user-sandbox
primary
Data
Data source
AwsDataCatalog
Database
cleanzonedb
Tables and views
Create
Filter tables and views
Tables (1)
csvtoparquet
Views (0)
Query 1
SQL
Ln 1, Col 1
Run
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Results
Copy
Download results
Search rows
No results
Run a query to view results

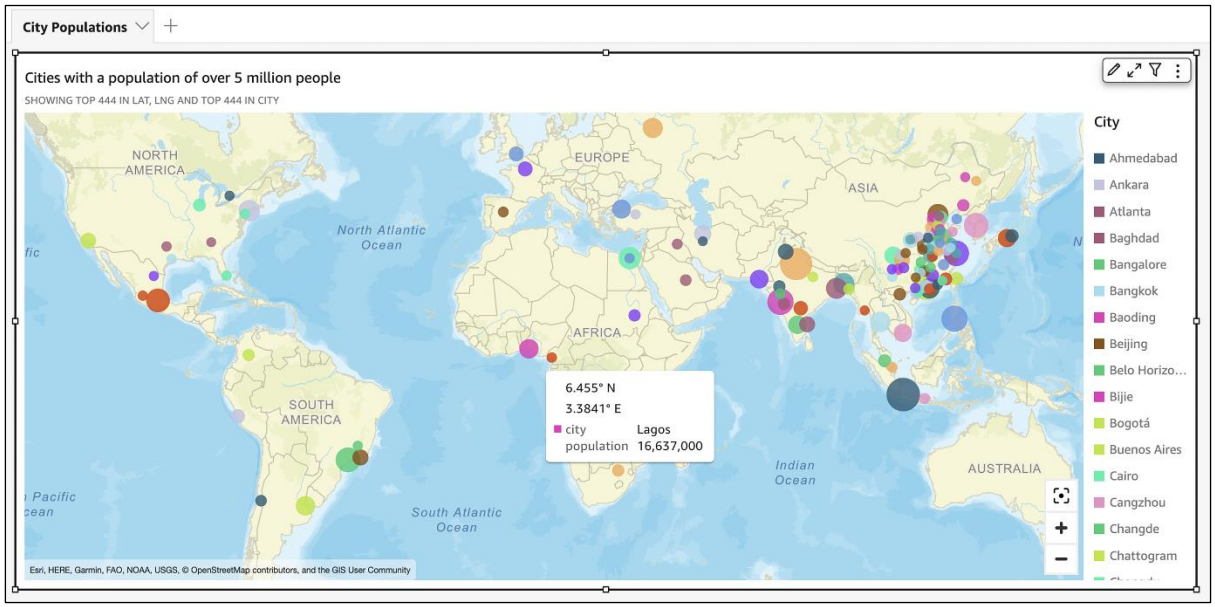

Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
datalake-user-sand...
Data
Data source
AwsDataCatalog
Database
curatedzonedb
Tables and views
Create
Filter tables and views
Tables (2)
film_category
streaming_films
Views (0)
Query 1
1 SELECT category_name,
2 count(category_name) streams
3 FROM streaming_films
4 GROUP BY category_name
5 ORDER BY streams DESC
SQL Ln 5, Col 22
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 112 ms
Run time: 456 ms
Data scanned: 3.63 KB
Results (16)
Copy
Download results
Search rows
category_name streams
1 Sports 662
2 Foreign 648
3 Children 603
4 Documentary 558

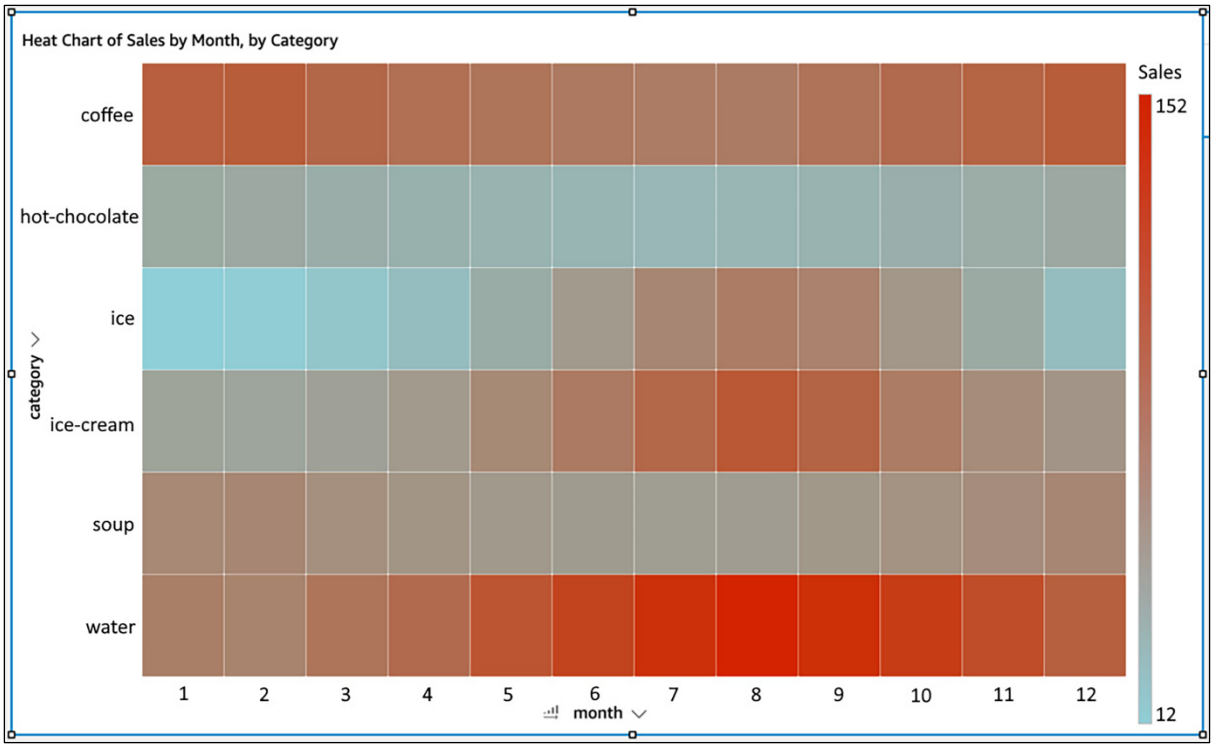

Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
datalake-user-sand...
Data
Data source
AwsDataCatalog
Database
curatedzonedb
Tables and views
Create
Filter tables and views
Tables (2)
film_category
streaming_films
Views (0)
Query 1
Save
Save as
Rename
Delete
SQL Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 115 ms
Run time: 444 ms
Data scanned: 3.63 KB
Results (16)
Copy
Download results
Search rows
category_name streams
1 Sports 662

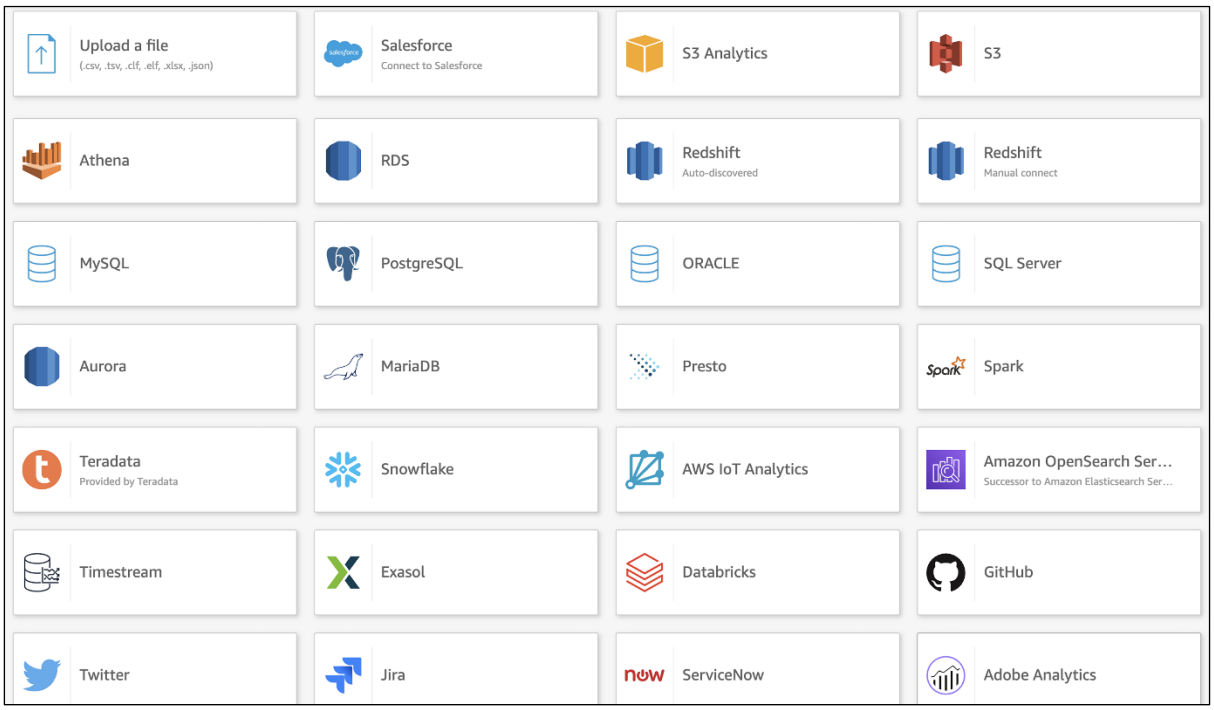

Amazon Athena
Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
datalake-user-sand...
Recent queries (1/12)
Cancel
Download results
Download CSV
Search recent queries
1
Execution ID
Query
Start time
Status
Run time
Cache
Data sc...
Query engine versi...
Encryption
490a4d28-5dfa...
select name from film_cat...
2023-06-11T21...
Failed
150 ms
-
0 MB
Athena engine version 3
SSE_S3
e135fae1-2980...
SELECT state, count(state)...
2023-06-11T21...
Completed
455 ms
-
8.65 KB
Athena engine version 3
SSE_S3
74acb79b-6b6...
SELECT category_name, c...
2023-06-11T21...
Completed
444 ms
-
3.63 KB
Athena engine version 3
SSE_S3
ef508c93-d889...
SELECT category_name, c...
2023-06-11T21...
Completed
194 ms
Result reuse
0 MB
Athena engine version 3
SSE_S3
390b4229-a90...
SELECT category_name, c...
2023-06-11T21...
Completed
170 ms
Result reuse
0 MB
Athena engine version 3
SSE_S3
f7f0b4ed-2216...
SELECT category_name, c...
2023-06-11T21...
Completed
192 ms
Result reuse
0 MB
Athena engine version 3
SSE_S3
968b8529-3ea...
SELECT category_name, c...
2023-06-11T21...
Completed
182 ms
Result reuse
0 MB
Athena engine version 3
SSE_S3
98b25897-d28...
SELECT category_name, c...
2023-06-11T21...
Completed
172 ms
Result reuse
0 MB
Athena engine version 3
SSE_S3
69267b6a-9ab...
SELECT category_name, c...
2023-06-11T21...
Completed
424 ms
-
3.63 KB
Athena engine version 3
SSE_S3

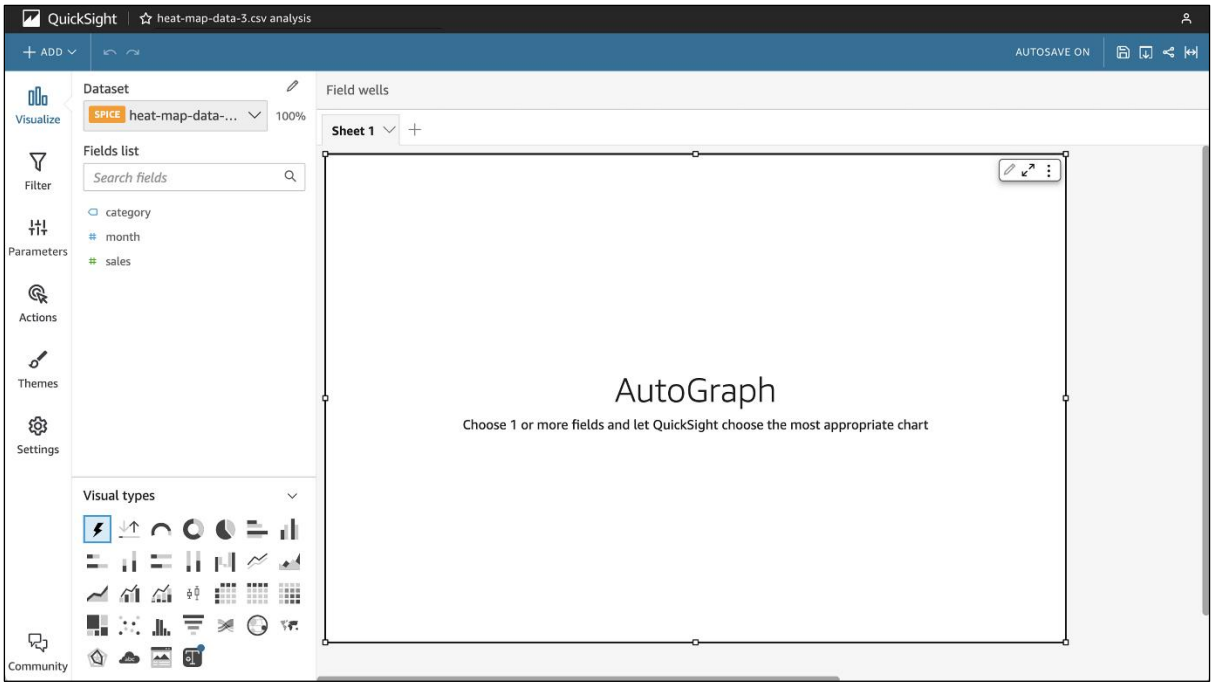
Chapter 12: Visualizing Data with Amazon QuickSight

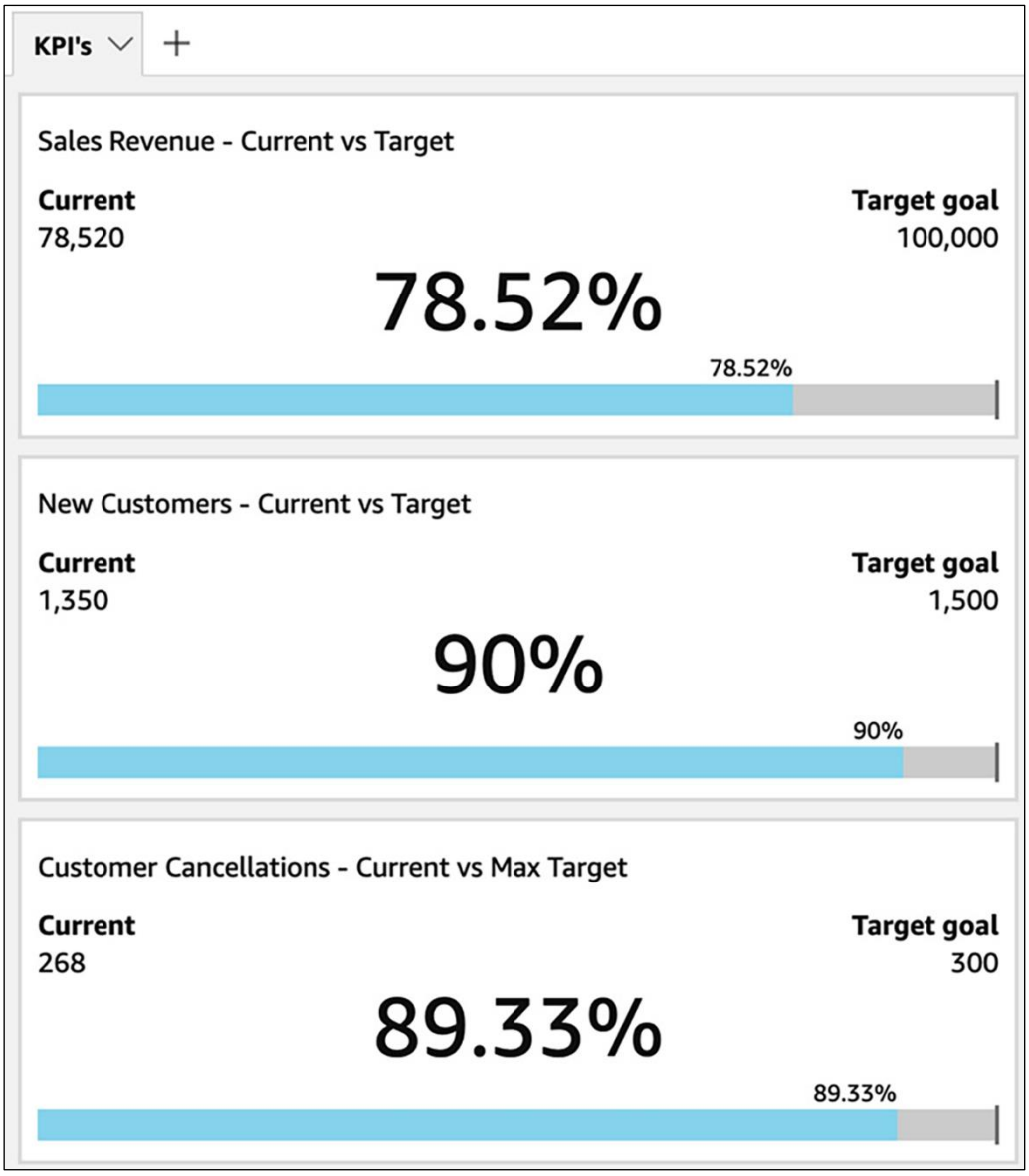





Heat Chart of Sales by Month, by Category
coffee
hot-chocolate
ice
ice-cream
soup
water
category
1
2
3
4
5
6
7
8
9
10
11
12
month
Sales
152
12


Upload a file
(.csv, .tsv, .clf, .elf, .xlsx, .json)
Salesforce
Connect to Salesforce
S3 Analytics
S3
Athena
RDS
Redshift
Auto-discovered
Redshift
Manual connect
MySQL
PostgreSQL
ORACLE
SQL Server
Aurora
MariaDB
Presto
Spark
Teradata
Provided by Teradata
Snowflake
AWS IoT Analytics
Amazon OpenSearch Ser…
Successor to Amazon Elasticsearch Ser…
Timestream
Exasol
Databricks
GitHub
Twitter
Jira
ServiceNow
Adobe Analytics


QuickSight
heat-map-data-3.csv analysis
ADD
AUTOSAVE ON
Visualize
Filter
Parameters
Actions
Themes
Settings
Community
Dataset
SPICE
heat-map-data-...
100%
Fields list
Search fields
category
month
sales
Visual types
Field wells
Sheet 1
AutoGraph
Choose 1 or more fields and let QuickSight choose the most appropriate chart


KPI's
Sales Revenue - Current vs Target
Current
78,520
Target goal
100,000
78.52%
78.52%
New Customers - Current vs Target
Current
1,350
Target goal
1,500
90%
90%
Customer Cancellations - Current vs Max Target
Current
268
Target goal
300
89.33%
89.33%

Edition	Enterprise	Enterprise + Q Learn more
Team trial for 30 days (4 authors)*	FREE	FREE
Author per month (yearly)**	$18	$28
Author per month (monthly)**	$24	$34
Readers (pay-per-Session)	$0.30 / session (max $5)****	$0.30 / session (max $10)****
Additional SPICE per month	$0.38 per GB	$0.38 per GB
QuickSight Q regional fee	N/A	$250 / mo / region
Natural language query with QuickSight Q	N/A	INCLUDED
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	Learn more

* Trial authors are auto-converted to month-to-month subscription upon trial expiry
** Each additional author includes 10GB of SPICE capacity
*** Active Directory groups are available in accounts connected to Active Directory
**** Sessions of 30-minute duration. Total charges for each reader are capped at $5 per month. Conditions apply.

Sign up for Standard Edition here.

QuickSight Standard Sign-up

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)

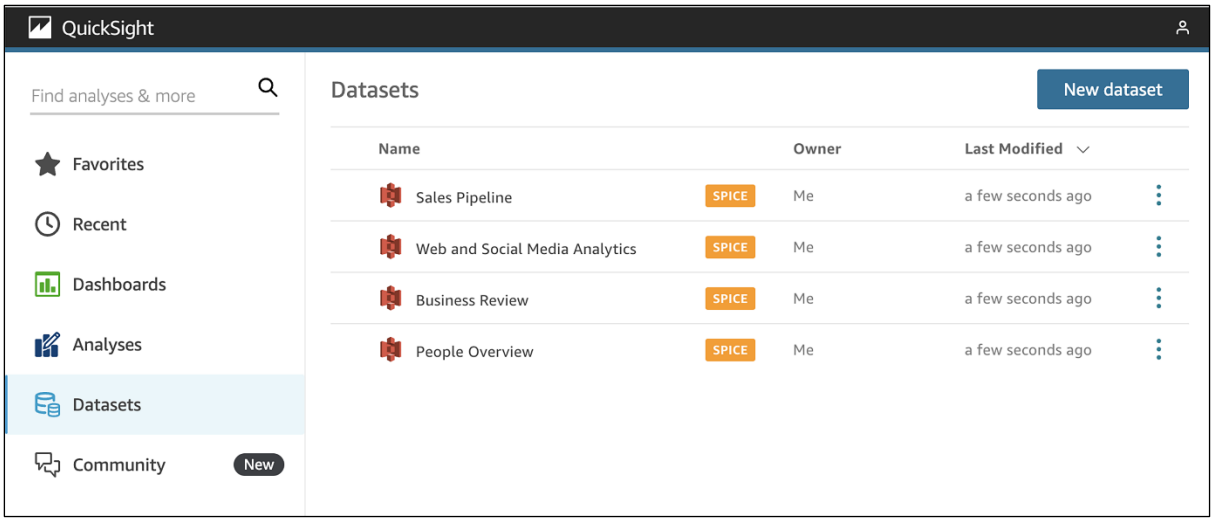
Account info

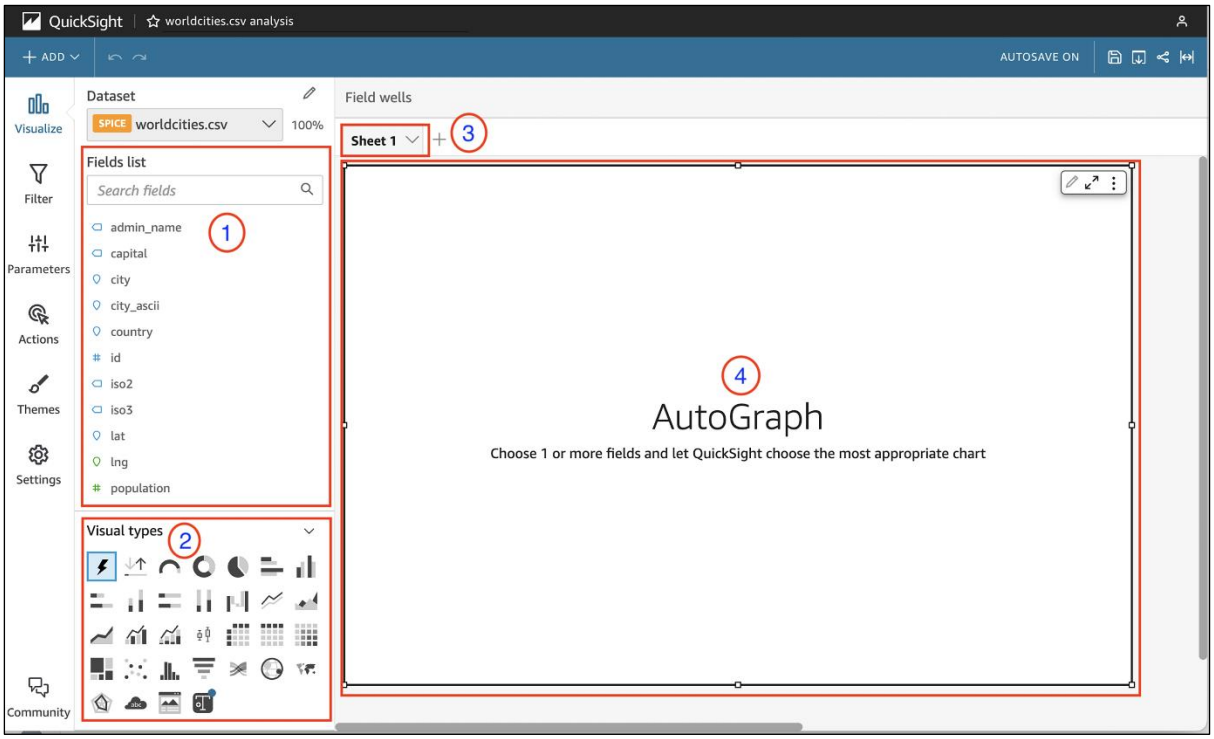
QuickSight account name
You will need this for you and others to sign in

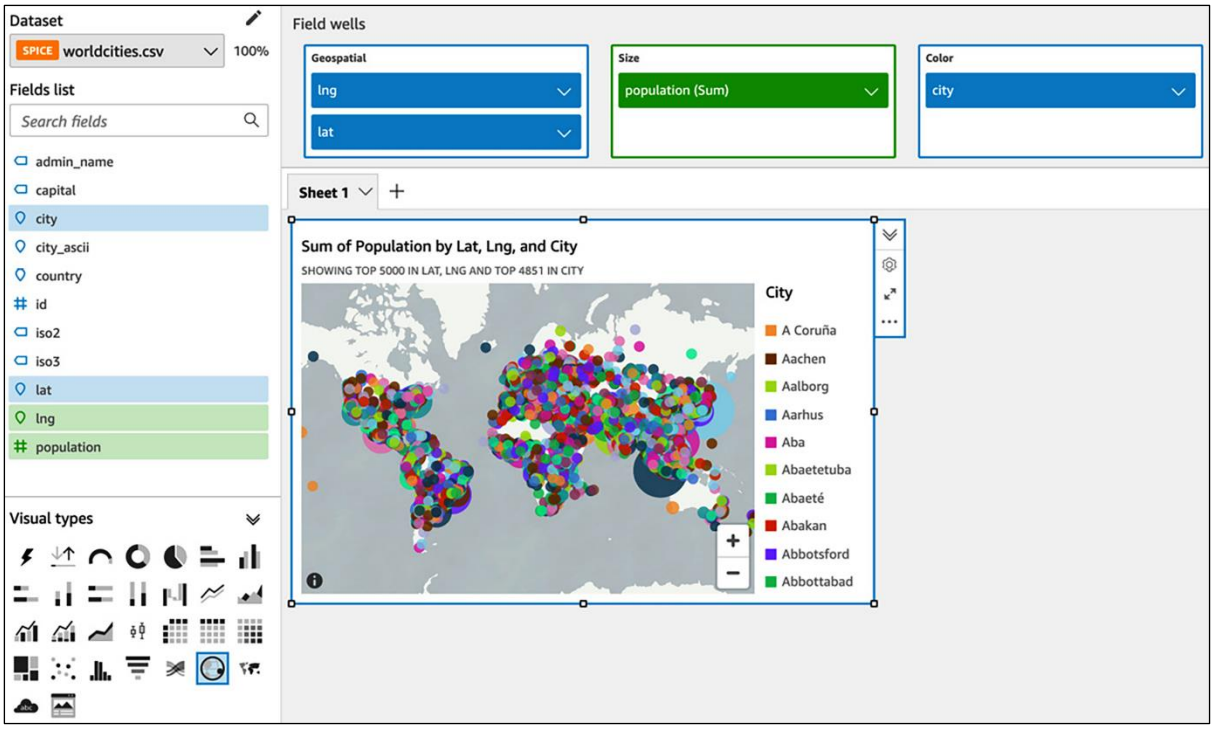
data-

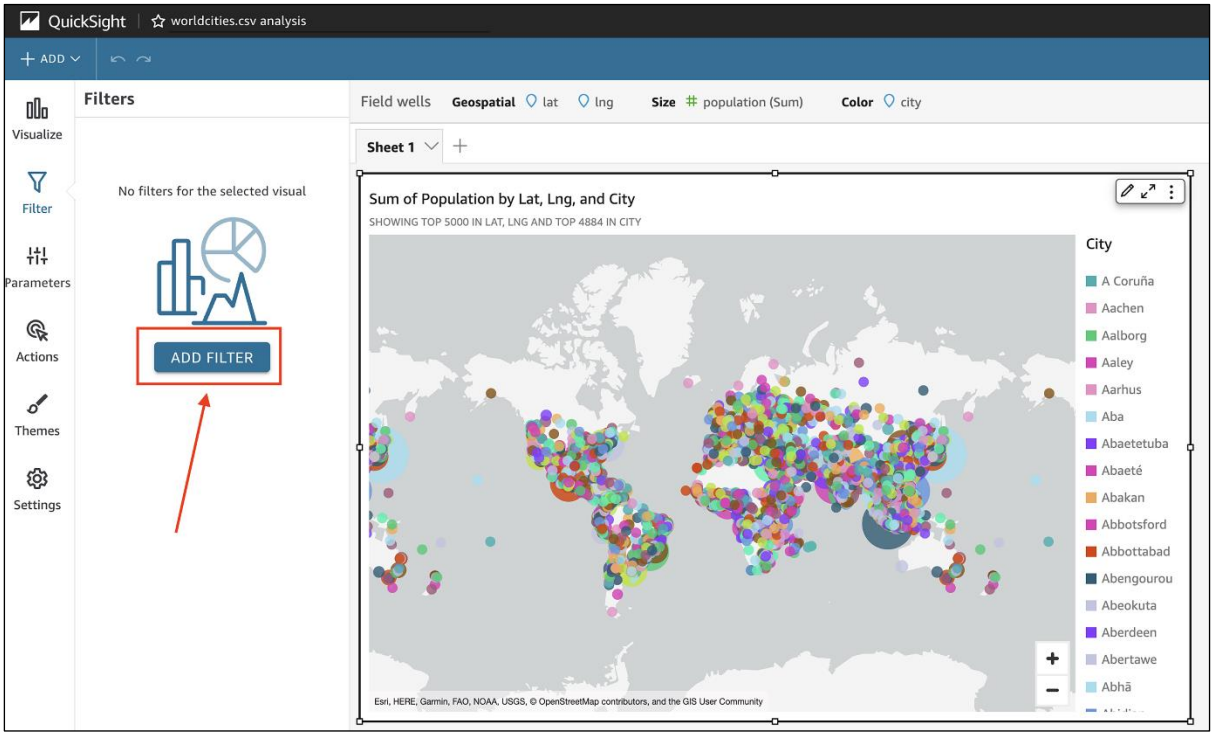
Notification email address
For QuickSight to send important notifications

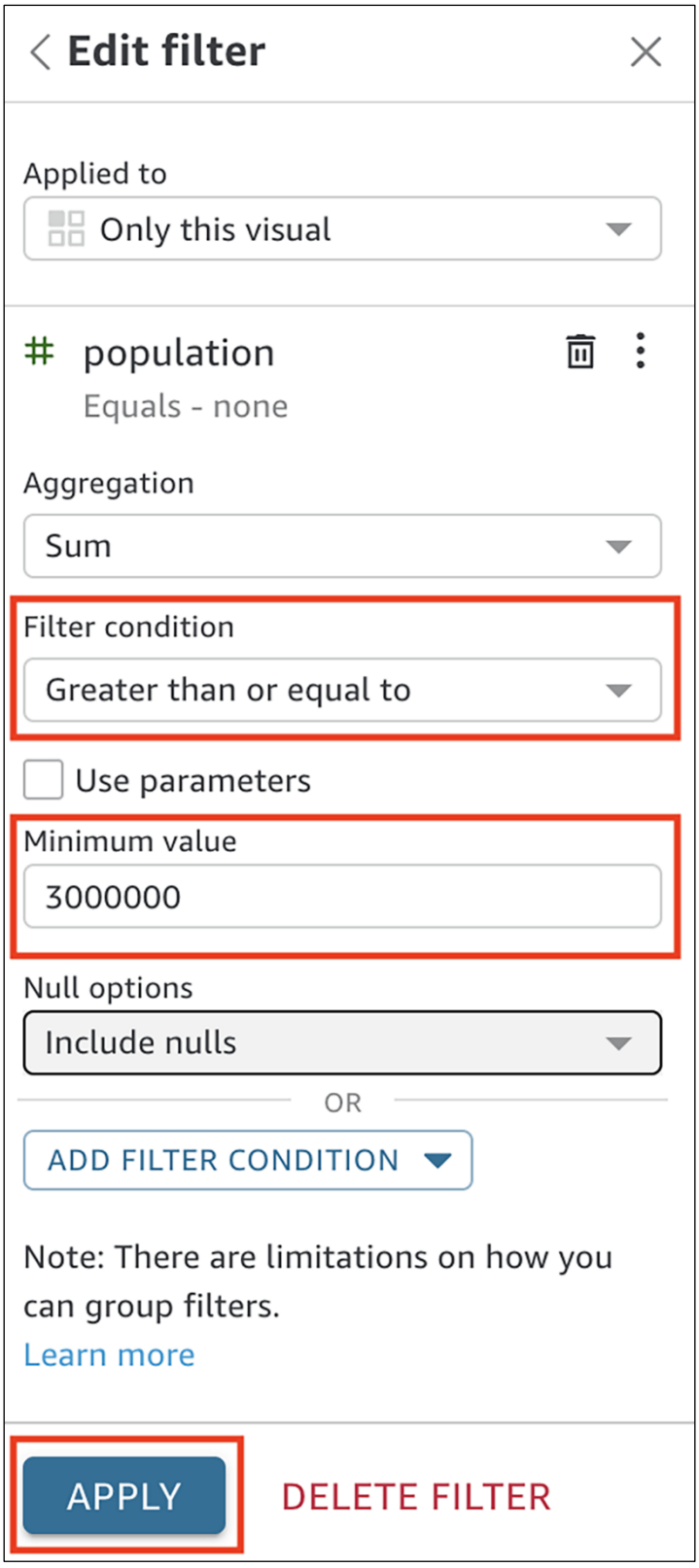
gare l.com

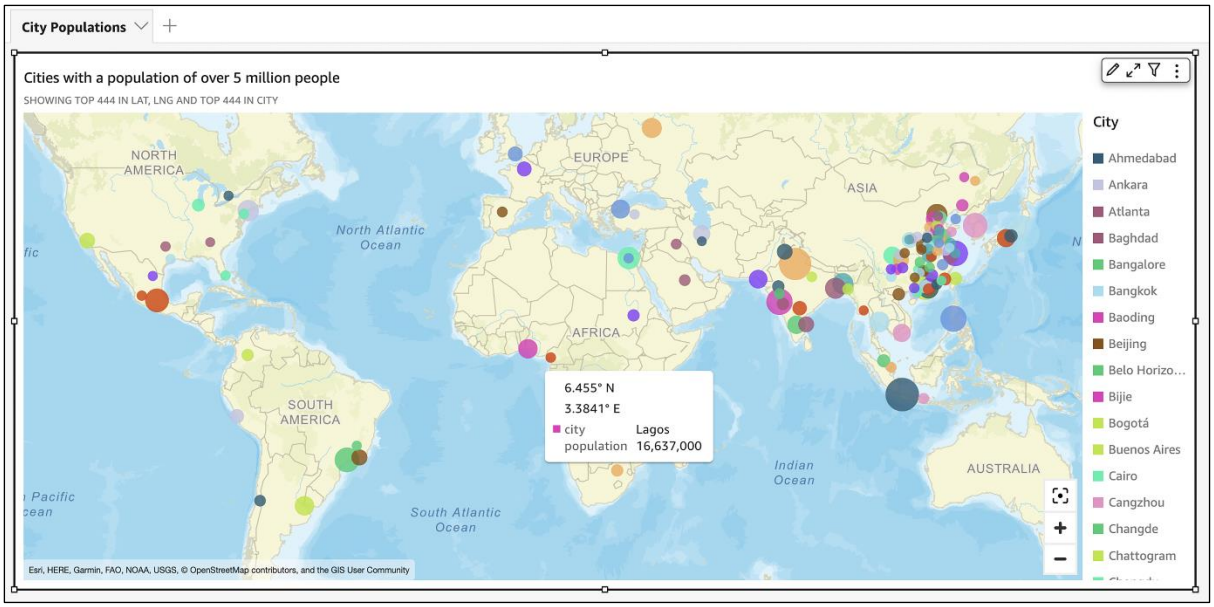

QuickSight
Find analyses & more
Favorites
Recent
Dashboards
Analyses
Datasets
Community
New
Datasets
New dataset
Name
Owner
Last Modified
Sales Pipeline
SPICE
Me
a few seconds ago
Web and Social Media Analytics
SPICE
Me
a few seconds ago
Business Review
SPICE
Me
a few seconds ago
People Overview
SPICE
Me
a few seconds ago

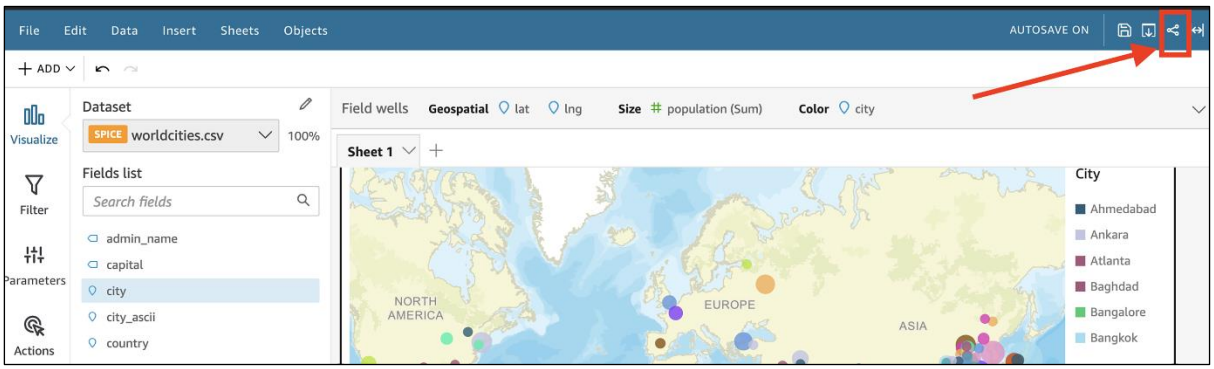

QuickSight
worldcities.csv analysis
ADD
AUTOSAVE ON
Visualize
Filter
Parameters
Actions
Themes
Settings
Community
Dataset
SPICE worldcities.csv
100%
Fields list
Search fields
admin_name
capital
city
city_ascii
country
id
iso2
iso3
lat
lng
population
1
Visual types
2
Field wells
Sheet 1
3
4
AutoGraph
Choose 1 or more fields and let QuickSight choose the most appropriate chart


Dataset
SPICE worldcities.csv
100%
Fields list
Search fields
admin_name
capital
city
city_ascii
country
id
iso2
iso3
lat
lng
population
Visual types
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

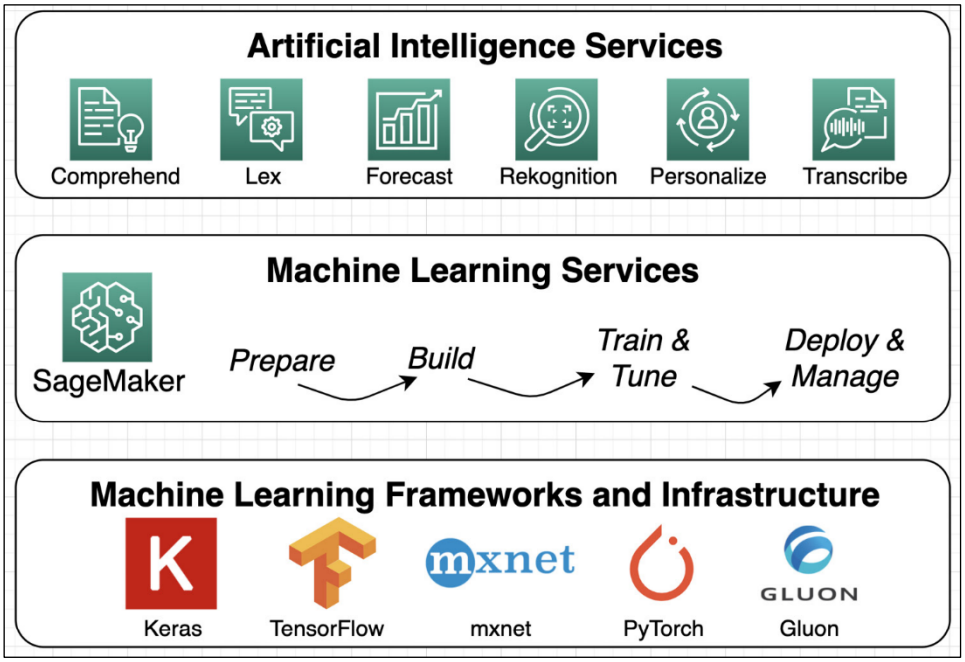

QuickSight
worldcities.csv analysis
ADD
Visualize
Filter
Parameters
Actions
Themes
Settings
Filters
No filters for the selected visual
ADD FILTER
Field wells
Geospatial
lat
lng
Size
population (Sum)
Color
city
Sheet 1
Sum of Population by Lat, Lng, and City
SHOWING TOP 5000 IN LAT, LNG AND TOP 4884 IN CITY
City
A Coruña
Aachen
Aalborg
Aaley
Aarhus
Aba
Abaetetuba
Abaeté
Abakan
Abbotsford
Abbottabad
Abengourou
Abeokuta
Aberdeen
Abertawe
Abhā
Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community


Edit filter
Applied to
Only this visual
population
Equals - none
Aggregation
Sum
Filter condition
Greater than or equal to
Use parameters
Minimum value
3000000
Null options
Include nulls
OR
ADD FILTER CONDITION
Note: There are limitations on how you can group filters.
Learn more
APPLY
DELETE FILTER


City Populations
Cities with a population of over 5 million people
SHOWING TOP 444 IN LAT, LNG AND TOP 444 IN CITY
NORTH AMERICA
EUROPE
ASIA
North Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
South Atlantic Ocean
6.455° N
3.3841° E
city Lagos
population 16,637,000
Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community
City
Ahmedabad
Ankara
Atlanta
Baghdad
Bangalore
Bangkok
Baoding
Beijing
Belo Horizo...
Bijie
Bogotá
Buenos Aires
Cairo
Cangzhou
Changde
Chattogram


File
Edit
Data
Insert
Sheets
Objects
AUTOSAVE ON
ADD
Visualize
Filter
Parameters
Actions
Dataset
SPICE worldcities.csv
100%
Fields list
Search fields
admin_name
capital
city
city_ascii
country
Field wells
Geospatial lat lng
Size population (Sum)
Color city
Sheet 1
NORTH AMERICA
EUROPE
ASIA
City
Ahmedabad
Ankara
Atlanta
Baghdad
Bangalore
Bangkok

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	($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	($212.85)	

Amazon SQS > Queues > Create queue

Create queue

Details

Type

Choose the queue type for your application or cloud infrastructure.

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

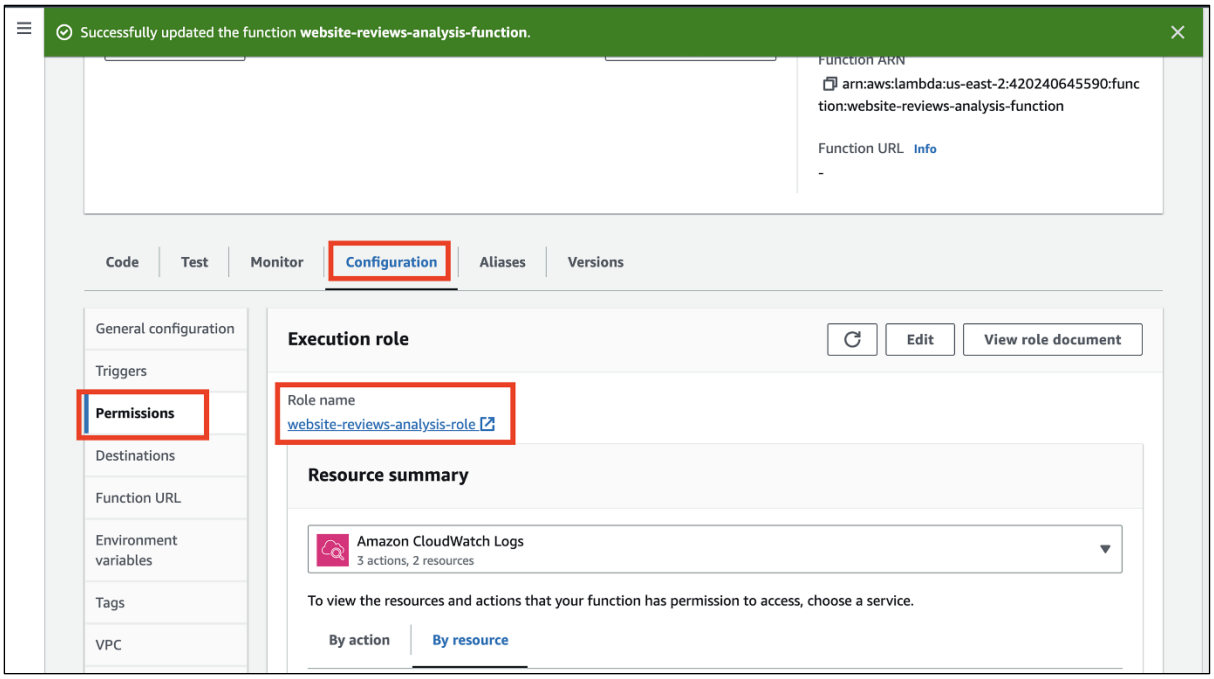
- First-in-first-out delivery
- Exactly-once processing

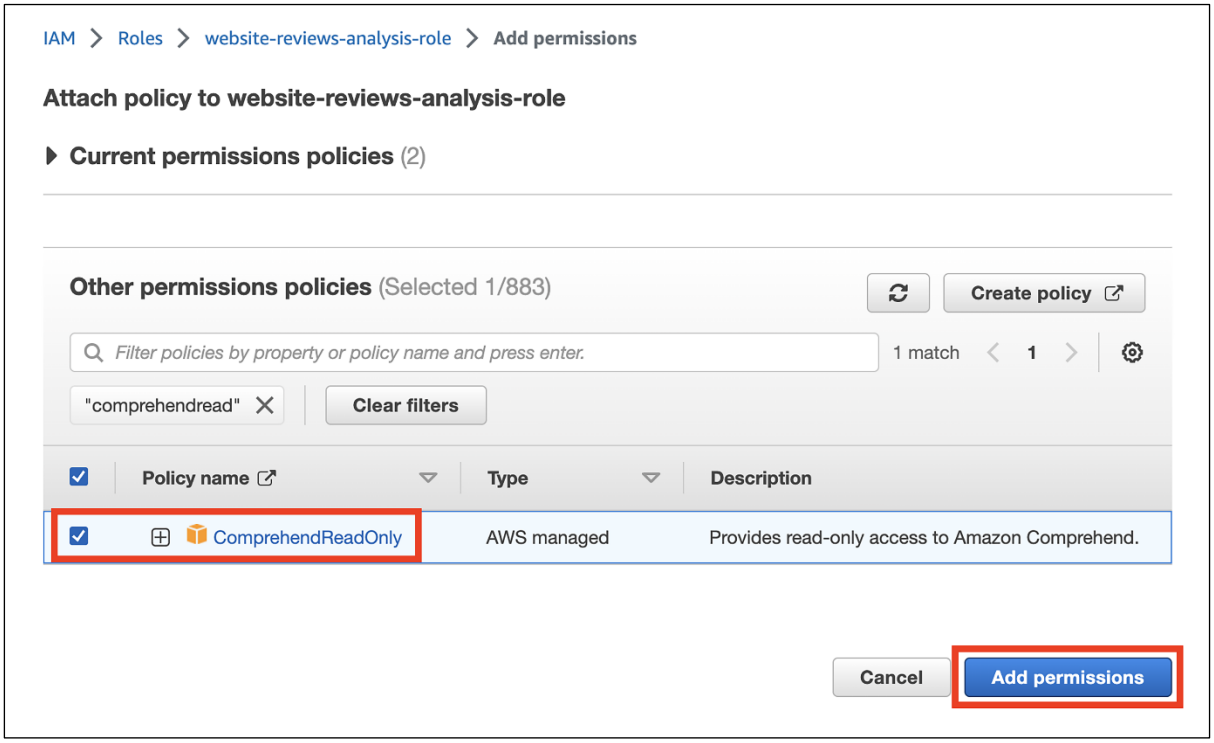
You can't change the queue type after you create a queue.

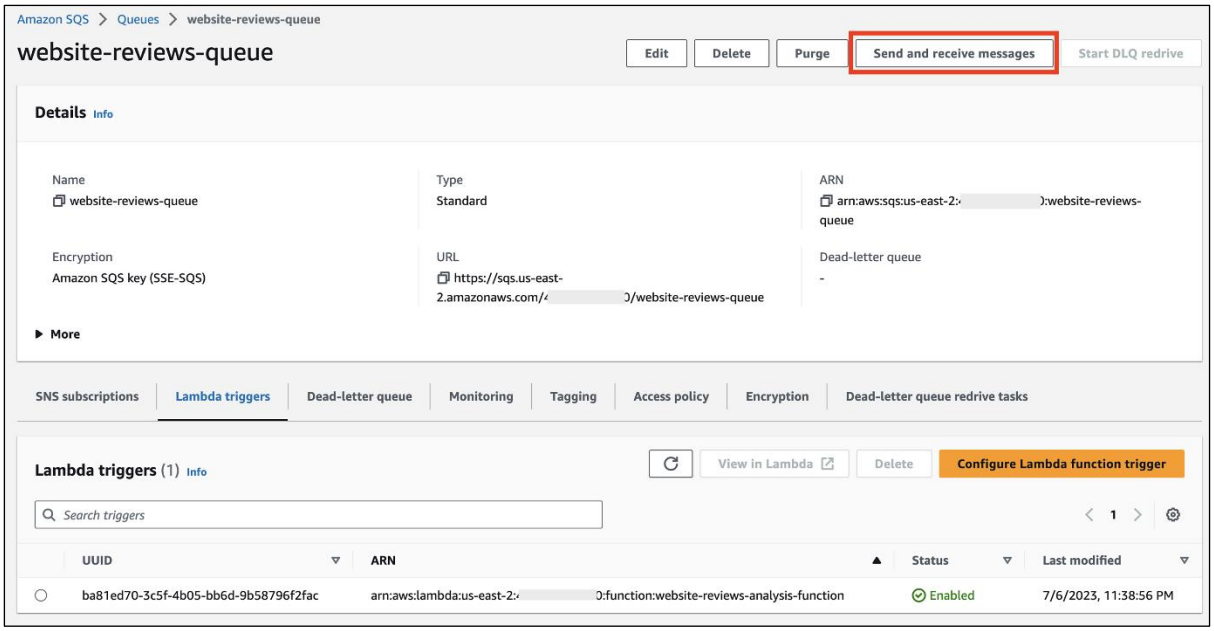
Name

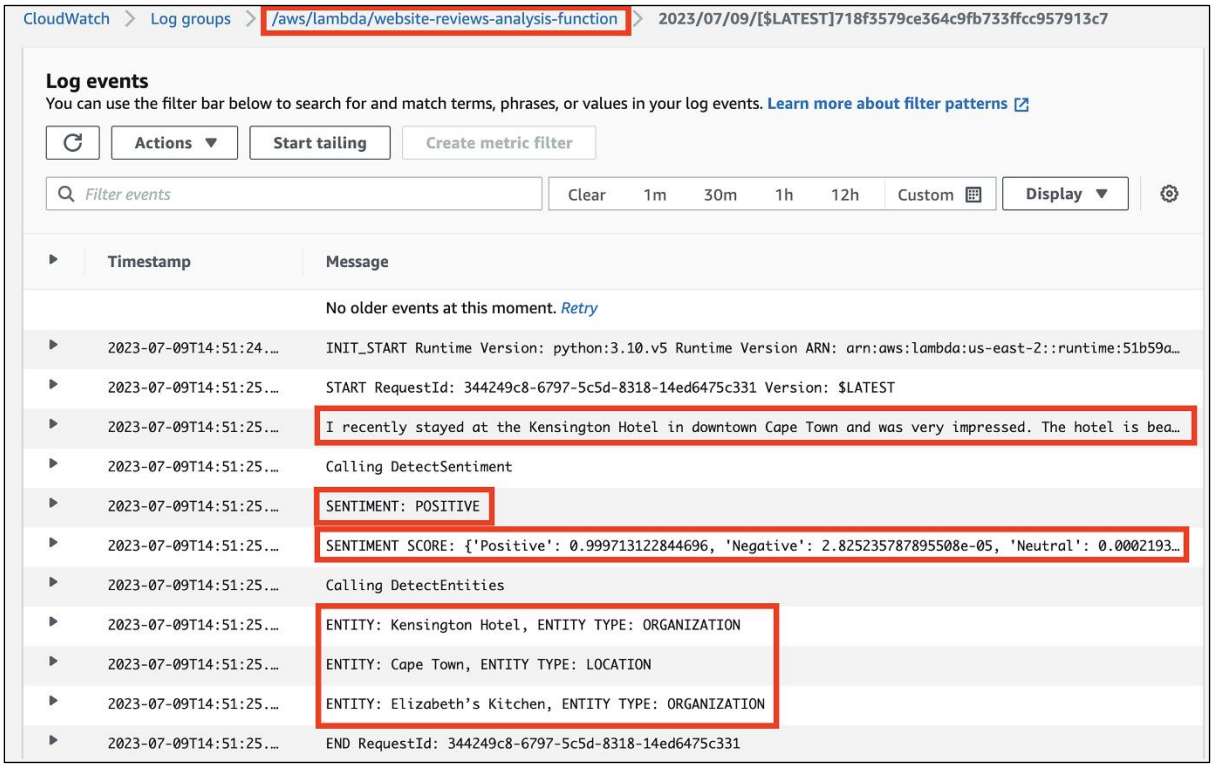
website-reviews-queue

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


Successfully updated the function website-reviews-analysis-function.
Function ARN
arn:aws:lambda:us-east-2:420240645590:function:website-reviews-analysis-function
Function URL Info
-
Code
Test
Monitor
Configuration
Aliases
Versions
General configuration
Triggers
Permissions
Destinations
Function URL
Environment variables
Tags
VPC
Execution role
Edit
View role document
Role name
website-reviews-analysis-role
Resource summary
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

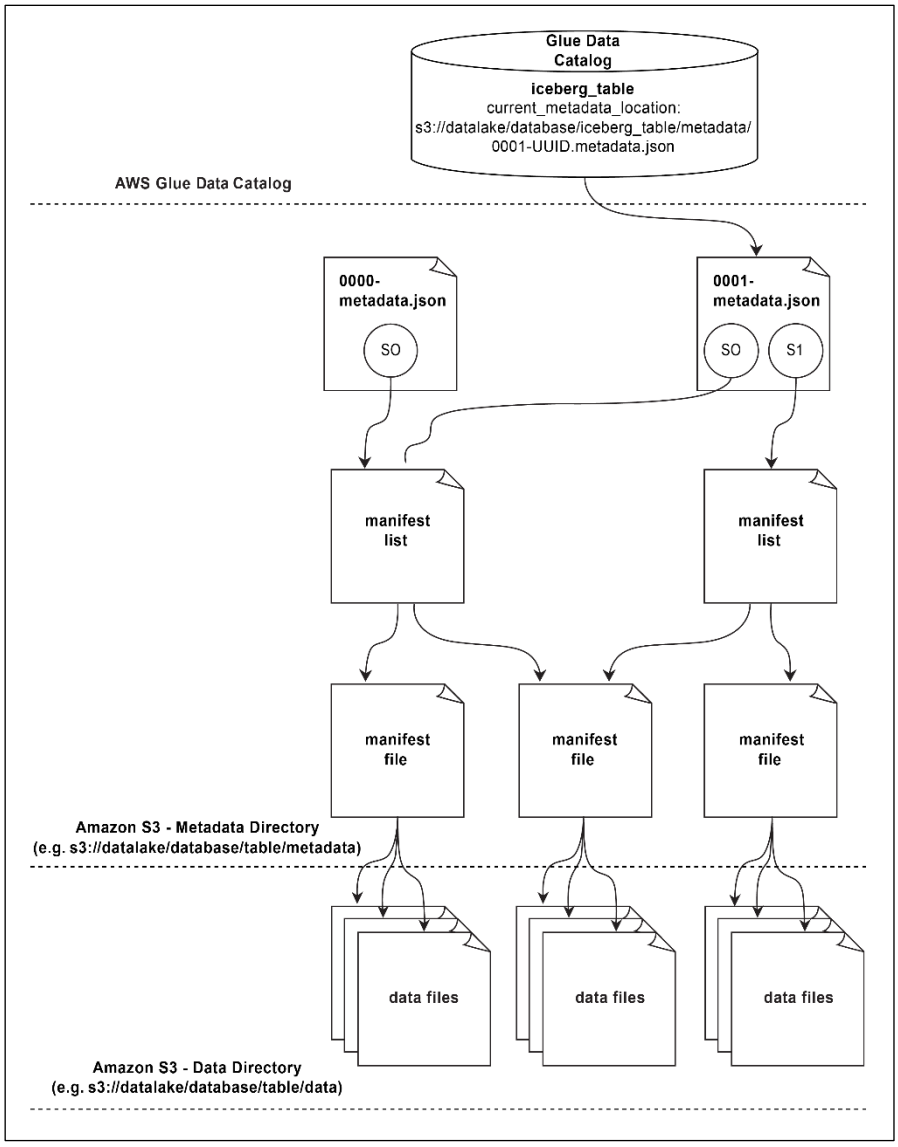

IAM > Roles > website-reviews-analysis-role > Add permissions
Attach policy to website-reviews-analysis-role
Current permissions policies (2)
Other permissions policies (Selected 1/883)
Create policy
Filter policies by property or policy name and press enter.
1 match
1
"comprehendread"
Clear filters
Policy name
Type
Description
ComprehendReadOnly
AWS managed
Provides read-only access to Amazon Comprehend.
Cancel
Add permissions

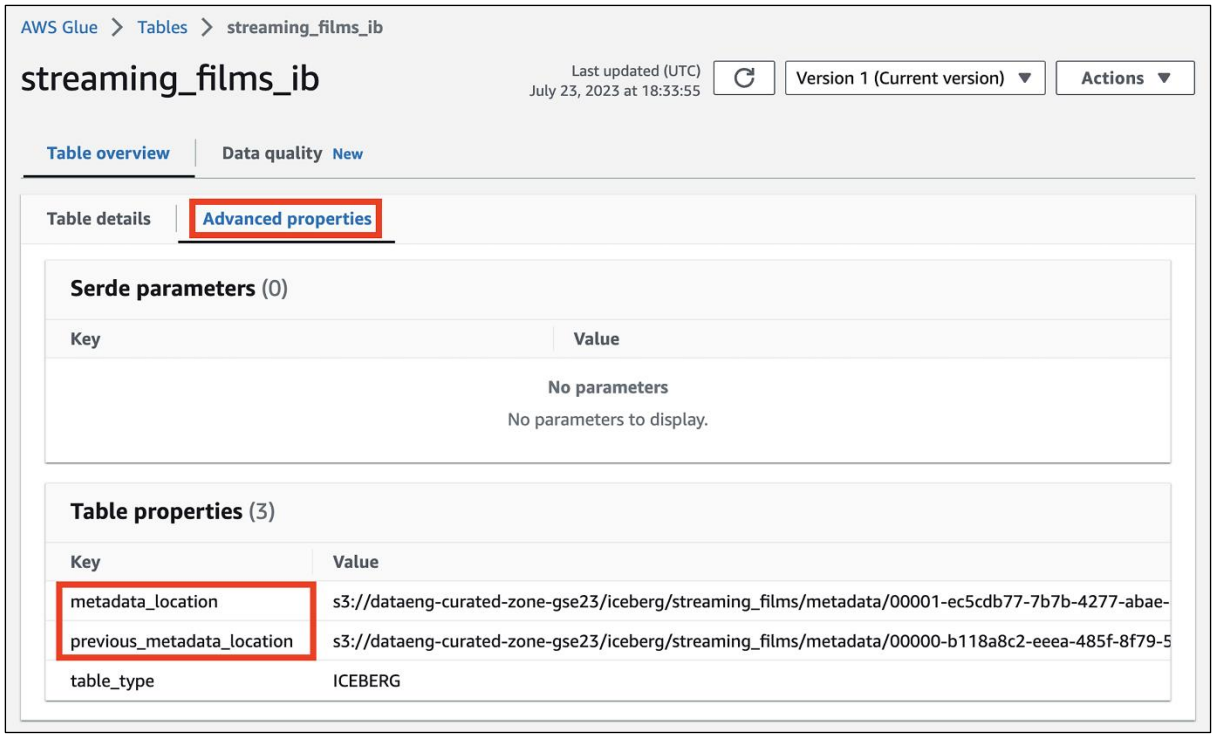

Amazon SQS > Queues > website-reviews-queue
website-reviews-queue
Edit
Delete
Purge
Send and receive messages
Start DLQ redrive
Details Info
Name
website-reviews-queue
Type
Standard
ARN
arn:aws:sqs:us-east-2:
:website-reviews-queue
Encryption
Amazon SQS key (SSE-SQS)
URL
https://sqs.us-east-2.amazonaws.com/
/website-reviews-queue
Dead-letter queue
-
More
SNS subscriptions
Lambda triggers
Dead-letter queue
Monitoring
Tagging
Access policy
Encryption
Dead-letter queue redrive tasks
Lambda triggers (1) Info
View in Lambda
Delete
Configure Lambda function trigger
Search triggers
UUID
ARN
Status
Last modified
ba81ed70-3c5f-4b05-bb6d-9b58796f2fac
arn:aws:lambda:us-east-2:
:function:website-reviews-analysis-function
Enabled
7/6/2023, 11:38:56 PM

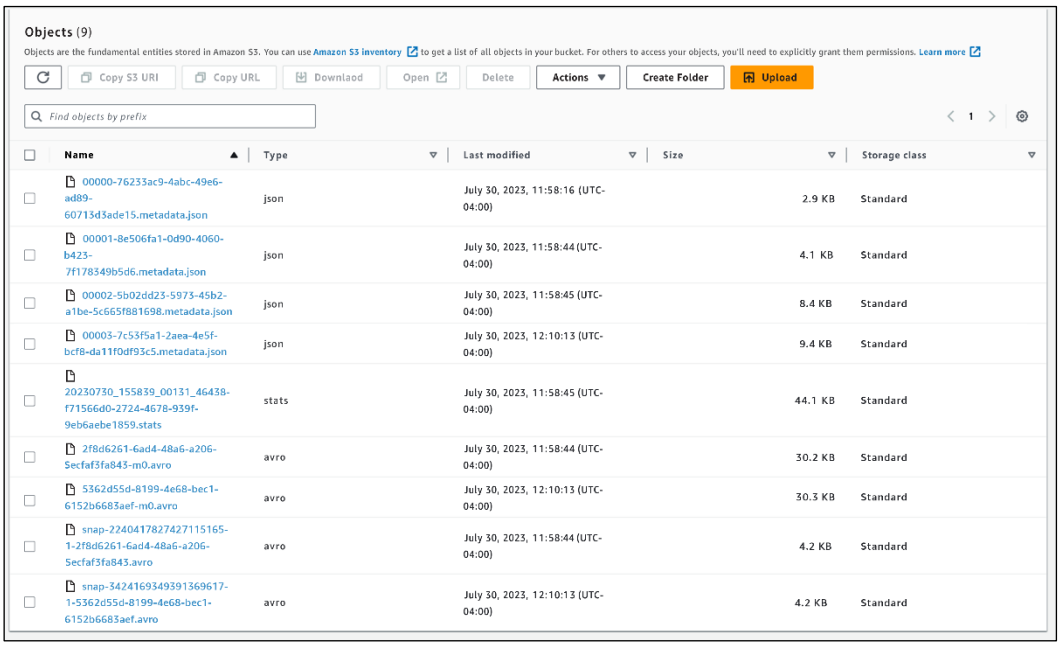

CloudWatch > Log groups > /aws/lambda/website-reviews-analysis-function > 2023/07/09/[$LATEST]718f3579ce364c9fb733ffcc957913c7
Log events
You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns
Actions
Start tailing
Create metric filter
Filter events
Clear
1m
30m
1h
12h
Custom
Display
Timestamp
Message
No older events at this moment. Retry
2023-07-09T14:51:24.…
INIT_START Runtime Version: python:3.10.v5 Runtime Version ARN: arn:aws:lambda:us-east-2::runtime:51b59a…
2023-07-09T14:51:25.…
START RequestId: 344249c8-6797-5c5d-8318-14ed6475c331 Version: $LATEST
2023-07-09T14:51:25.…
I recently stayed at the Kensington Hotel in downtown Cape Town and was very impressed. The hotel is bea…
2023-07-09T14:51:25.…
Calling DetectSentiment
2023-07-09T14:51:25.…
SENTIMENT: POSITIVE
2023-07-09T14:51:25.…
SENTIMENT SCORE: {'Positive': 0.999713122844696, 'Negative': 2.825235787895508e-05, 'Neutral': 0.0002193…
2023-07-09T14:51:25.…
Calling DetectEntities
2023-07-09T14:51:25.…
ENTITY: Kensington Hotel, ENTITY TYPE: ORGANIZATION
2023-07-09T14:51:25.…
ENTITY: Cape Town, ENTITY TYPE: LOCATION
2023-07-09T14:51:25.…
ENTITY: Elizabeth's Kitchen, ENTITY TYPE: ORGANIZATION
2023-07-09T14:51:25.…
END RequestId: 344249c8-6797-5c5d-8318-14ed6475c331

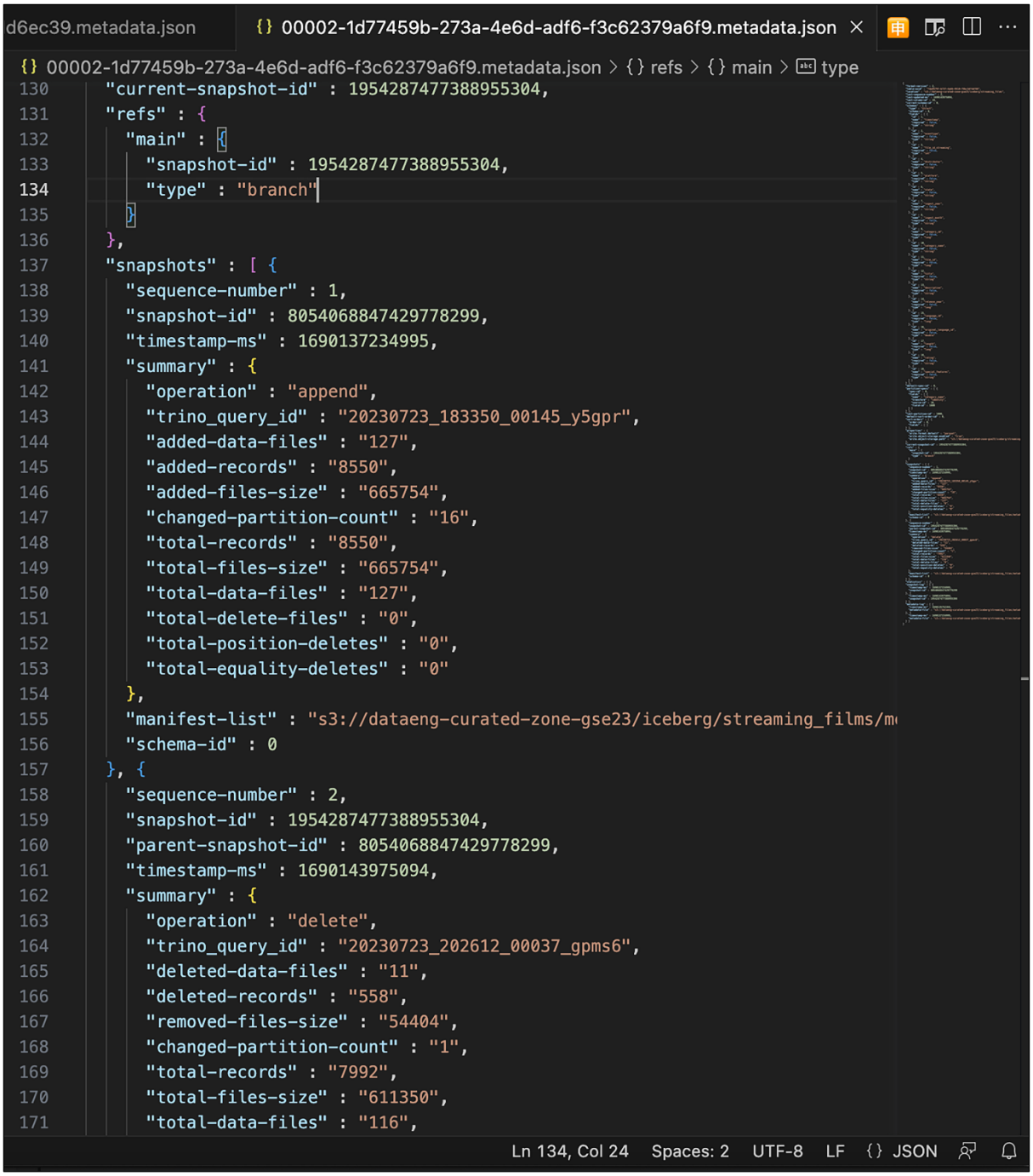
Chapter 14: Building Transactional Data Lakes

	Copy-On-Write (COW)	Merge-On-Read (MOR)
Action on record update/delete	New version of affected file is created containing newly updated records, or skipping deleted records	Deleted and updated rows are written to a deletion tracking file, and updated records are written to a new file
Action when table is read	The metadata for the table tracks the files that have the latest data, and only those files are read	The metadata for the table tracks the original file, the file tracking deleted records, and files containing new/updated data. The query engine needs to merge each of these files to query the data.
Optimized for reads or writes?	Optimized for table reads	Optimized for table writes



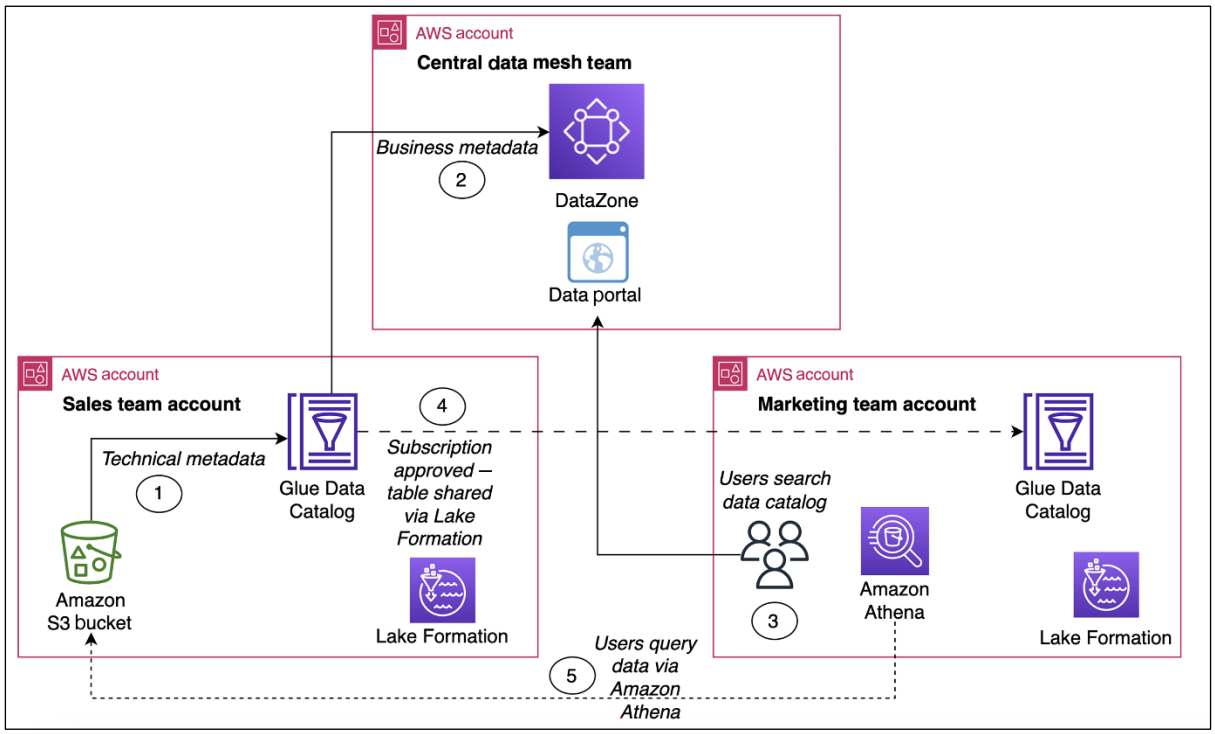

AWS Glue > Tables > streaming_films_ib
streaming_films_ib
Last updated (UTC)
July 23, 2023 at 18:33:55
Version 1 (Current version)
Actions
Table overview
Data quality New
Table details
Advanced properties
Serde parameters (0)
Key
Value
No parameters
No parameters to display.
Table properties (3)
Key
Value
metadata_location
s3://dataeng-curated-zone-gse23/iceberg/streaming_films/metadata/00001-ec5cdb77-7b7b-4277-abae-
previous_metadata_location
s3://dataeng-curated-zone-gse23/iceberg/streaming_films/metadata/00000-b118a8c2-eeea-485f-8f79-5
table_type
ICEBERG

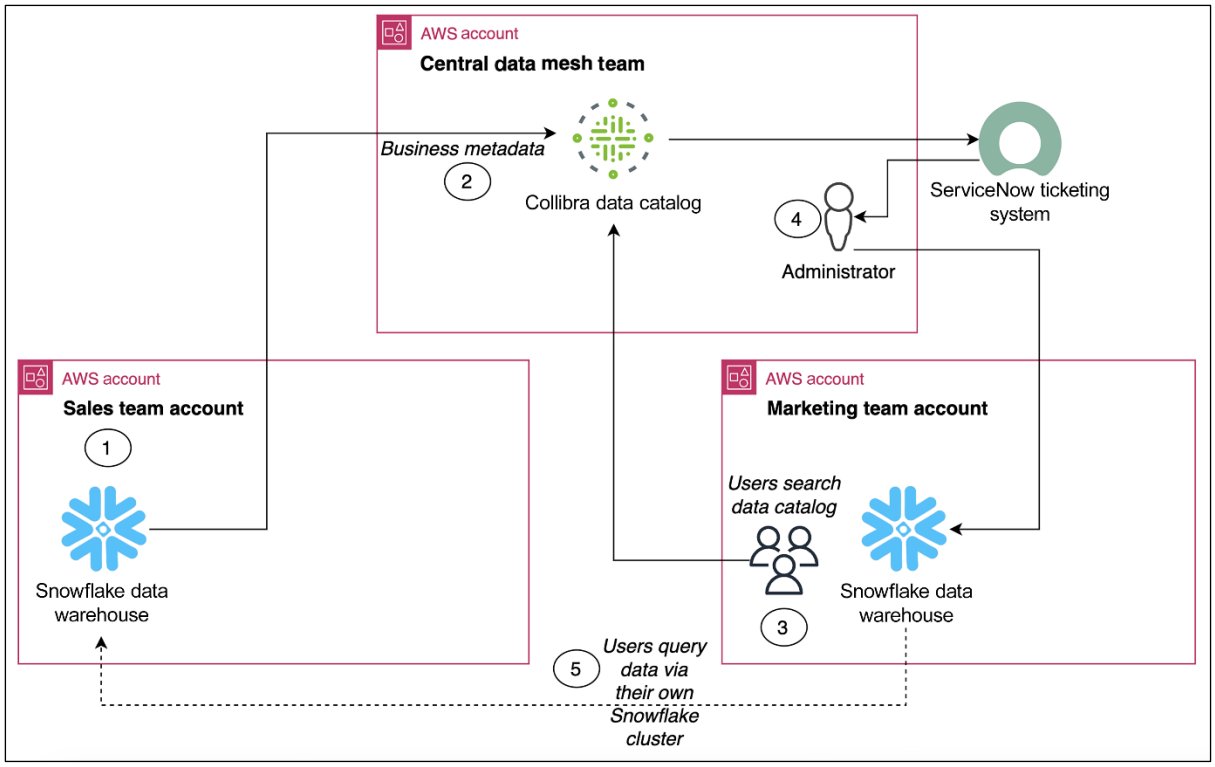

Objects (9)
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create Folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
00000-76233ac9-4abc-49e6-ad89-60713d3ade15.metadata.json
json
July 30, 2023, 11:58:16 (UTC-04:00)
2.9 KB
Standard
00001-8e506fa1-0d90-4060-b423-7f178349b5d6.metadata.json
json
July 30, 2023, 11:58:44 (UTC-04:00)
4.1 KB
Standard
00002-5b02dd23-5973-45b2-a1be-5c665f881698.metadata.json
json
July 30, 2023, 11:58:45 (UTC-04:00)
8.4 KB
Standard
00003-7c53f5a1-2aea-4e5f-bcf8-da11f0df93c5.metadata.json
json
July 30, 2023, 12:10:13 (UTC-04:00)
9.4 KB
Standard
20230730_155839_00131_46438-f71566d0-2724-4678-939f-9eb6aebe1859.stats
stats
July 30, 2023, 11:58:45 (UTC-04:00)
44.1 KB
Standard
2f8d6261-6ad4-48a6-a206-5ecfaf3fa843-m0.avro
avro
July 30, 2023, 11:58:44 (UTC-04:00)
30.2 KB
Standard
5362d55d-8199-4e68-bec1-6152b6683aef-m0.avro
avro
July 30, 2023, 12:10:13 (UTC-04:00)
30.3 KB
Standard
snap-2240417827427115165-1-2f8d6261-6ad4-48a6-a206-5ecfaf3fa843.avro
avro
July 30, 2023, 11:58:44 (UTC-04:00)
4.2 KB
Standard
snap-3424169349391369617-1-5362d55d-8199-4e68-bec1-6152b6683aef.avro
avro
July 30, 2023, 12:10:13 (UTC-04:00)
4.2 KB
Standard

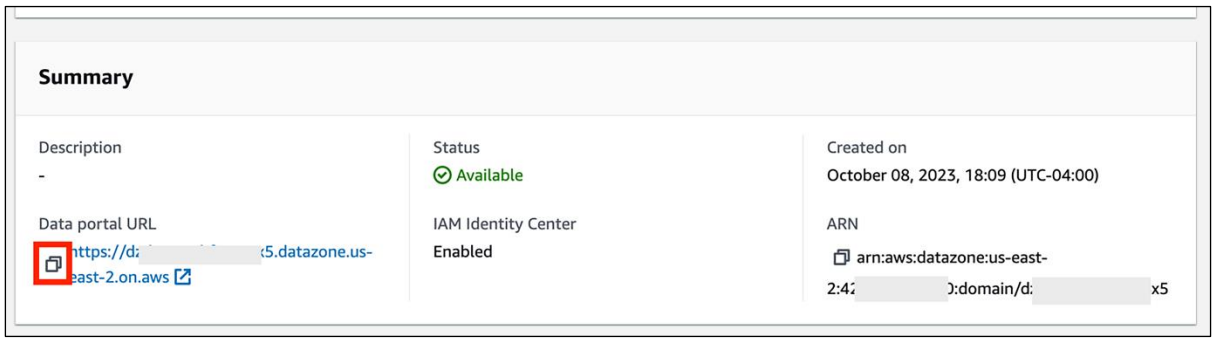

d6ec39.metadata.json
{} 00002-1d77459b-273a-4e6d-adf6-f3c62379a6f9.metadata.json
{} 00002-1d77459b-273a-4e6d-adf6-f3c62379a6f9.metadata.json > {} refs > {} main > type
"current-snapshot-id" : 1954287477388955304,
"refs" : {
"main" : {
"snapshot-id" : 1954287477388955304,
"type" : "branch"
}
},
"snapshots" : [{
"sequence-number" : 1,
"snapshot-id" : 8054068847429778299,
"timestamp-ms" : 1690137234995,
"summary" : {
"operation" : "append",
"trino_query_id" : "20230723_183350_00145_y5gpr",
"added-data-files" : "127",
"added-records" : "8550",
"added-files-size" : "665754",
"changed-partition-count" : "16",
"total-records" : "8550",
"total-files-size" : "665754",
"total-data-files" : "127",
"total-delete-files" : "0",
"total-position-deletes" : "0",
"total-equality-deletes" : "0"
},
"manifest-list" : "s3://dataeng-curated-zone-gse23/iceberg/streaming_films/m
"schema-id" : 0
}, {
"sequence-number" : 2,
"snapshot-id" : 1954287477388955304,
"parent-snapshot-id" : 8054068847429778299,
"timestamp-ms" : 1690143975094,
"summary" : {
"operation" : "delete",
"trino_query_id" : "20230723_202612_00037_gpms6",
"deleted-data-files" : "11",
"deleted-records" : "558",
"removed-files-size" : "54404",
"changed-partition-count" : "1",
"total-records" : "7992",
"total-files-size" : "611350",
"total-data-files" : "116",
Ln 134, Col 24
Spaces: 2
UTF-8
LF
{} JSON

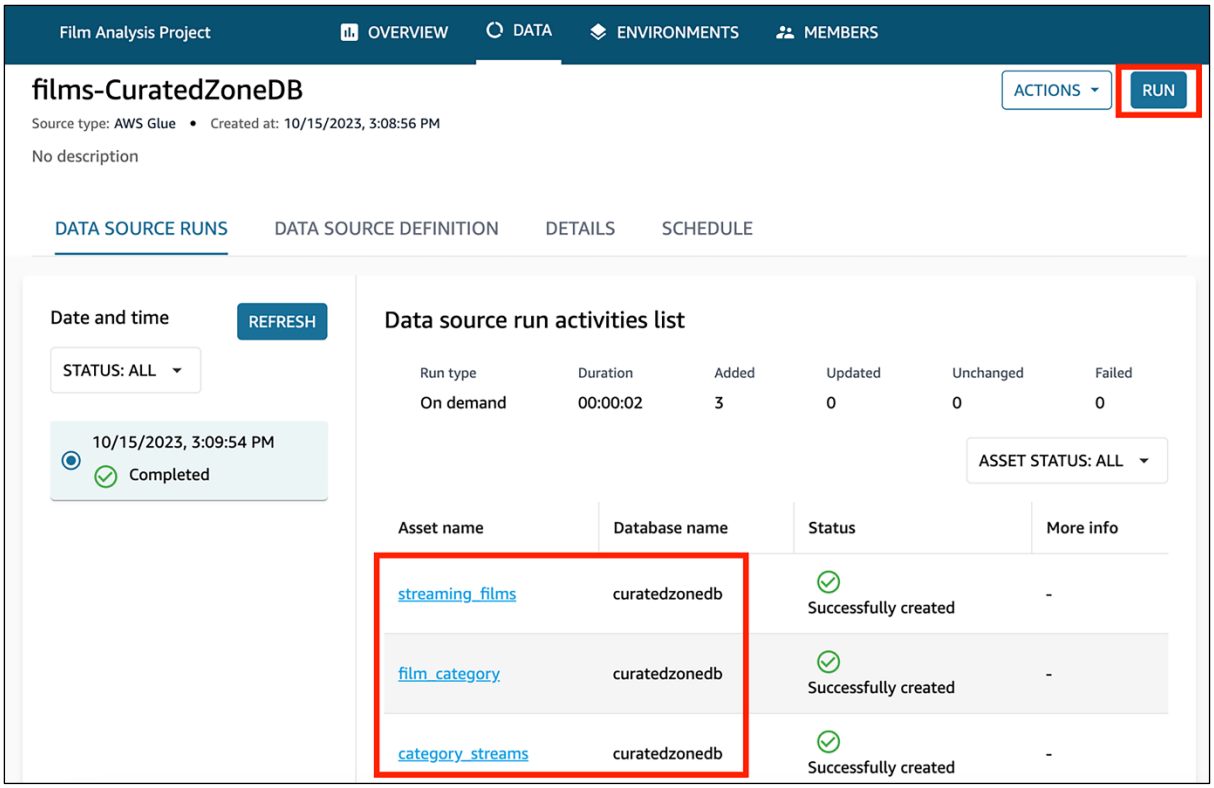
Chapter 15: Implementing a Data Mesh Strategy

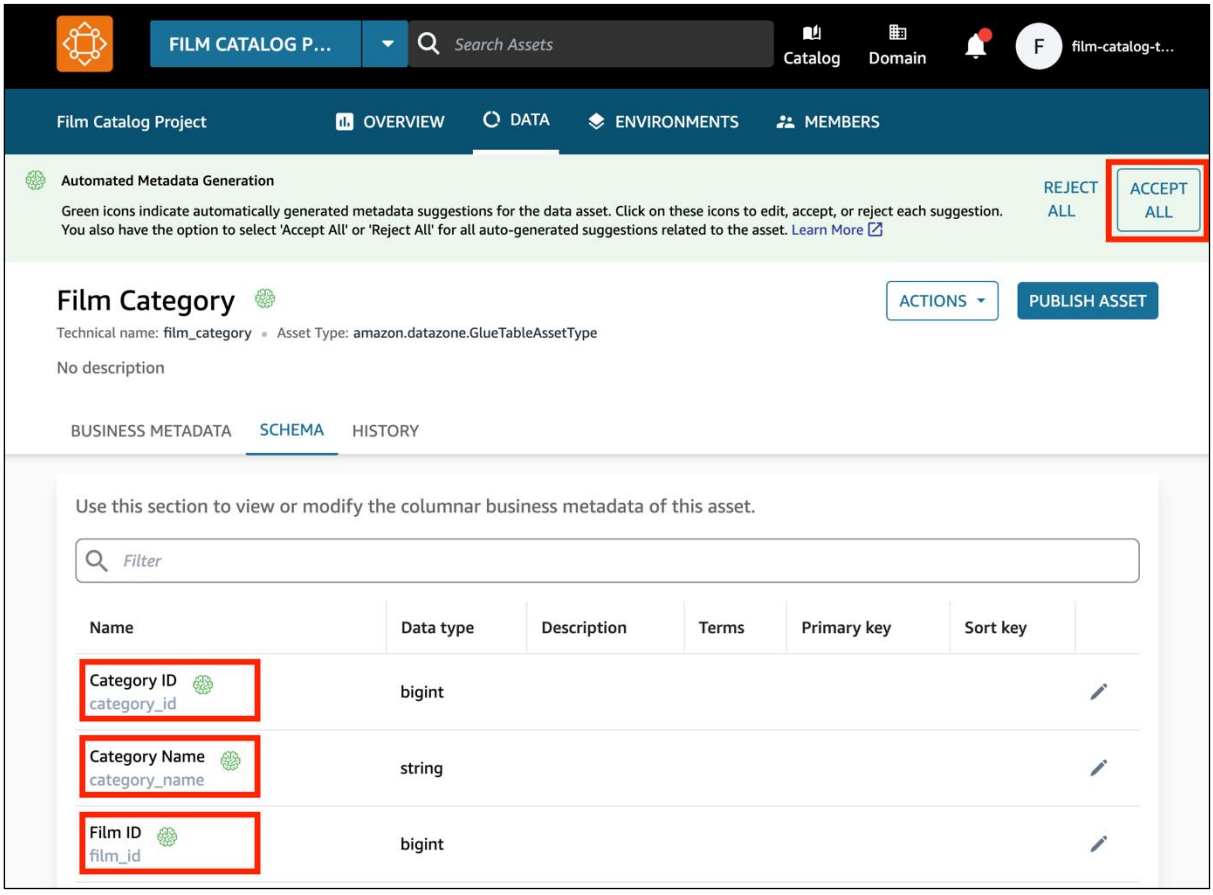
Field Name	Field Type	Description
Sales Region	*Country* Business Glossary	The sales region for this dataset
Sales Quarter	*Quarter* Business Glossary	The calendar quarter that this dataset covers
Dataset Support	String	Provide the email address to use for queries about this dataset
Last Audit Date	Date	Provide the date that this dataset was last audited



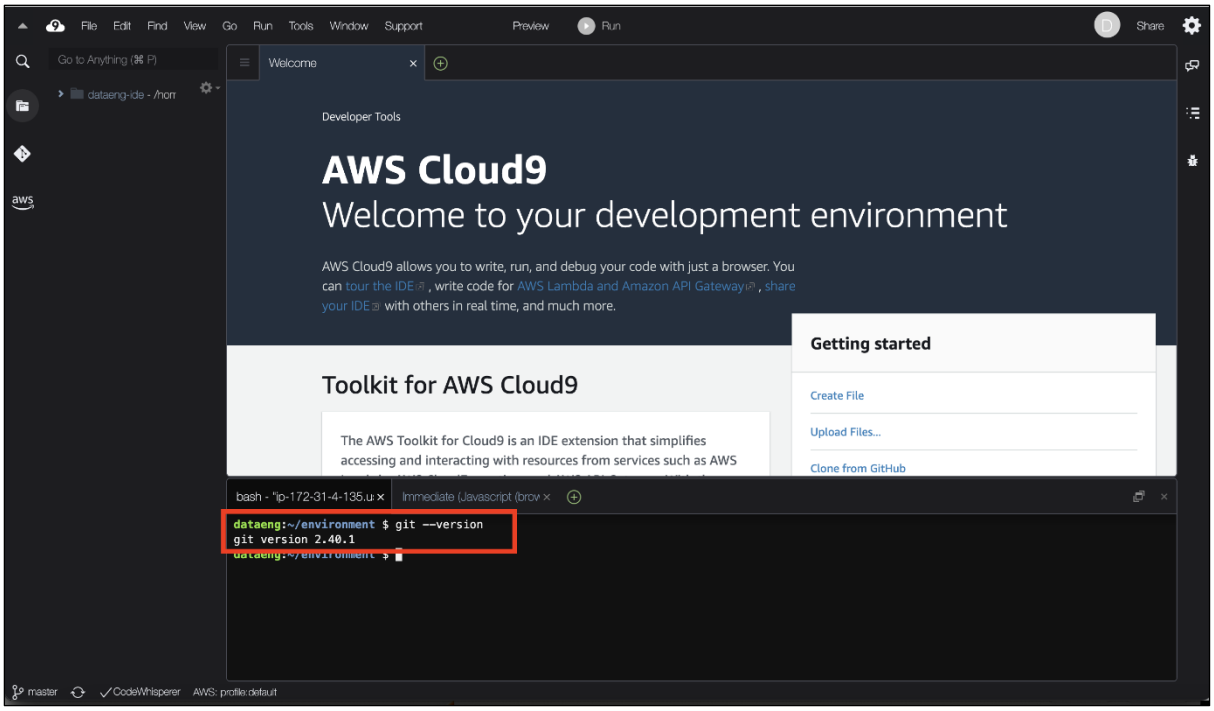

AWS account
Central data mesh team
Business metadata
2
Collibra data catalog
ServiceNow ticketing system
4
Administrator
AWS account
Sales team account
1
Snowflake data warehouse
AWS account
Marketing team account
Users search data catalog
3
Snowflake data warehouse
5
Users query data via their own Snowflake cluster

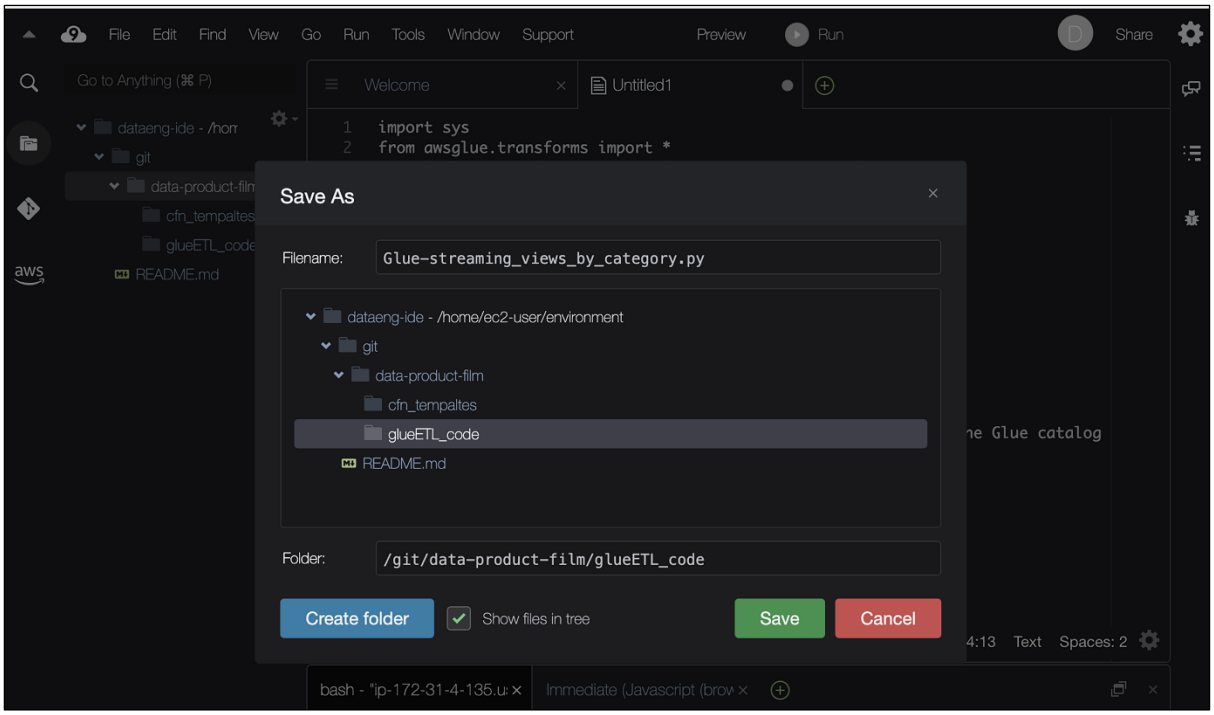

Summary
Description
-
Status
Available
Created on
October 08, 2023, 18:09 (UTC-04:00)
Data portal URL
IAM Identity Center
Enabled
ARN
arn:aws:datazone:us-east-


Film Analysis Project
OVERVIEW
DATA
ENVIRONMENTS
MEMBERS
films-CuratedZoneDB
ACTIONS
RUN
Source type: AWS Glue • Created at: 10/15/2023, 3:08:56 PM
No description
DATA SOURCE RUNS
DATA SOURCE DEFINITION
DETAILS
SCHEDULE
Date and time
REFRESH
STATUS: ALL
10/15/2023, 3:09:54 PM
Completed
Data source run activities list
Run type
Duration
Added
Updated
Unchanged
Failed
On demand
00:00:02
3
0
0
0
ASSET STATUS: ALL
Asset name
Database name
Status
More info
streaming_films
curatedzonedb
Successfully created
-
film_category
curatedzonedb
Successfully created
-
category_streams
curatedzonedb
Successfully created
-

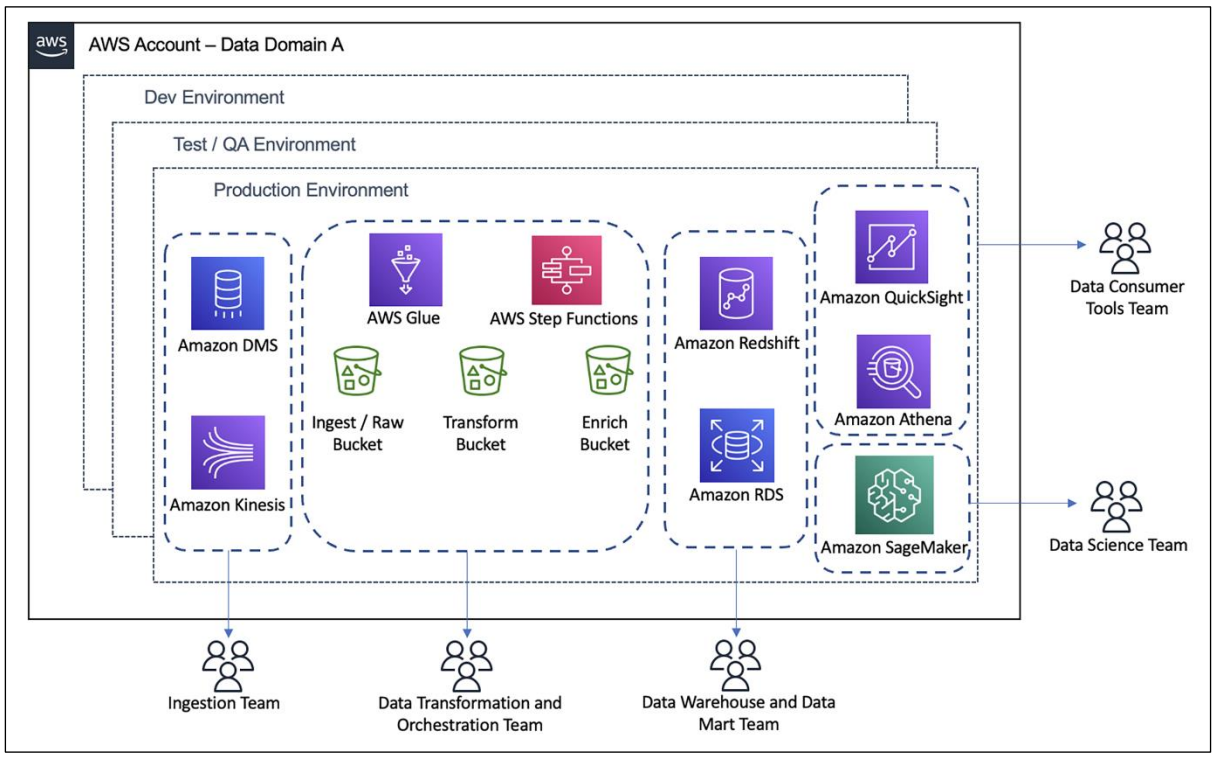

FILM CATALOG P...
Search Assets
Catalog
Domain
F
film-catalog-t...
Film Catalog Project
OVERVIEW
DATA
ENVIRONMENTS
MEMBERS
Automated Metadata Generation
Green icons indicate automatically generated metadata suggestions for the data asset. Click on these icons to edit, accept, or reject each suggestion.
You also have the option to select 'Accept All' or 'Reject All' for all auto-generated suggestions related to the asset. Learn More
REJECT ALL
ACCEPT ALL
Film Category
ACTIONS
PUBLISH ASSET
Technical name: film_category • Asset Type: amazon.datazone.GlueTableAssetType
No description
BUSINESS METADATA
SCHEMA
HISTORY
Use this section to view or modify the columnar business metadata of this asset.
Filter
Name
Data type
Description
Terms
Primary key
Sort key
Category ID
category_id
bigint
Category Name
category_name
string
Film ID
film_id
bigint

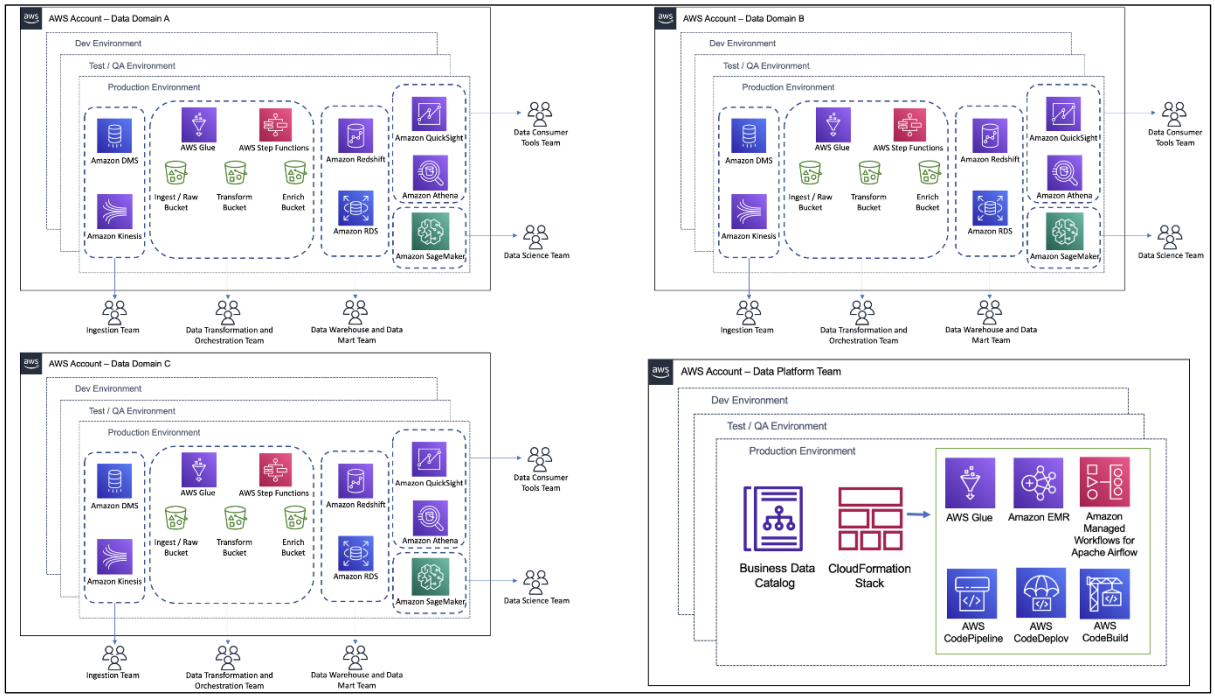
Chapter 16: Building a Modern Data Platform on AWS

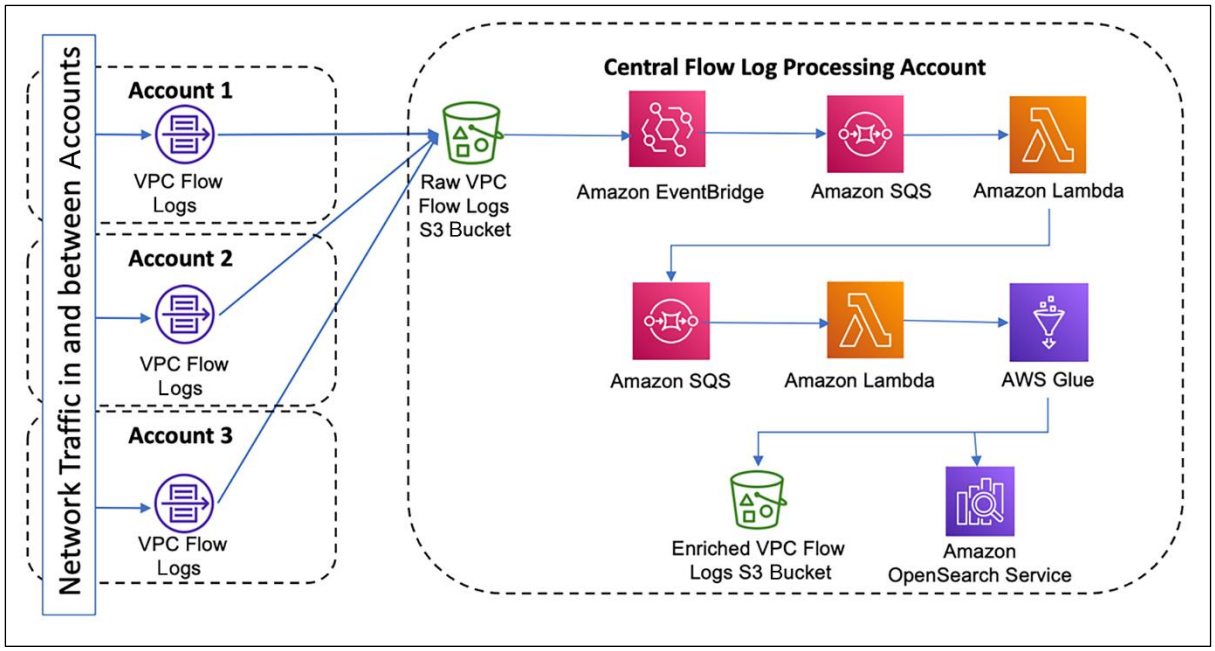


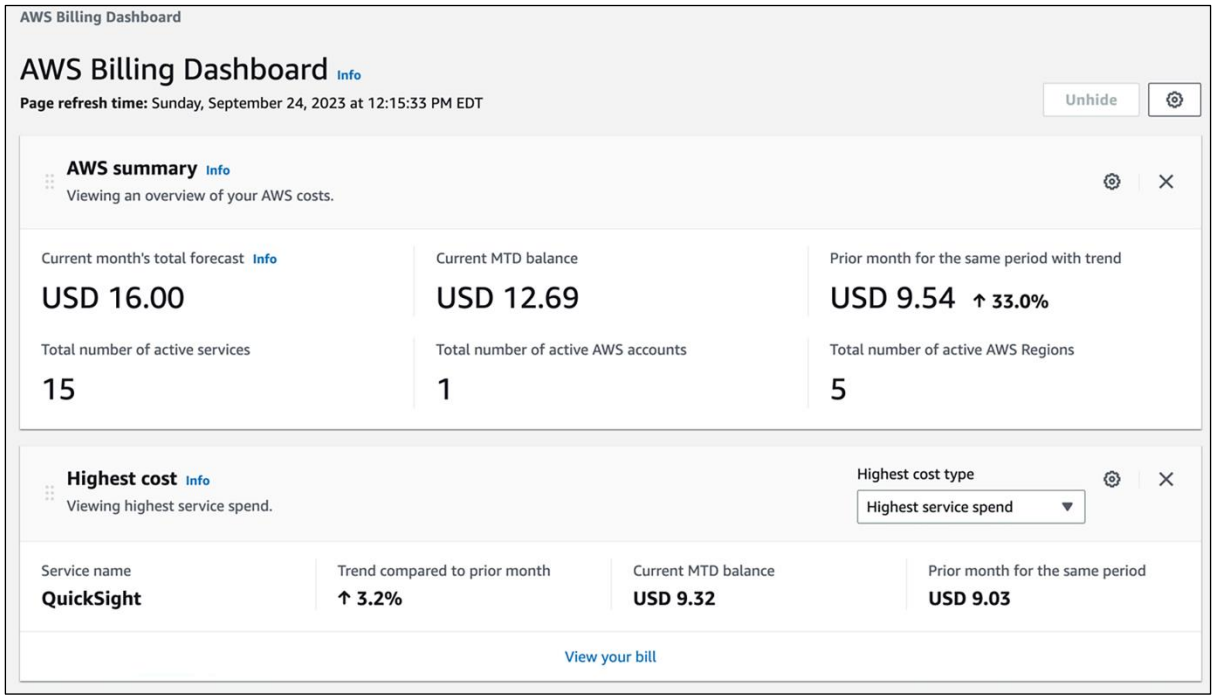


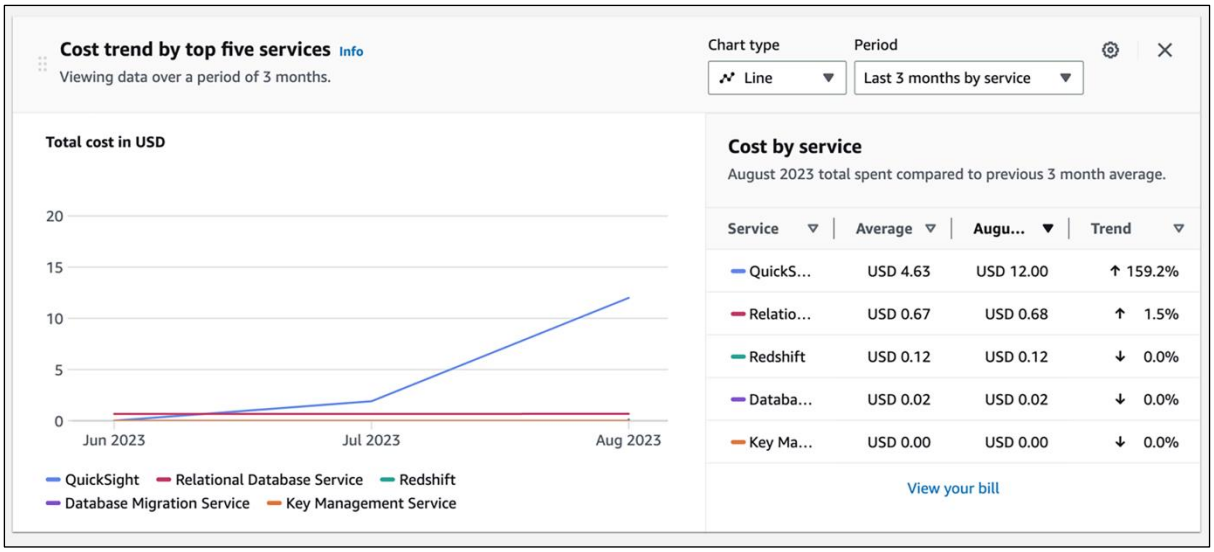
Chapter 17: Wrapping Up the First Part of Your Learning Journey

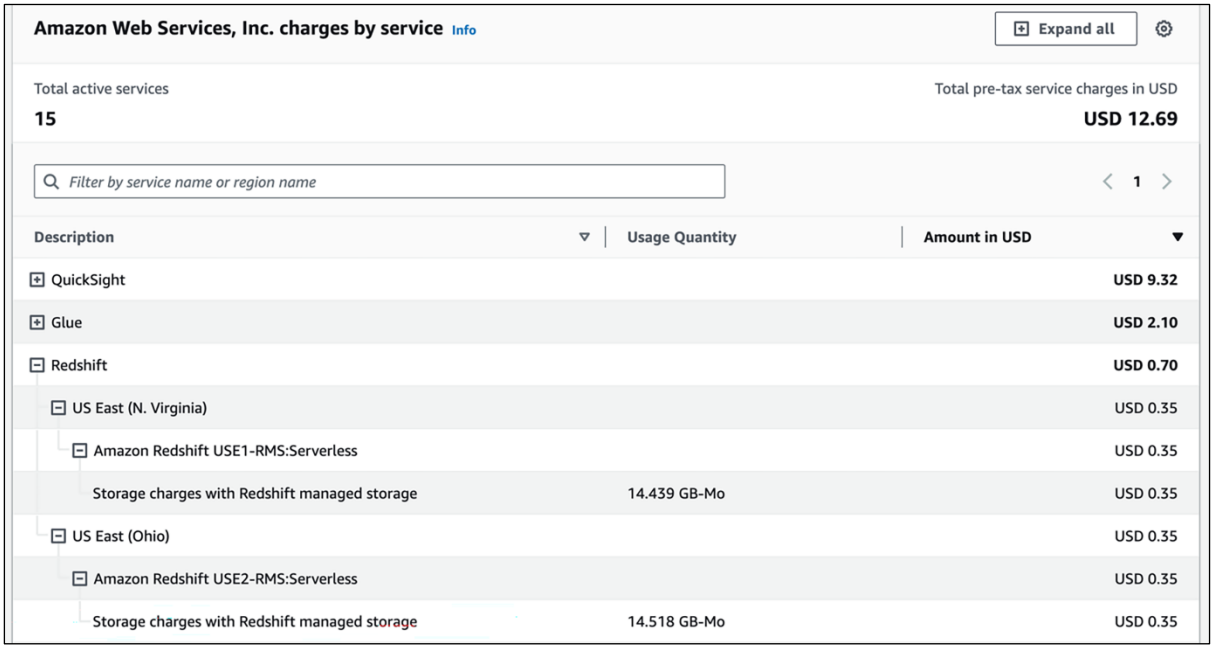


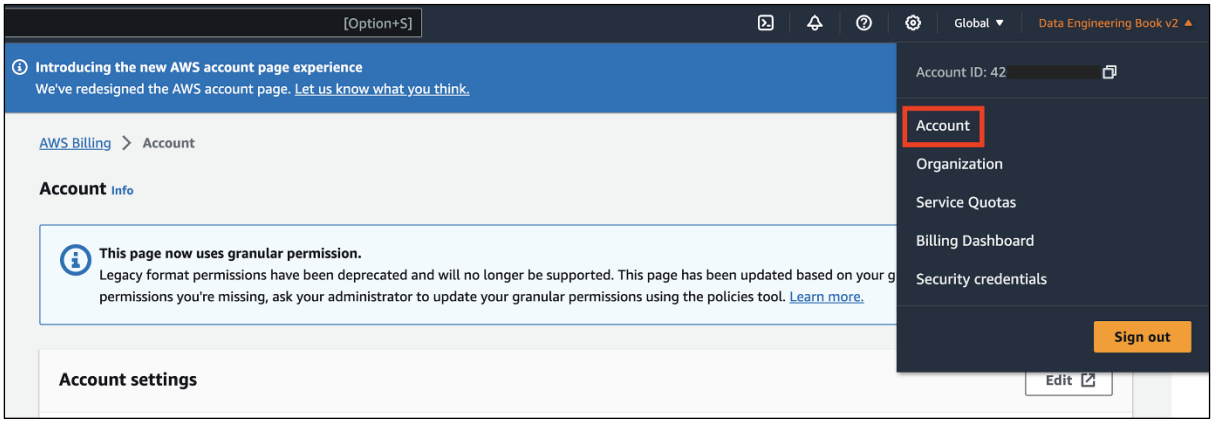



Network Traffic in and between Accounts
Account 1
VPC Flow Logs
Account 2
VPC Flow Logs
Account 3
VPC Flow Logs
Central Flow Log Processing Account
Raw VPC Flow Logs S3 Bucket
Amazon EventBridge
Amazon SQS
Amazon Lambda
Amazon SQS
Amazon Lambda
AWS Glue
Enriched VPC Flow Logs S3 Bucket
Amazon OpenSearch Service


AWS Billing Dashboard
AWS Billing Dashboard Info
Page refresh time: Sunday, September 24, 2023 at 12:15:33 PM EDT
Unhide
AWS summary Info
Viewing an overview of your AWS costs.
Current month's total forecast Info
USD 16.00
Current MTD balance
USD 12.69
Prior month for the same period with trend
USD 9.54 ↑ 33.0%
Total number of active services
15
Total number of active AWS accounts
1
Total number of active AWS Regions
5
Highest cost Info
Viewing highest service spend.
Highest cost type
Highest service spend
Service name
QuickSight
Trend compared to prior month
↑ 3.2%
Current MTD balance
USD 9.32
Prior month for the same period
USD 9.03
View your bill


Cost trend by top five services Info
Viewing data over a period of 3 months.
Chart type
Line
Period
Last 3 months by service
Total cost in USD
20
15
10
5
0
Jun 2023
Jul 2023
Aug 2023
QuickSight
Relational Database Service
Redshift
Database Migration Service
Key Management Service
Cost by service
August 2023 total spent compared to previous 3 month average.
Service
Average
Augu...
Trend
QuickS... USD 4.63 USD 12.00 ↑ 159.2%
Relatio... USD 0.67 USD 0.68 ↑ 1.5%
Redshift USD 0.12 USD 0.12 ↓ 0.0%
Databa... USD 0.02 USD 0.02 ↓ 0.0%
Key Ma... USD 0.00 USD 0.00 ↓ 0.0%
View your bill


Amazon Web Services, Inc. charges by service Info
Expand all
Total active services
15
Total pre-tax service charges in USD
USD 12.69
Filter by service name or region name
1
Description
Usage Quantity
Amount in USD
QuickSight USD 9.32
Glue USD 2.10
Redshift USD 0.70
US East (N. Virginia) USD 0.35
Amazon Redshift USE1-RMS:Serverless USD 0.35
Storage charges with Redshift managed storage 14.439 GB-Mo USD 0.35
US East (Ohio) USD 0.35
Amazon Redshift USE2-RMS:Serverless USD 0.35
Storage charges with Redshift managed storage 14.518 GB-Mo USD 0.35


[Option+S]
Global
Data Engineering Book v2
Introducing the new AWS account page experience
We've redesigned the AWS account page. Let us know what you think.
Account ID: 42
Account
Organization
Service Quotas
Billing Dashboard
Security credentials
Sign out
AWS Billing > Account
Account Info
This page now uses granular permission.
Legacy format permissions have been deprecated and will no longer be supported. This page has been updated based on your g
permissions you're missing, ask your administrator to update your granular permissions using the policies tool. Learn more.
Account settings
Edit

Close account

Please review this important account closure guidance. Specifically:

- **Agreement termination:** Closing your account will serve as your notice of termination of the AWS Customer Agreement (or any other AWS agreement governing this account) for this account.
- **Billing:** You remain responsible for all outstanding fees and charges, including this month's usage and active subscriptions (such as Reserved Instances).
- **Reactivation:** You may reopen your AWS account for 90 days after closure. If you reopen your account, you may be charged for any active resources. After 90 days your account will be permanently closed, any remaining content will be deleted, and unused credits will be lost.
- **GovCloud:** Closing this account will also close any linked GovCloud accounts.

If you are experiencing unexpected charges, review how to troubleshoot common changes, including Free Tier changes, unwanted resources and unauthorized activity, without closing your account.

Close account