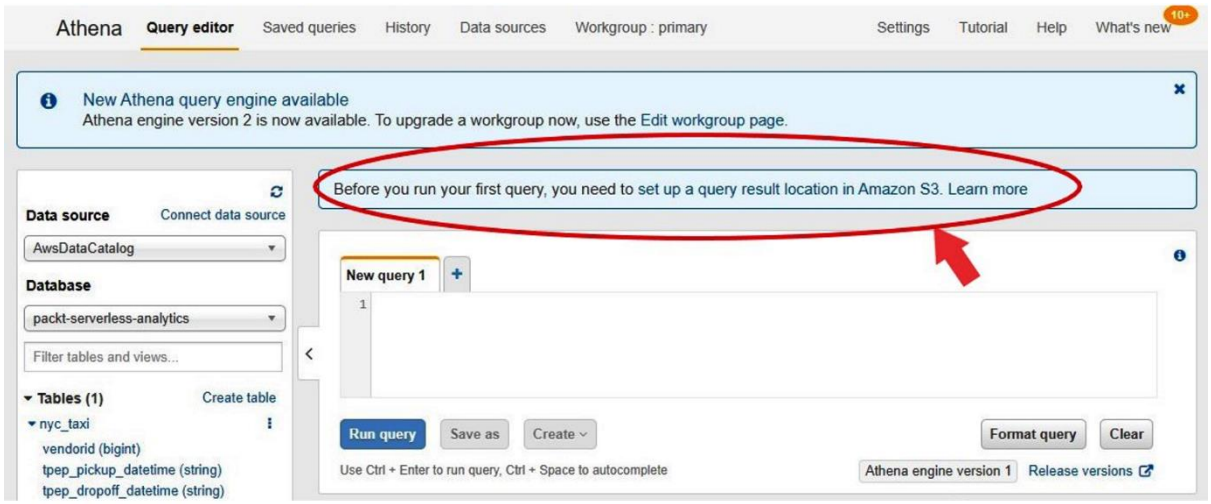
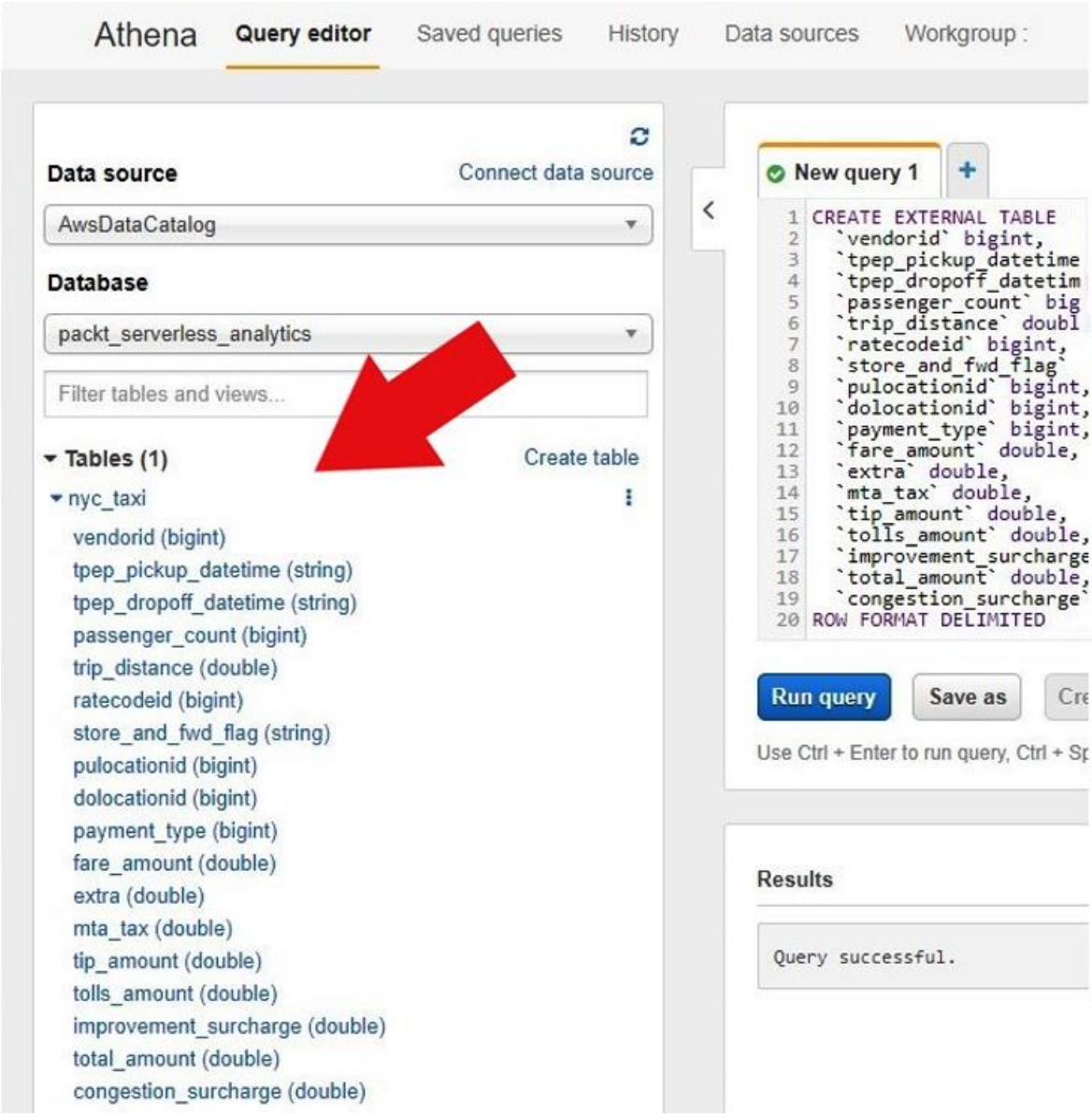
Chapter 1: Your First Query





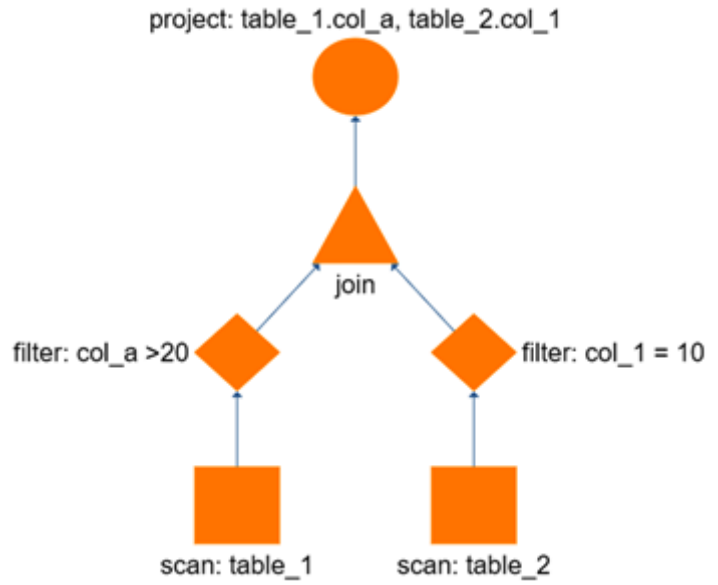
Results

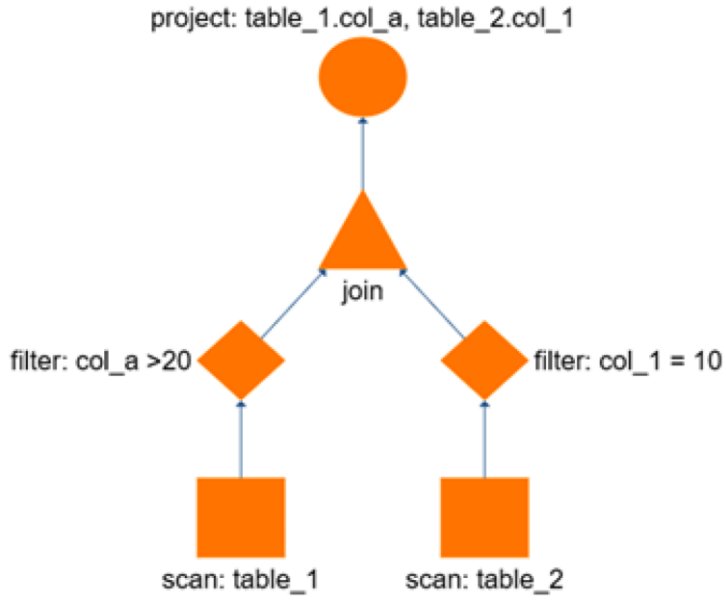
	vendorid	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	ratecodeid	store_and_fwd_flag	pul
1	2	2020-06-01 00:08:05	2020-06-01 00:21:33	1	3.41	1	N	
2	1	2020-06-01 01:41:39	2020-06-01 02:01:44	1	7.6	1	N	
3	2	2020-06-01 03:00:53	2020-06-01 03:09:52	1	2.13	1	N	
4	1	2020-06-01 03:08:27	2020-06-01 03:10:38	1	0.7	1	N	
5	2	2020-06-01 06:00:08	2020-06-01 06:01:43	2	0.61	1	N	
6	2	2020-06-01 06:42:11	2020-06-01 06:54:49	1	3.41	1	N	
7	1	2020-06-01 06:39:50	2020-06-01 06:48:05	1	3.7	1	N	
8	1	2020-06-01 06:25:16	2020-06-01 06:34:37	1	3.4	1	N	
9	2	2020-06-01 06:35:36	2020-06-01 06:46:24	1	2.03	1	N	
10	2	2020-06-01 06:35:49	2020-06-01 06:43:31	1	2.2	1	N	
11	1	2020-06-01 07:12:52	2020-06-01 07:16:14	1	1.0	1	N	
12	1	2020-06-01 07:57:04	2020-06-01 08:11:15	1	7.7	1	N	
13	1	2020-06-01 07:12:45	2020-06-01 07:18:15	1	1.0	1	N	
14	1	2020-06-01 07:27:43	2020-06-01 07:29:53	1	0.8	1	N	

ride_minutes	number_rides
-531231.0	1.0
6.328061	356705.0
16.226482	123621.0
29.021547	56296.0
51.916306	12283.0
518.5143	70.0
995.4878	41.0
1408.4696	741.0
2687.0	1.0
4497.0	1.0

trip_distance (miles)	number_rides
0.2919	31882.0
0.8105	98200.0
1.4027	56680.0
2.5416	5788.0
56.5666	3.0

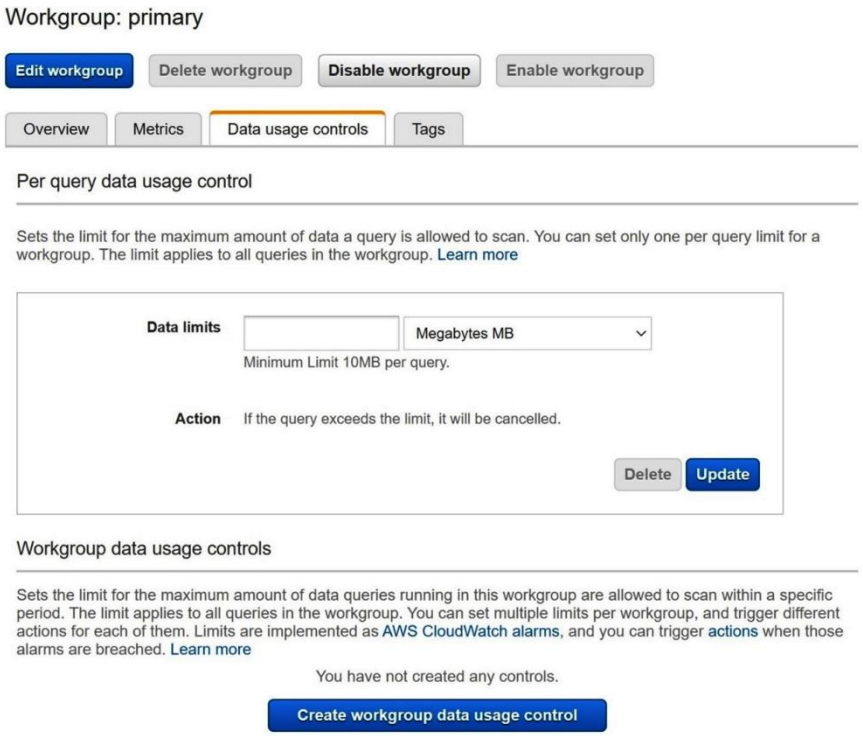
Chapter 2: Introduction to Amazon Athena



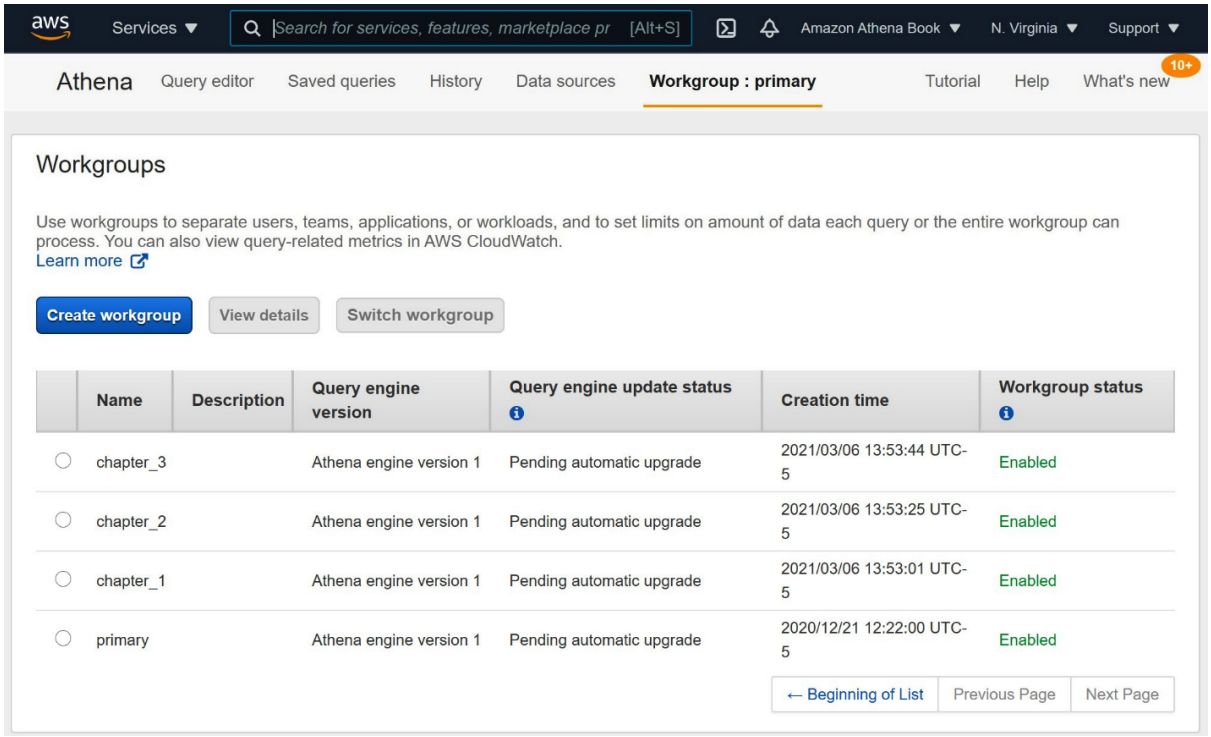


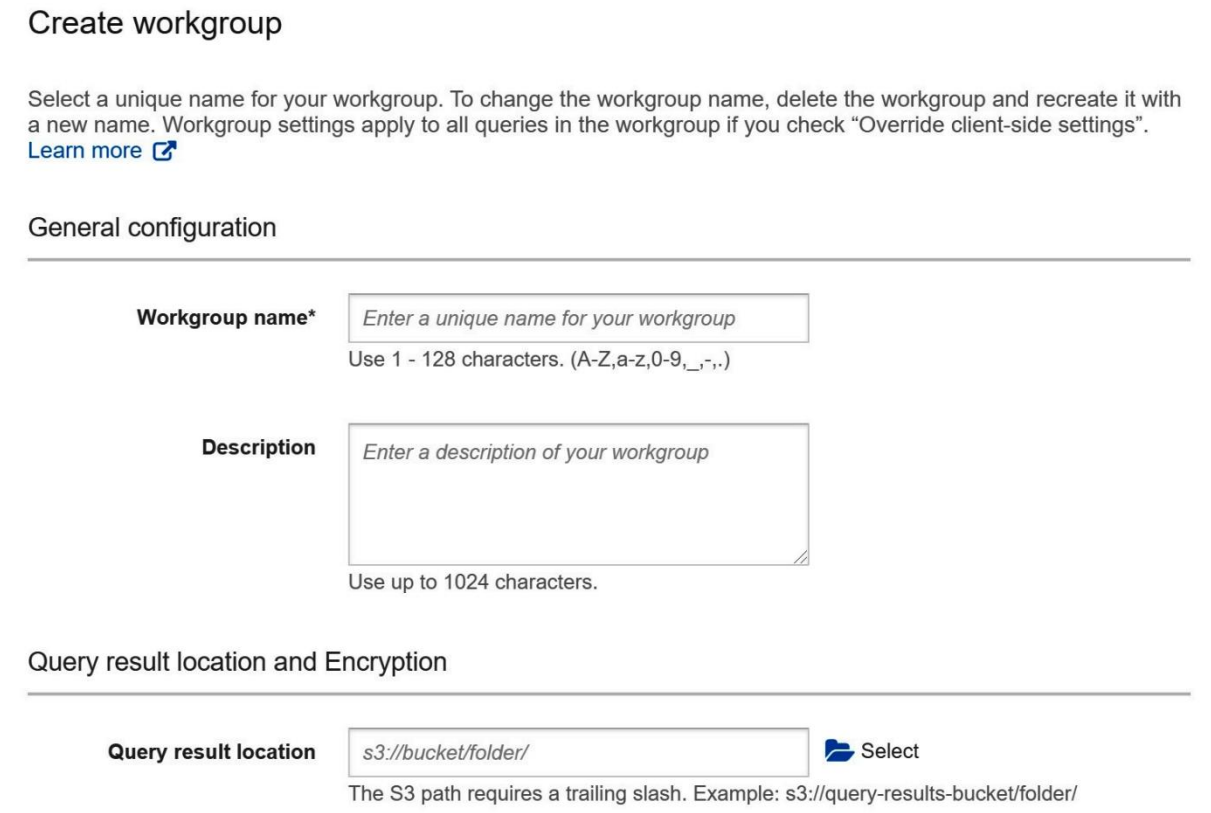
Format	Compression	Size	Performance Ranking
CSV	None	50.7 MB	6
CSV with GZIP	GZIP	9.7 MB	5
Parquet	Run-length encoding	7.9 MB	1
Parquet	Run-length encoding + SNAPPY	10.9 MB	3
ORC	Run-length encoding	10 MB	2
ORC	Run-length enconding + SNAPPY	15.6 MB	4

Format	Compression	Size	Performance Ranking
CSV	None	50.7 MB	6
CSV with GZIP	GZIP	9.7 MB	5
Parquet	Run-length encoding	7.9 MB	1
Parquet	Run-length encoding + SNAPPY	10.9 MB	3
ORC	Run-length encoding	10 MB	2
ORC	Run-length enconding + SNAPPY	15.6 MB	4



Chapter 3: Key Features, Query Types, and Functions





Query engine version

Specify an Athena engine version to use, or let Athena choose when to upgrade your workgroup. Athena occasionally releases a new engine version to provide improved performance, functionality, and code fixes. Learn more

Current Athena engine version

This workgroup is on Athena engine version 1 and is set to let Athena choose when to automatically upgrade to Athena engine version 2, but your engine version will remain on Athena engine version 1 pending the automatic upgrade. Athena engine version 1 will be deprecated in the near future. Workgroups still on Athena engine version 1 will be upgraded to Athena engine version 2 at that time. Athena will notify you of the final deprecation date in the Athena console 30 days ahead of time.

Update query engine

- ○ Let Athena choose when to upgrade your workgroup.
- ◉ Manually choose an engine version now.
  - ◉ **Athena engine version 2 (recommended)**
    This version includes capabilities such as support for nested schema evolution for Parquet format, support for reading nested schema to reduce costs and performance improvements in Join and Aggregate operators. Learn more
  - ○ **Athena engine version 1**
    This is Athena's initial version.

Metrics

Metrics ☑ Publish query metrics to AWS CloudWatch

Settings

Override client-side settings ☐

Requester pays S3 buckets ☐ Enable queries on requester pays buckets in Amazon S3

Tags

A tag is a label that you assign to an Athena workgroup resource. It consists of a key and a value. Use tags to categorize workgroups by purpose, owner, or environment. You can also use tags in IAM policies to allow or deny access to workgroup actions based on a tag key/value pair, or on specific values for a tag key. Use best practices and create a consistent set of tags. Do not use duplicate tag keys the same workgroup. Learn more

Key	Value (Optional)
Enter key	*Enter value*
Use 1 - 128 characters. (A-Z,a-z,0-9, ,_.:/=+-@)	Use up to 256 characters. (A-Z,a-z,0-9, ,_.:/=+-@)

Cancel | Create workgroup

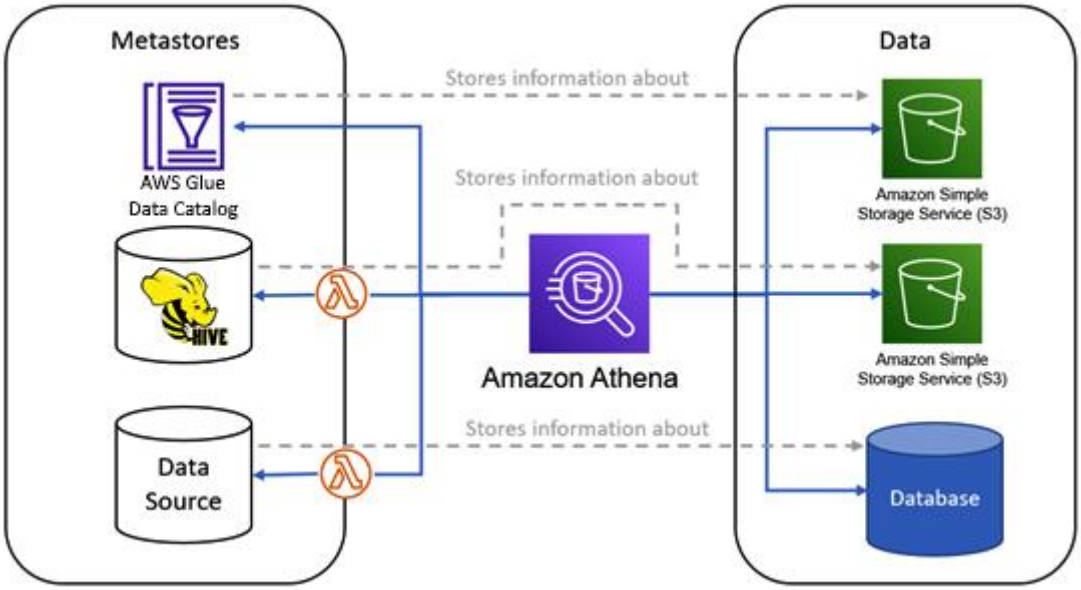
New query 1 +

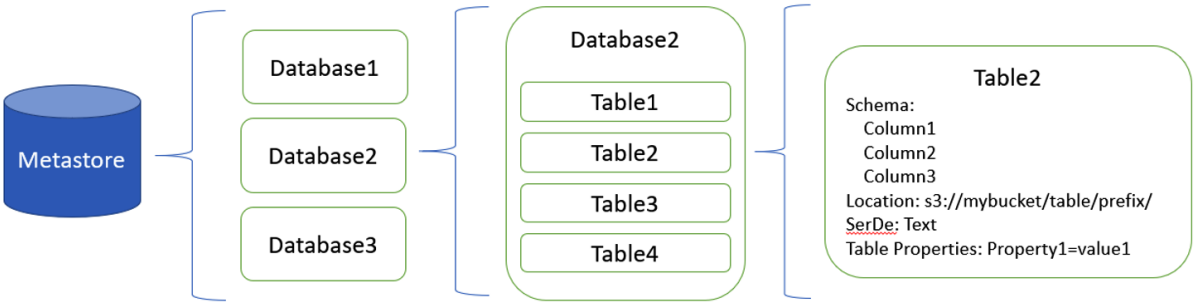
```
1 SELECT year, COUNT(*) from packt_serverless_analytics.chapter_3_nyc_taxi_parquet GROUP BY year
```

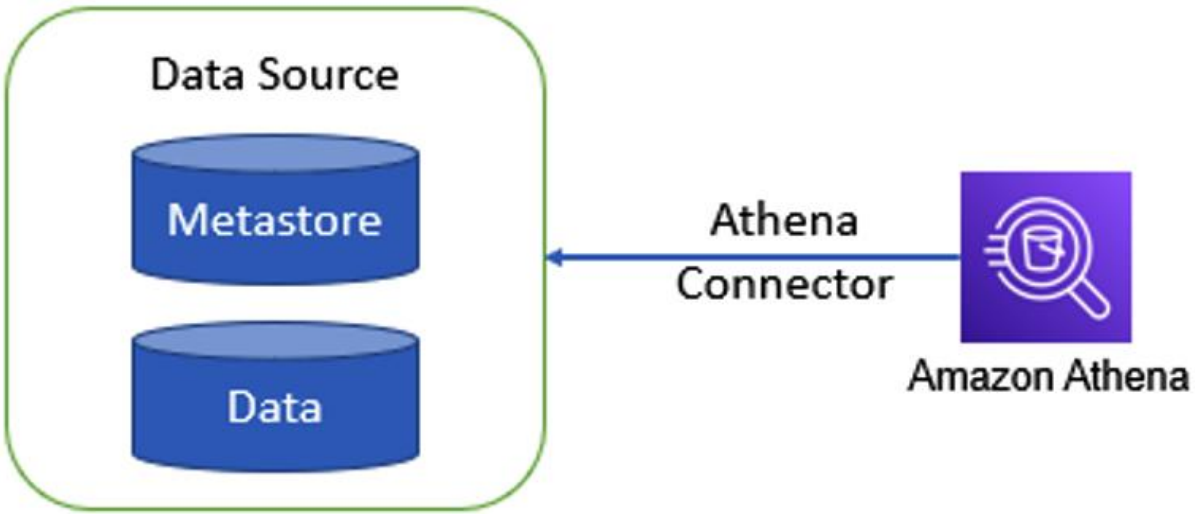
Run query | Save as | Create | (Run time: 5.44 seconds, Data scanned: 0 KB) | Format query | Clear

Hour	Actual Count	BERNOULLI	SYSTEM
17	18206794	18223390	15679710
18	20352240	20378540	19000690
19	19580912	19553590	17711660
20	17871911	17903240	16079030
21	17712882	17720660	16195500

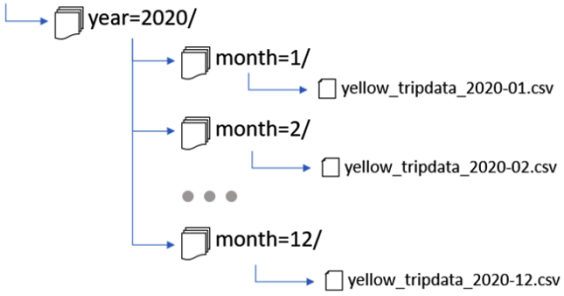
Chapter 4: Metastores, Data Sources, and Data Lakes

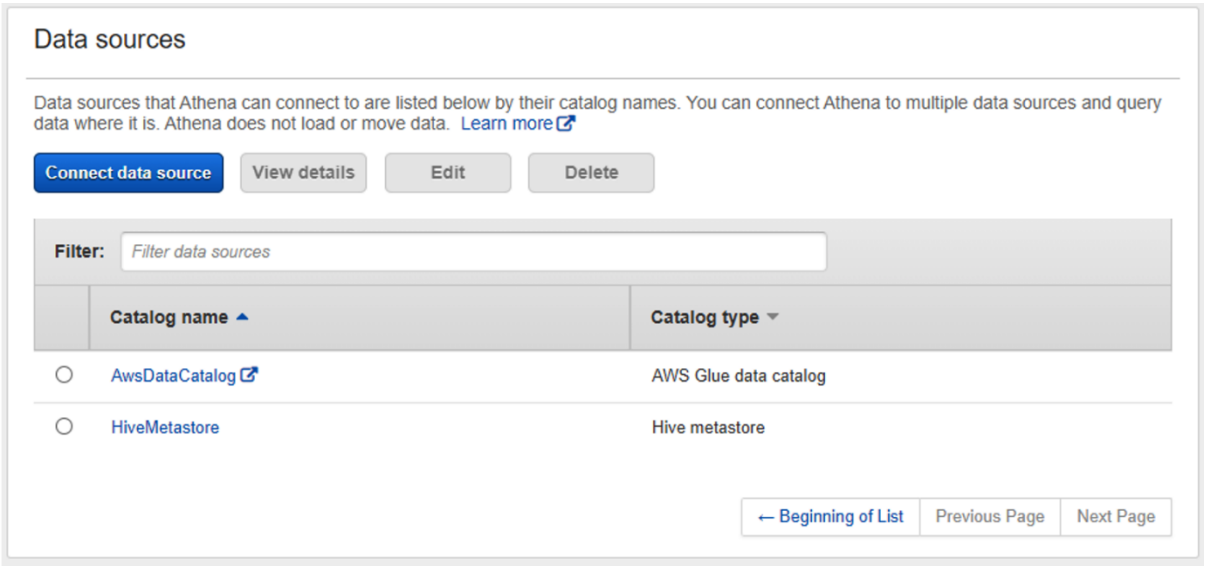


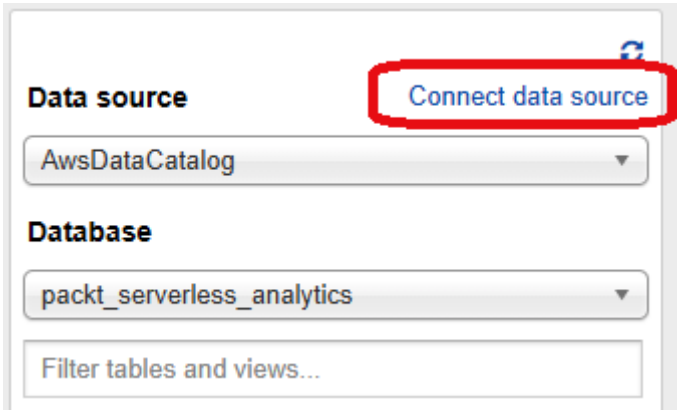


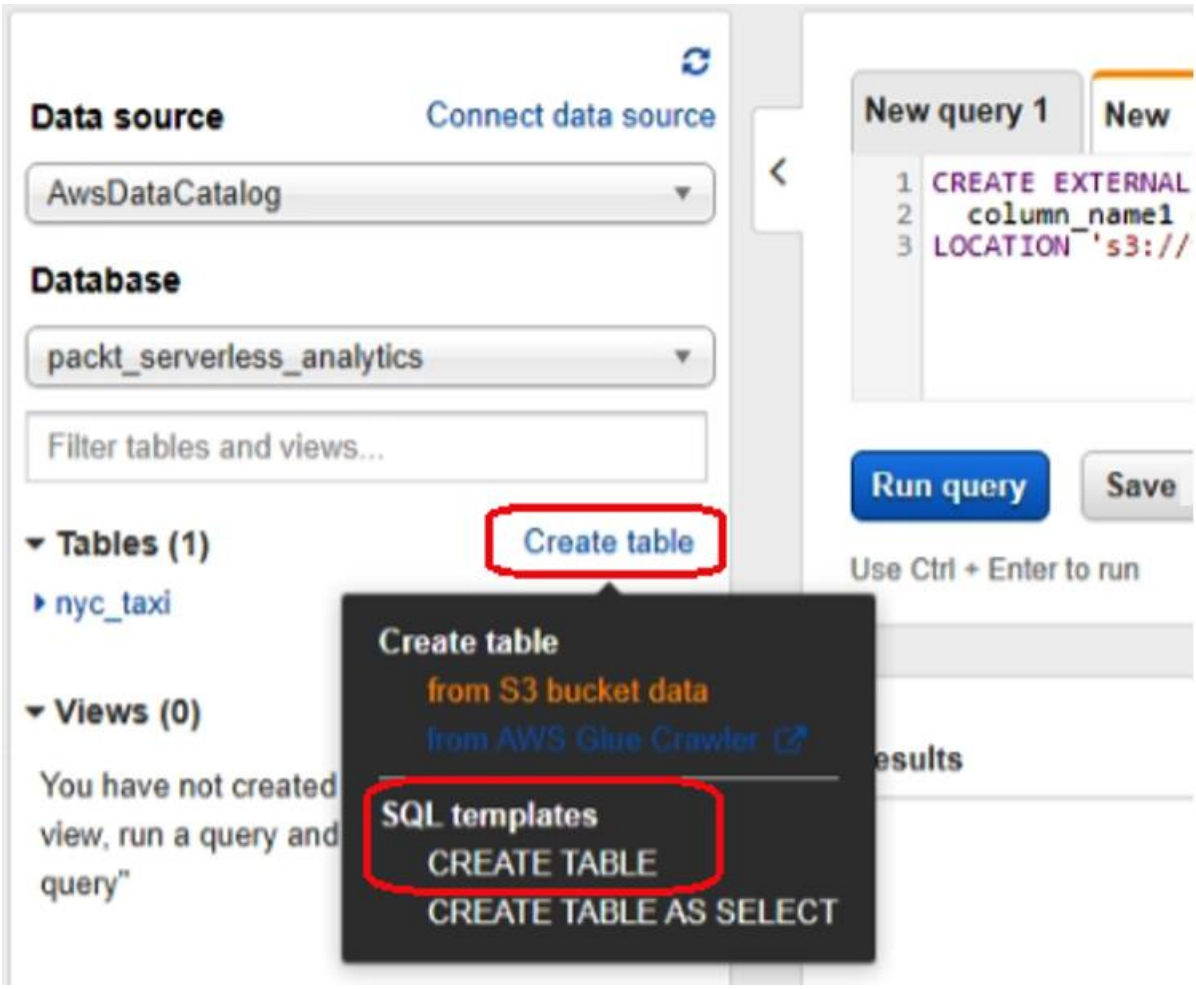


's3://packt-serverless-analytics/tables/nyc_taxi_partitioned/

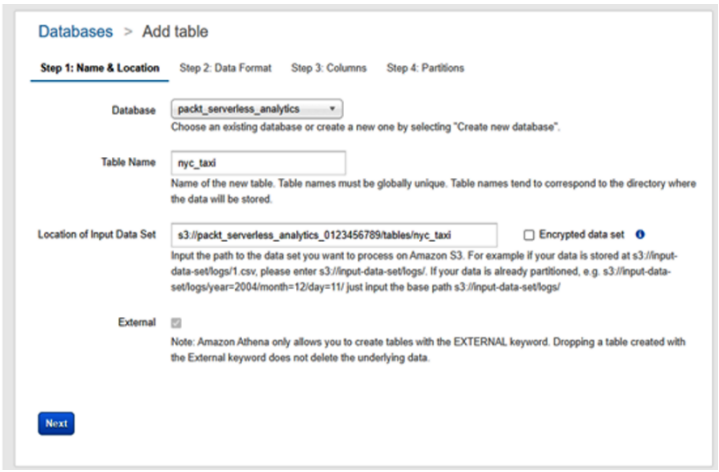


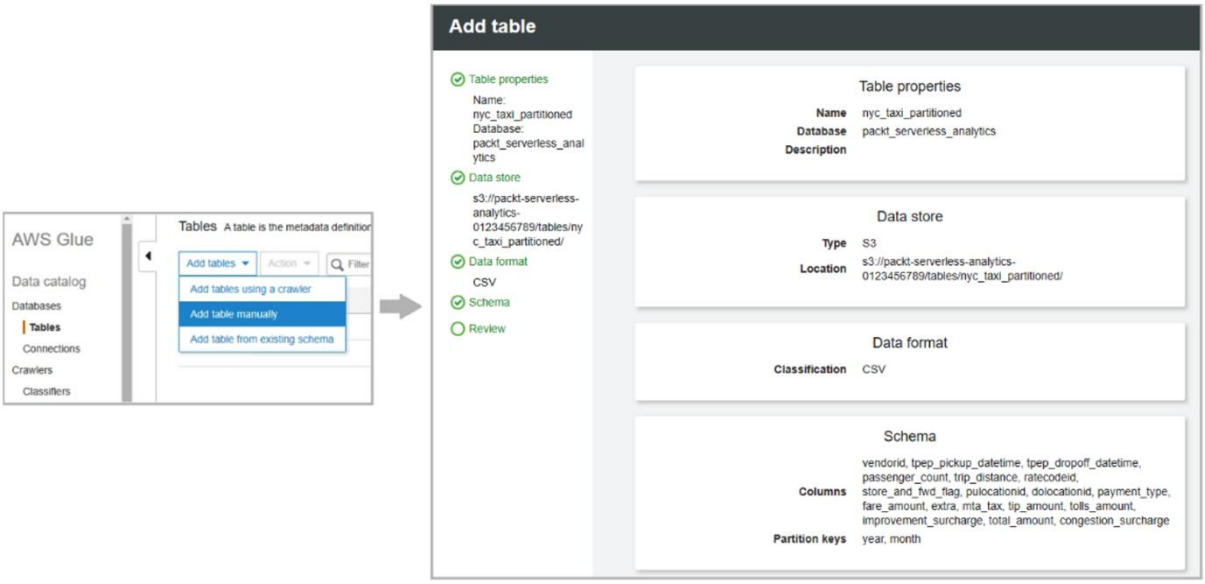


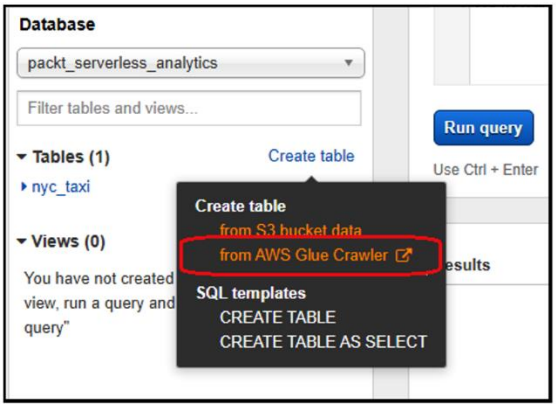


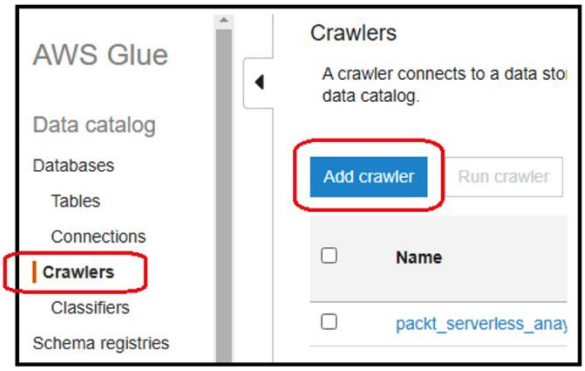

Data source
Connect data source
AwsDataCatalog
Database
packt_serverless_analytics
Filter tables and views...
Tables (1)
Create table
nyc_taxi
Views (0)
You have not created
view, run a query and
query"
Create table
from S3 bucket data
from AWS Glue Crawler
SQL templates
CREATE TABLE
CREATE TABLE AS SELECT
New query 1
New
1 CREATE EXTERNAL
2 column_name1
3 LOCATION 's3://
Run query
Save
Use Ctrl + Enter to run
esults

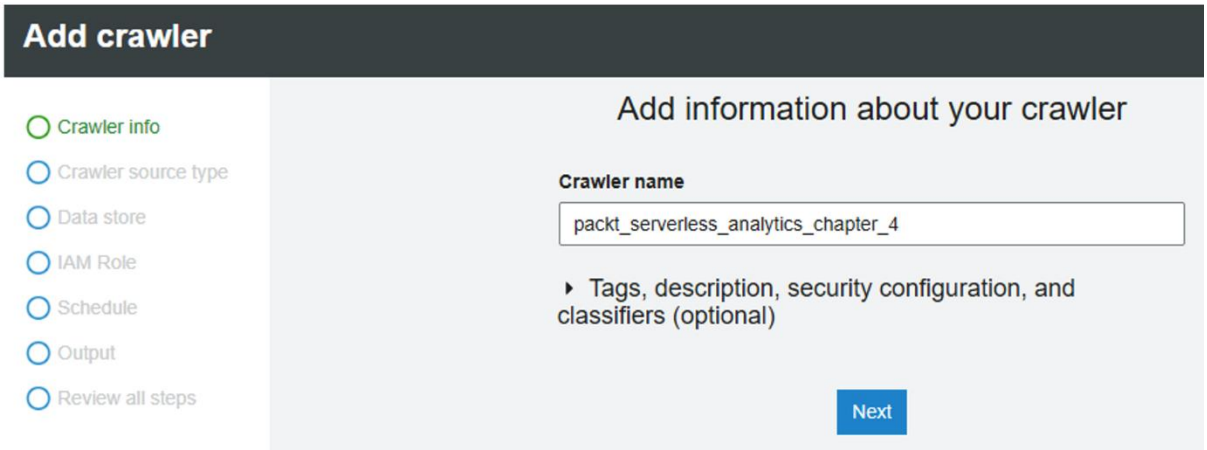

Data source
Connect data source
AwsDataCatalog
Database
packt_serverless_analytics
Filter tables and views...
Tables (1)
Create table
nyc_taxi


Databases > Add table
Step 1: Name & Location
Step 2: Data Format
Step 3: Columns
Step 4: Partitions
Database
packt_serverless_analytics
Choose an existing database or create a new one by selecting "Create new database".
Table Name
nyc_taxi
Name of the new table. Table names must be globally unique. Table names tend to correspond to the directory where the data will be stored.
Location of Input Data Set
s3://packt_serverless_analytics_0123456789/tables/nyc_taxi
Encrypted data set
Input the path to the data set you want to process on Amazon S3. For example if your data is stored at s3://input-data-set/logs/1.csv, please enter s3://input-data-set/logs/. If your data is already partitioned, e.g. s3://input-data-set/logs/year=2004/month=12/day=11/ just input the base path s3://input-data-set/logs/
External
Note: Amazon Athena only allows you to create tables with the EXTERNAL keyword. Dropping a table created with the External keyword does not delete the underlying data.
Next


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Tables A table is the metadata definition
Add tables
Action
Filter
Add tables using a crawler
Add table manually
Add table from existing schema
Add table
Table properties
Name: nyc_taxi_partitioned
Database: packt_serverless_analytics
Data store
s3://packt-serverless-analytics-0123456789/tables/nyc_taxi_partitioned/
Data format
CSV
Schema
Review
Table properties
Name nyc_taxi_partitioned
Database packt_serverless_analytics
Description
Data store
Type S3
Location s3://packt-serverless-analytics-0123456789/tables/nyc_taxi_partitioned/
Data format
Classification CSV
Schema
Columns vendorid, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, ratecodeid, store_and_fwd_flag, pulocationid, dolocationid, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge
Partition keys year, month


Database
packt_serverless_analytics
Filter tables and views...
Tables (1)
Create table
nyc_taxi
Views (0)
You have not created
view, run a query and
query"
Run query
Use Ctrl + Enter
esults
Create table
from S3 bucket data
from AWS Glue Crawler
SQL templates
CREATE TABLE
CREATE TABLE AS SELECT

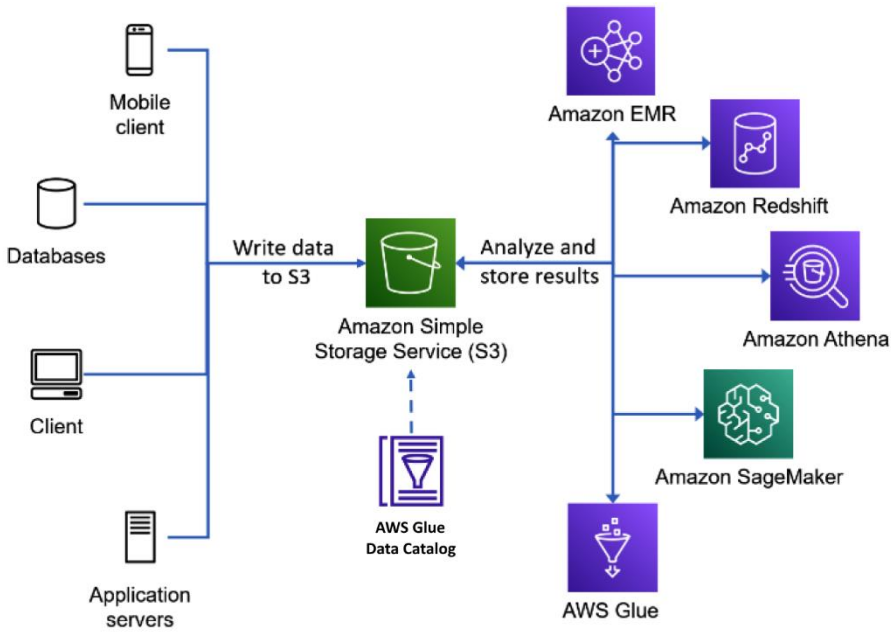

AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Crawlers
A crawler connects to a data sto
data catalog.
Add crawler
Run crawler
Name
packt_serverless_anay


Add crawler
Crawler info
Crawler source type
Data store
IAM Role
Schedule
Output
Review all steps
Add information about your crawler
Crawler name
packt_serverless_analytics_chapter_4
Tags, description, security configuration, and classifiers (optional)
Next

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
packt_serverless_analytics_chapter_4		Ready	Logs	1 min	1 min	0	2

Add tables | Action | Database : packt_serverless_analytics_chapter_4 | Save view | Showing: 1 - 2

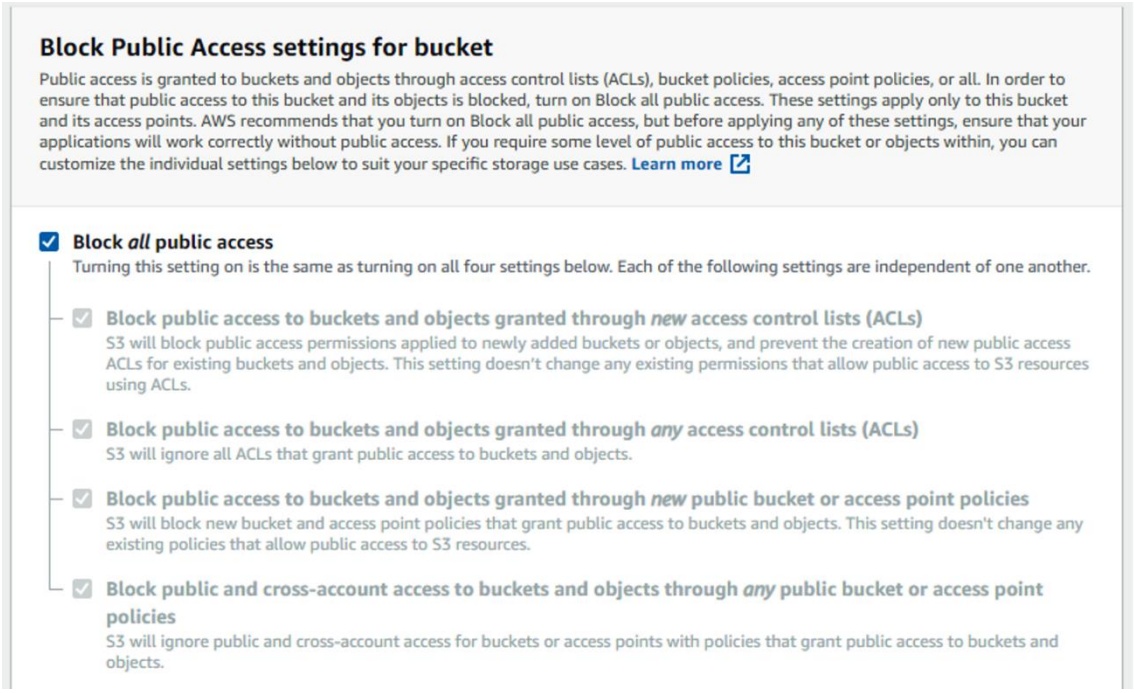
Name	Database	Location	Classificatio	Last updated	Deprecated
nyc_taxi	packt_serverless_analyt...	s3://packt-serverless-an...	csv	19 January 2021 2:36 A...	
nyc_taxi_partitioned	packt_serverless_analyt...	s3://packt-serverless-an...	csv	19 January 2021 2:36 A...	

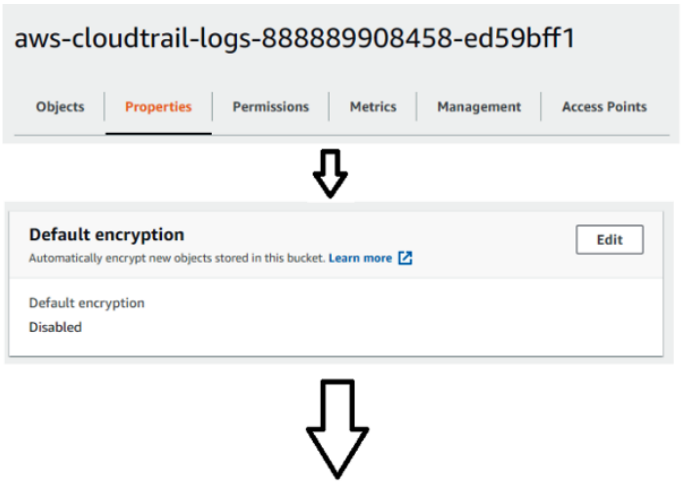


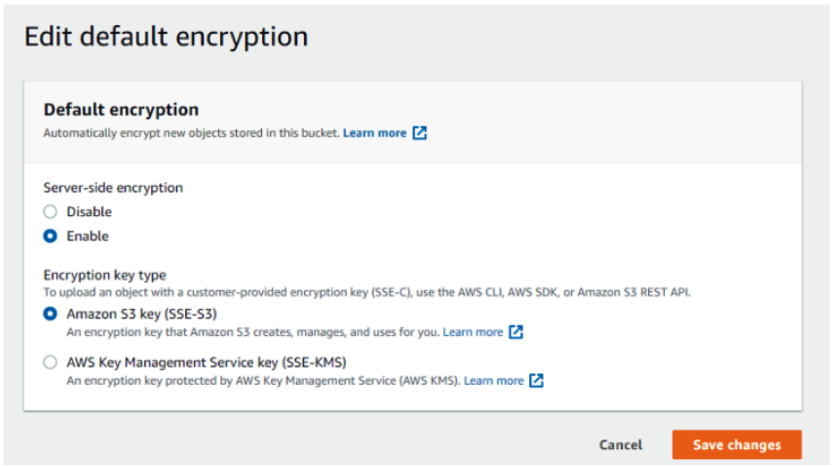
Feature	AWS Glue Data Catalog	Hive Metastore
Stores databases, tables, and columns	Yes.	Yes.
Stores table statistics	Yes.	Yes.
Supported operations	`SELECT`, `CTAS`, `CREATE TABLE`, `ALTER TABLE`, `SHOW COLUMNS`, `SHOW TABLES`, `SHOW SCHEMAS`, `SHOW CREATE TABLE`, `SHOW TBLPROPERTIES`, `SHOW PARTITIONS`.	`SHOW COLUMNS`, `SHOW TABLES`, `SHOW SCHEMAS`, `SHOW CREATE TABLE`, `SHOW TBLPROPERTIES`, `SHOW PARTITIONS`.

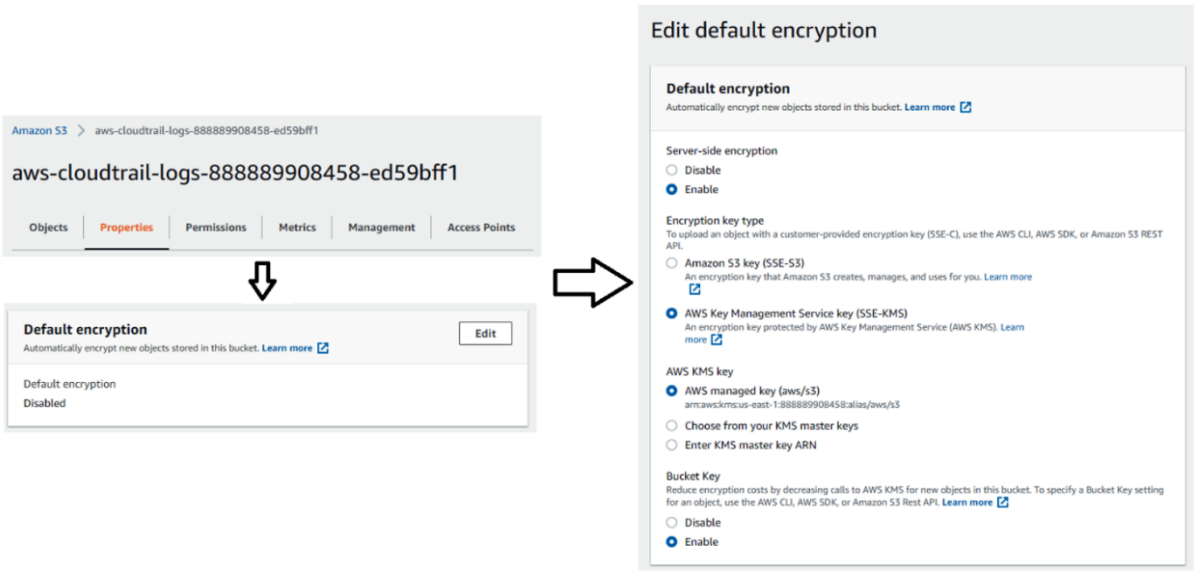
Feature	AWS Glue Data Catalog	Hive Metastore
Cost	First 1 million objects free, $1.00 per 100,000 objects. First 1 million object requests free, $1.00 per 1 million requests afterward.	Cost of hardware to host metastore processes, plus database costs. Athena usage will incur AWS Lambda usage costs.
Operational load	Low.	Medium to high.
Availability	High.	Low to high.
Performance	High.	Low to medium.
Open source engine compatibility	Apache Hive, Apache Spark, PrestoDB, Trino.	Apache Hive, Apache Spark, PrestoDB, Trino, Apache Pig.
AWS service compatibility	Amazon Athena, Amazon EMR, AWS Glue, Amazon Redshift.	Amazon Athena, Amazon EMR.
Table versioning	Yes.	No.
Auditing of changes	Yes, through AWS CloudTrail.	Yes, through log files.
Data access controls	Yes, IAM provides authentication, and authorization can be implemented using Lake Formation or IAM resource policies. See *Chapter 5*, *Securing Your Data*, for details.	Yes, Hive metastore processes can employ different authentication and authorization methods. See the Further reading section for links to the available mechanisms.

Chapter 5: Securing Your Data

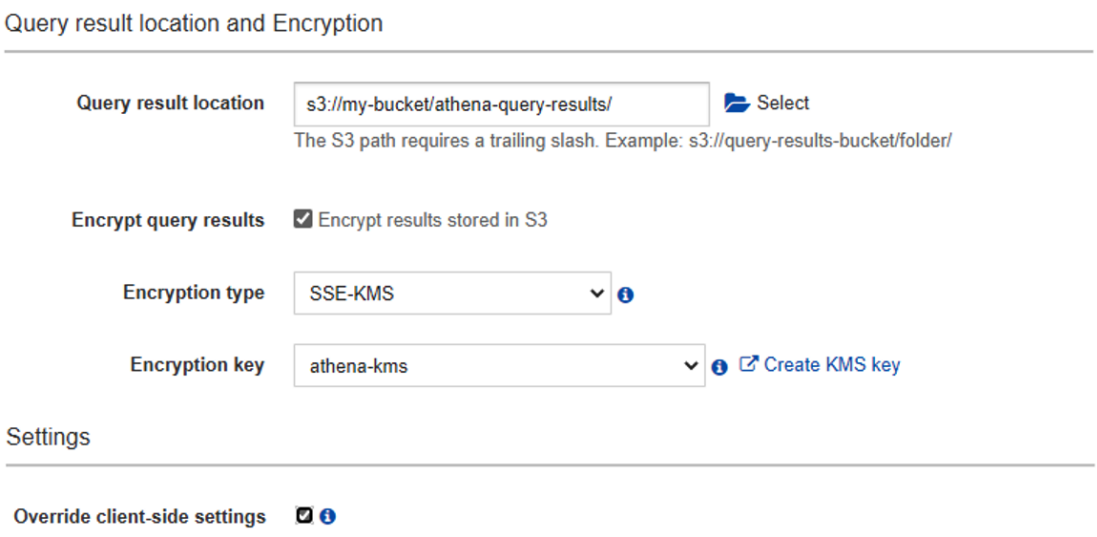


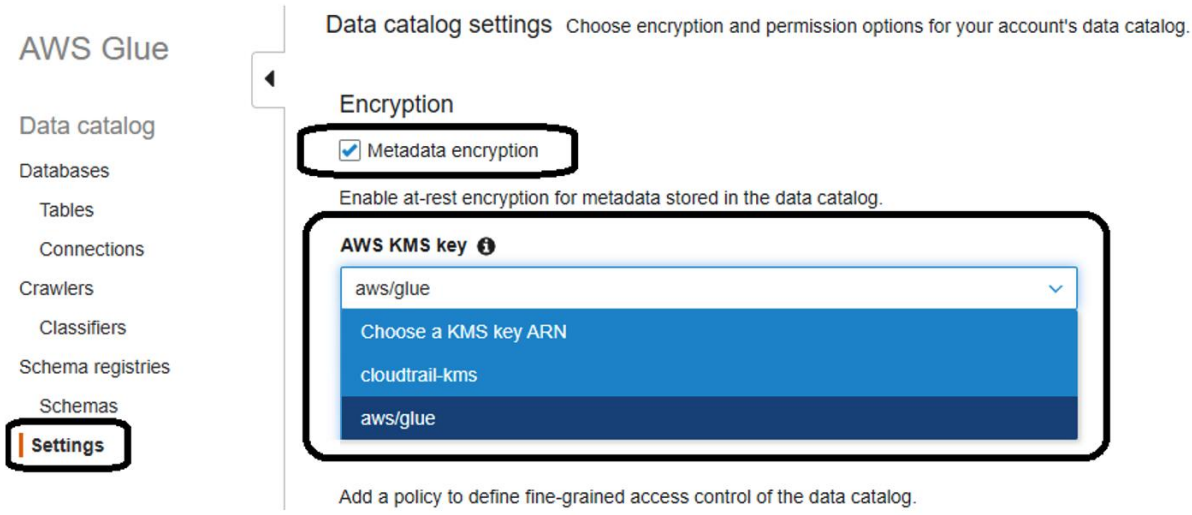


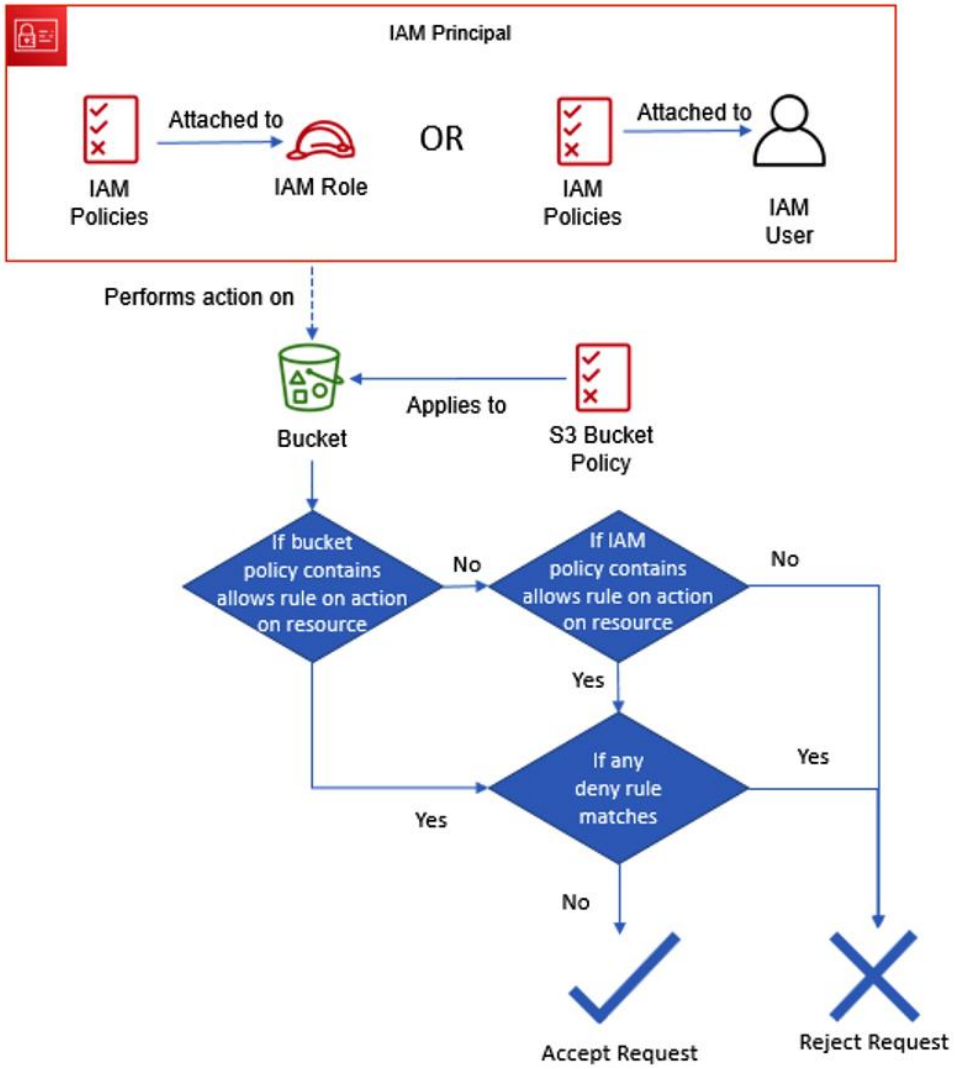


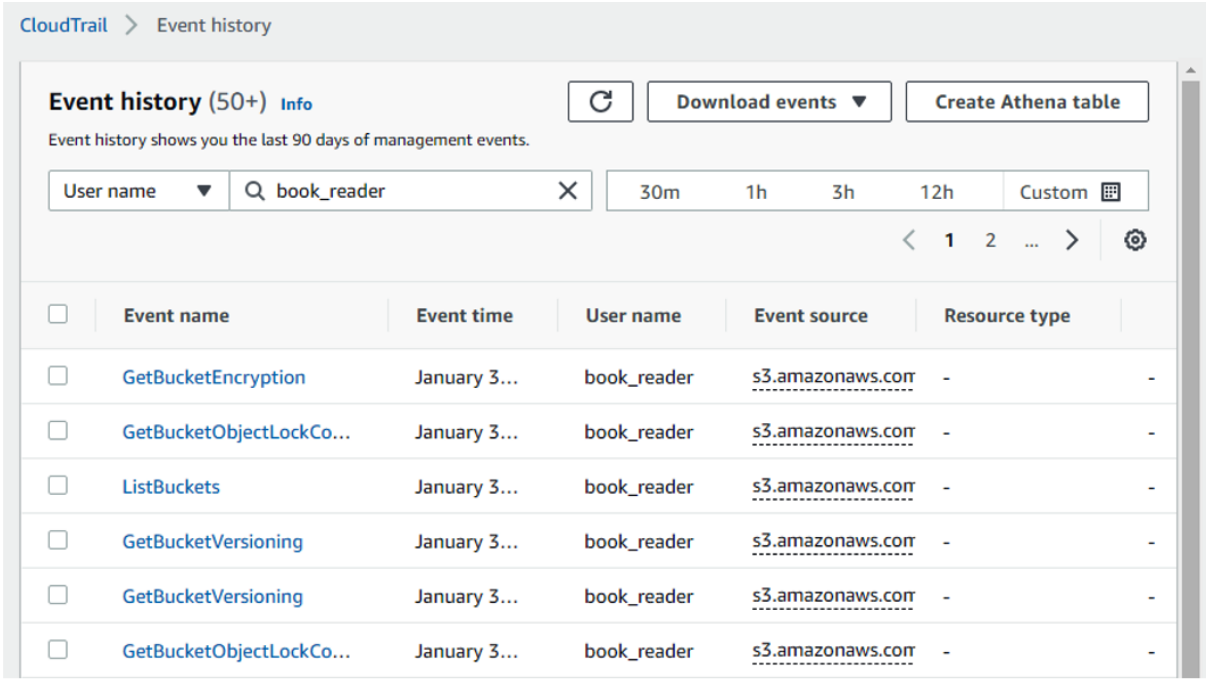


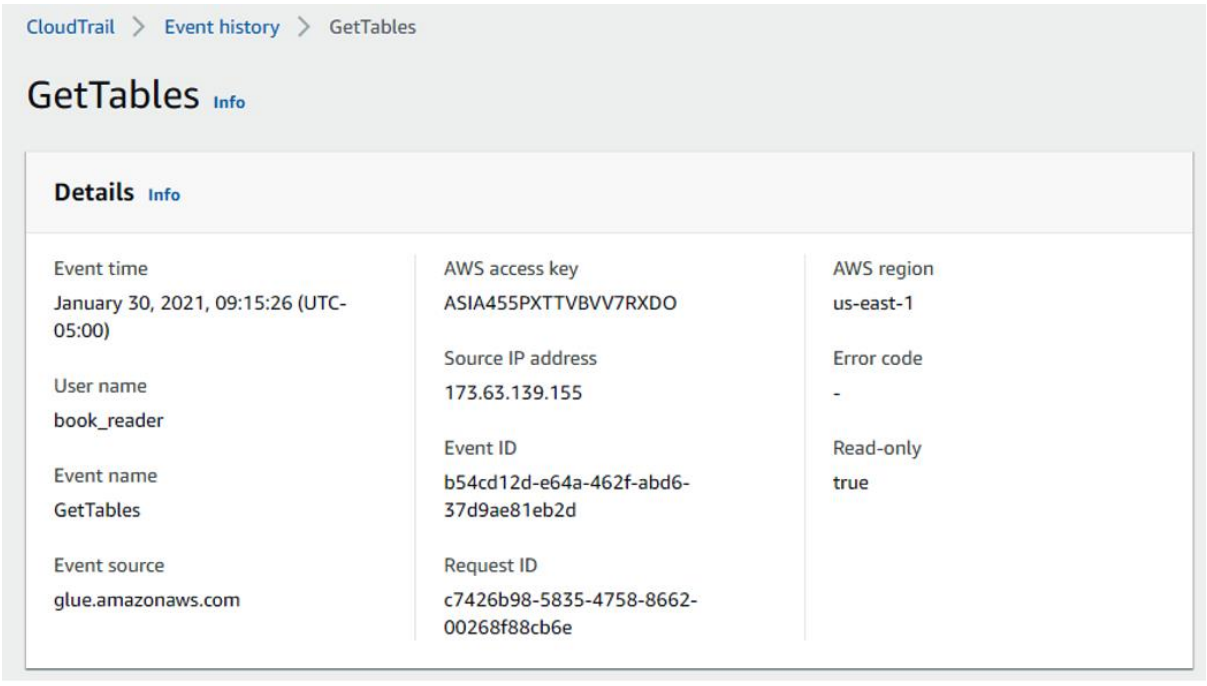
Encryption type	Performance	Cost	Level of security
SSE-S3	High	Free	Medium
SSE-KMS	Medium	Cost of KMS key + cost of API costs to KMS. KMS costs can be minimized if the same master key is used.	High
CSE-KMS	Low. Encryption and decryption may not be as parallelly done as if S3 does it.	Cost of KMS key + cost of API costs to KMS. KMS costs can be higher because KMS key caching cannot be done.	Highest

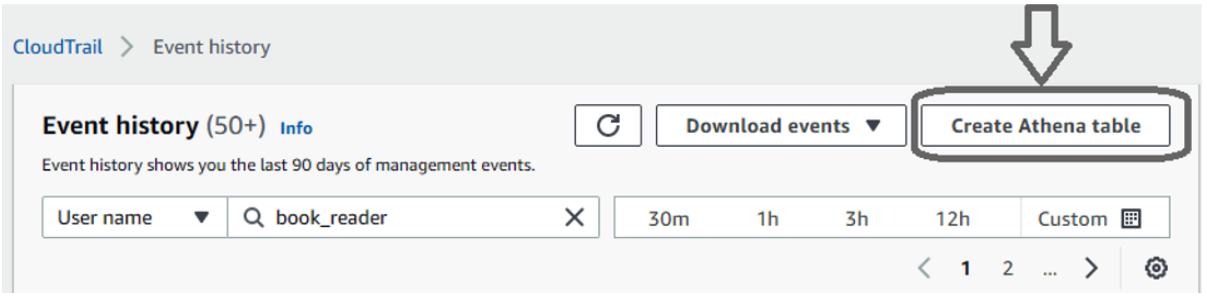


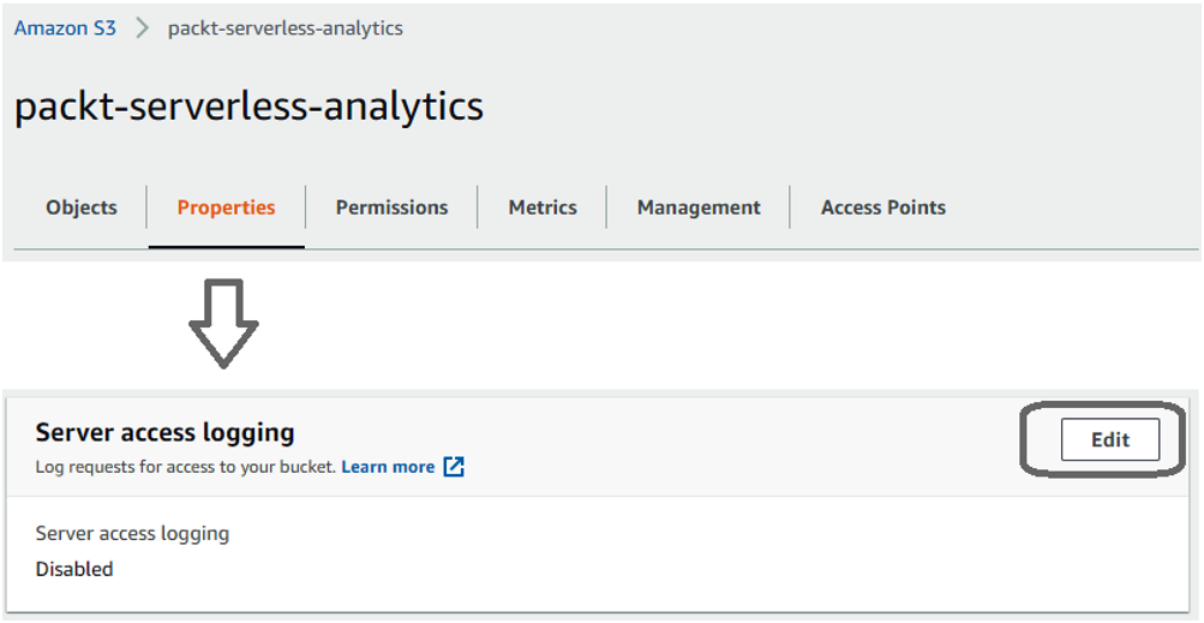

AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
Data catalog settings
Choose encryption and permission options for your account's data catalog.
Encryption
Metadata encryption
Enable at-rest encryption for metadata stored in the data catalog.
AWS KMS key
aws/glue
Choose a KMS key ARN
cloudtrail-kms
aws/glue
Add a policy to define fine-grained access control of the data catalog.


IAM Principal
Attached to
IAM Policies
IAM Role
OR
Attached to
IAM Policies
IAM User
Performs action on
Bucket
Applies to
S3 Bucket Policy
If bucket policy contains allows rule on action on resource
No
If IAM policy contains allows rule on action on resource
No
Yes
If any deny rule matches
Yes
Yes
No
Accept Request
Reject Request

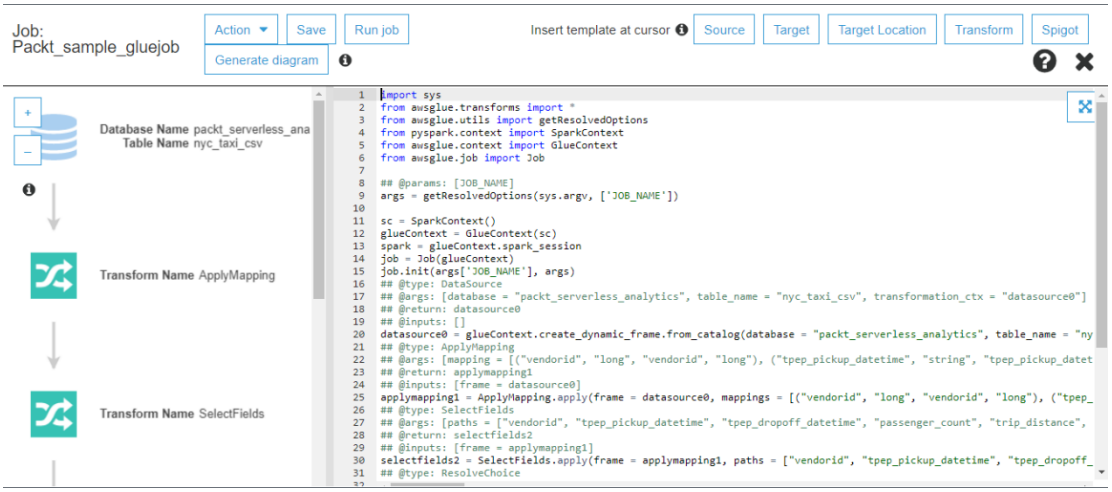

CloudTrail > Event history
Event history (50+) Info
Download events
Create Athena table
Event history shows you the last 90 days of management events.
User name
book_reader
30m
1h
3h
12h
Custom
1
2
...
Event name
Event time
User name
Event source
Resource type
GetBucketEncryption
January 3...
book_reader
s3.amazonaws.com
GetBucketObjectLockCo...
January 3...
book_reader
s3.amazonaws.com
ListBuckets
January 3...
book_reader
s3.amazonaws.com
GetBucketVersioning
January 3...
book_reader
s3.amazonaws.com
GetBucketVersioning
January 3...
book_reader
s3.amazonaws.com
GetBucketObjectLockCo...
January 3...
book_reader
s3.amazonaws.com

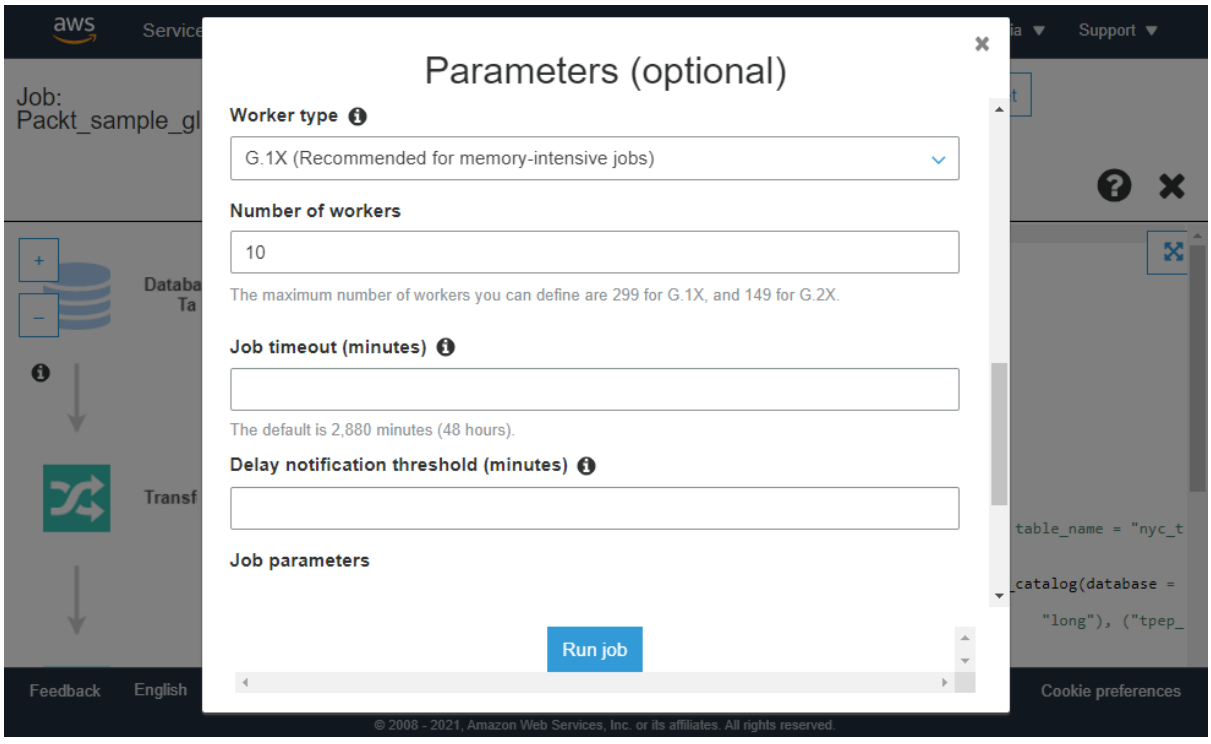

CloudTrail > Event history > GetTables
GetTables Info
Details Info
Event time
January 30, 2021, 09:15:26 (UTC-05:00)
User name
book_reader
Event name
GetTables
Event source
glue.amazonaws.com
AWS access key
ASIA455PXTTVBVV7RXDO
Source IP address
173.63.139.155
Event ID
b54cd12d-e64a-462f-abd6-37d9ae81eb2d
Request ID
c7426b98-5835-4758-8662-00268f88cb6e
AWS region
us-east-1
Error code
-
Read-only
true

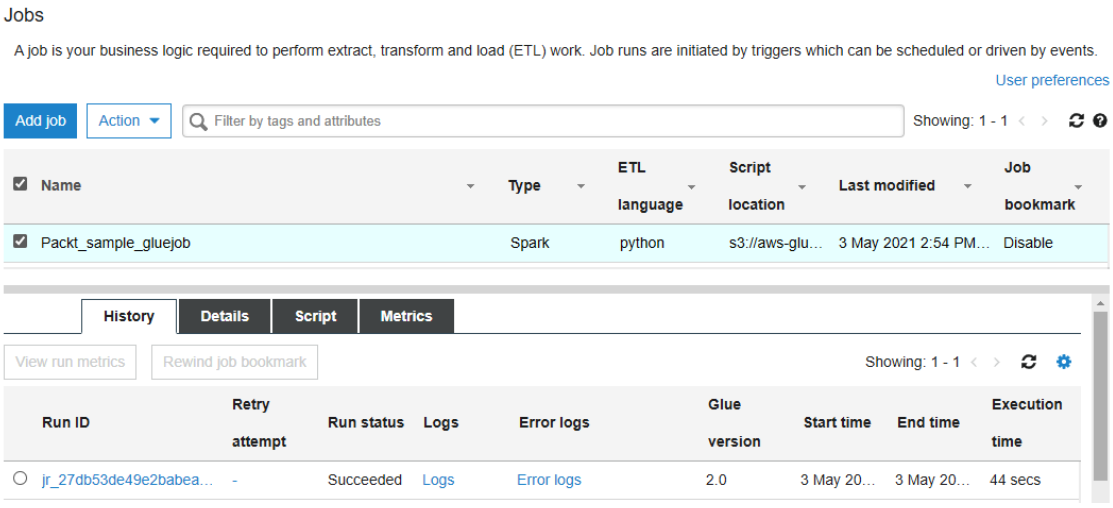

CloudTrail > Event history
Event history (50+) Info
Download events
Create Athena table
Event history shows you the last 90 days of management events.
User name
book_reader
30m
1h
3h
12h
Custom
1
2
...

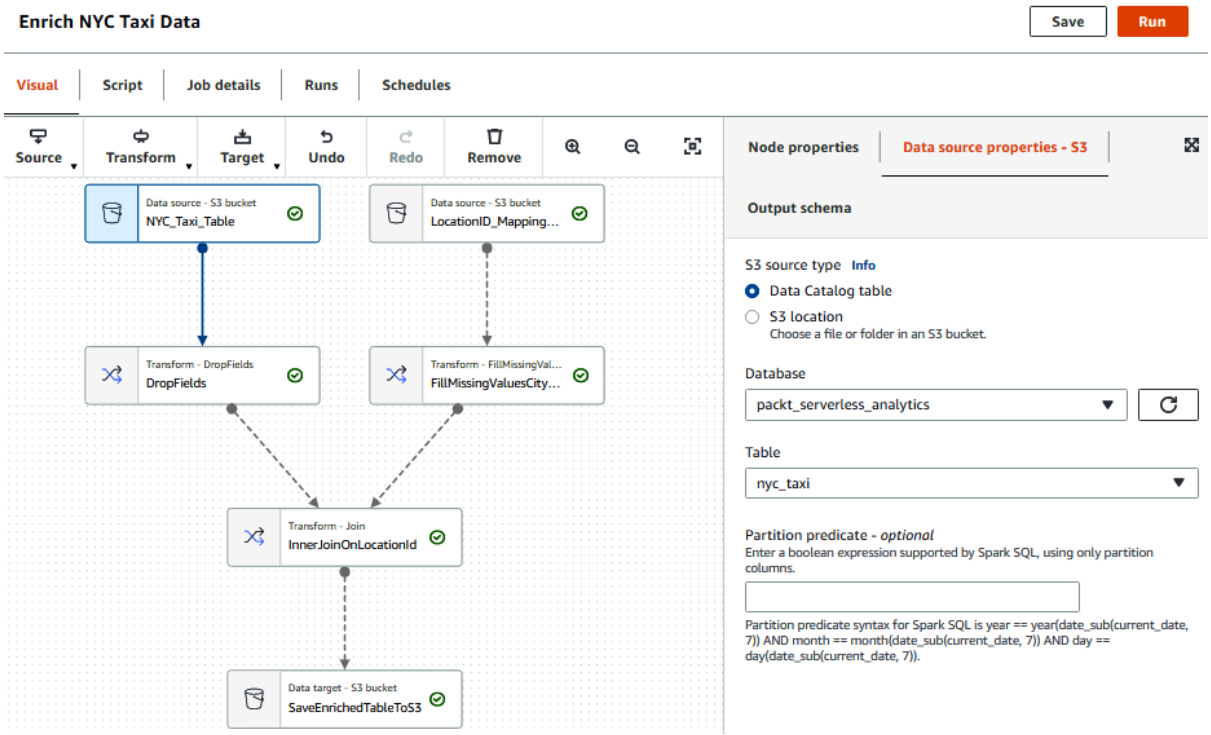

Amazon S3
packt-serverless-analytics
packt-serverless-analytics
Objects
Properties
Permissions
Metrics
Management
Access Points
Server access logging
Log requests for access to your bucket. Learn more
Edit
Server access logging
Disabled

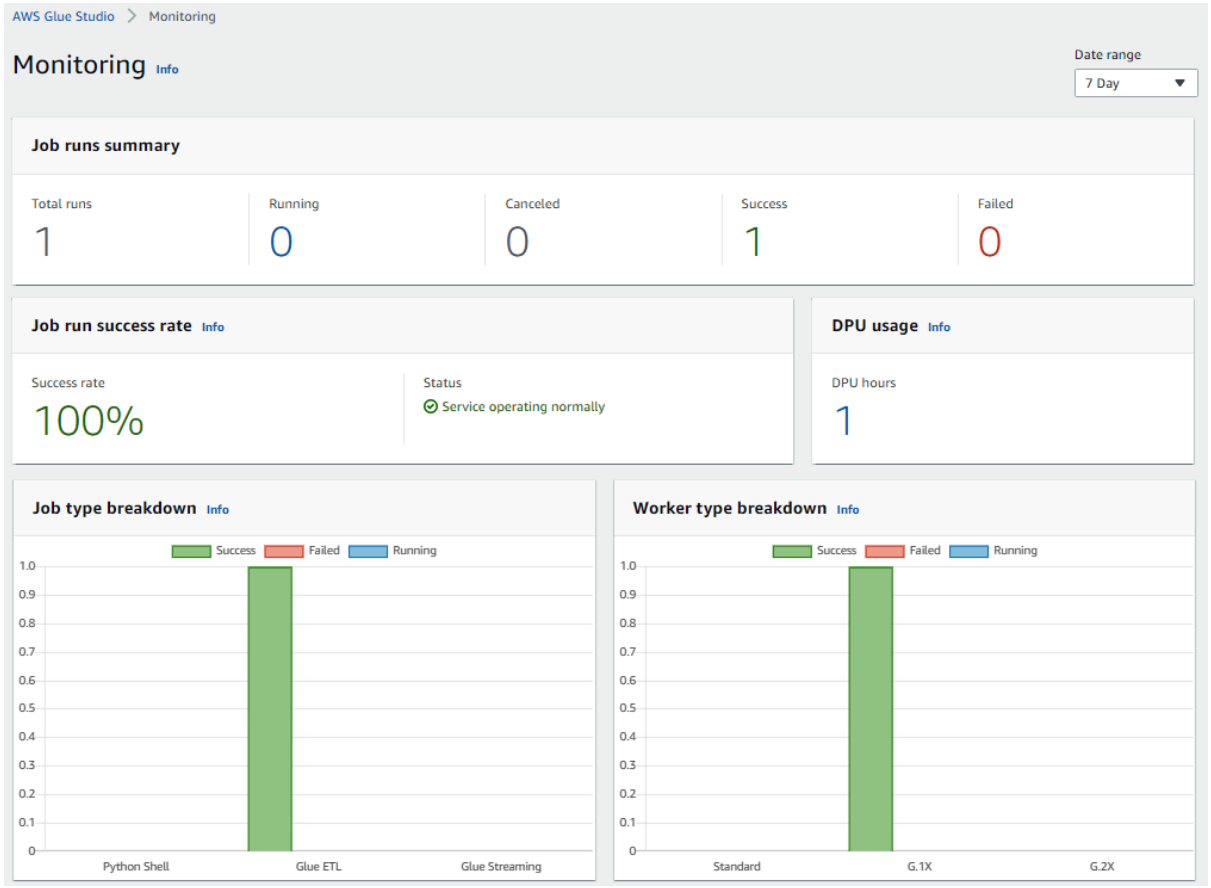
Chapter 6: AWS Glue and AWS Lake Formation

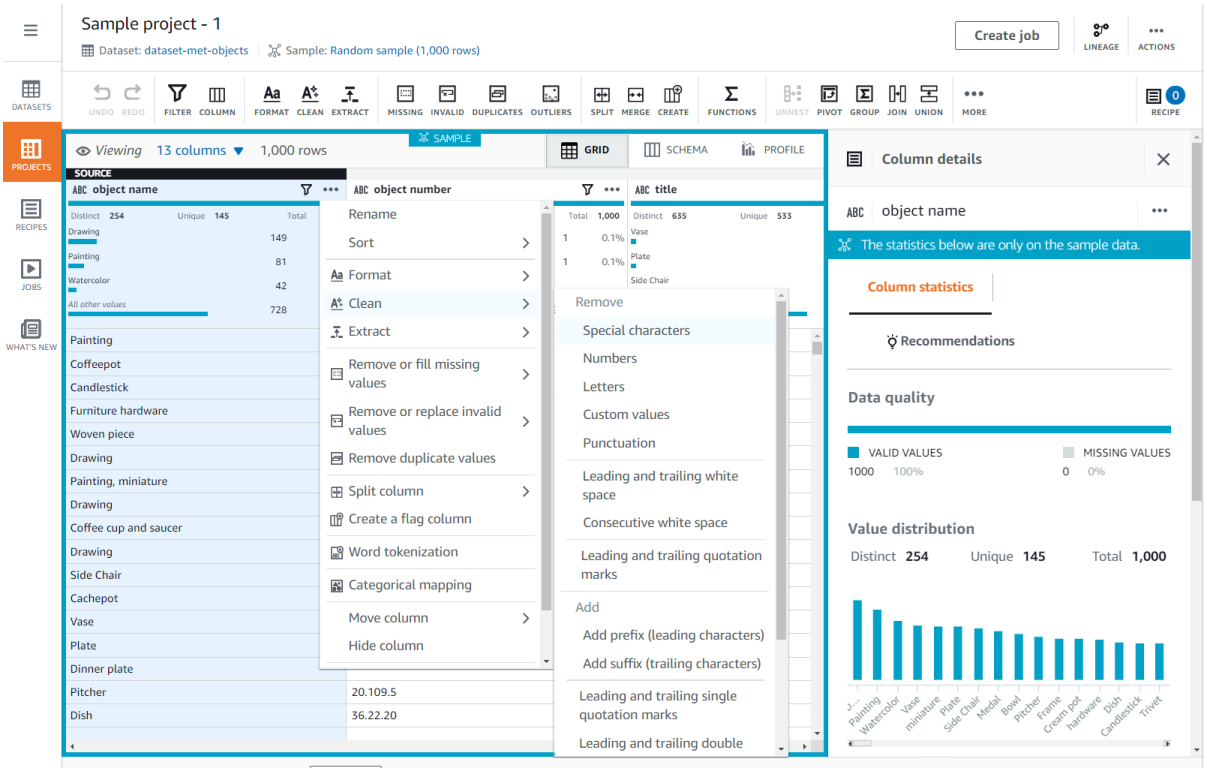


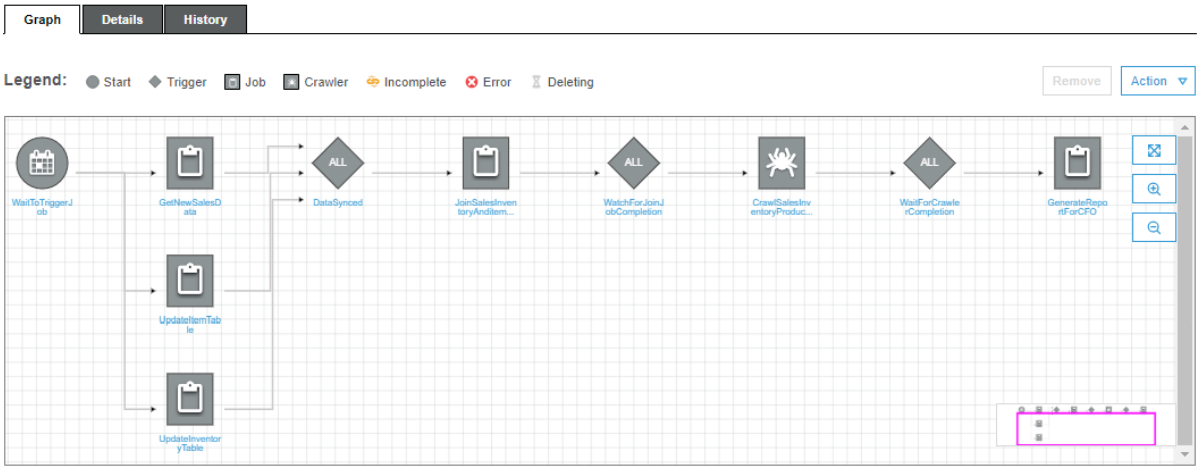





Enrich NYC Taxi Data
Save
Run
Visual
Script
Job details
Runs
Schedules
Source
Transform
Target
Undo
Redo
Remove
Data source - S3 bucket
NYC_Taxi_Table
Data source - S3 bucket
LocationID_Mapping...
Transform - DropFields
DropFields
Transform - FillMissingVal...
FillMissingValuesCity...
Transform - Join
InnerJoinOnLocationId
Data target - S3 bucket
SaveEnrichedTableToS3
Node properties
Data source properties - S3
Output schema
S3 source type Info
Data Catalog table
S3 location
Choose a file or folder in an S3 bucket.
Database
packt_serverless_analytics
Table
nyc_taxi
Partition predicate - optional
Enter a boolean expression supported by Spark SQL, using only partition columns.
Partition predicate syntax for Spark SQL is year == year(date_sub(current_date, 7)) AND month == month(date_sub(current_date, 7)) AND day == day(date_sub(current_date, 7)).


AWS Glue Studio > Monitoring
Monitoring Info
Date range
7 Day
Job runs summary
Total runs
1
Running
0
Canceled
0
Success
1
Failed
0
Job run success rate Info
Success rate
100%
Status
Service operating normally
DPU usage Info
DPU hours
1
Job type breakdown Info
Success
Failed
Running
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Python Shell
Glue ETL
Glue Streaming
Worker type breakdown Info
Success
Failed
Running
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Standard
G.1X
G.2X


Sample project - 1
Dataset: dataset-met-objects
Sample: Random sample (1,000 rows)
Create job
LINEAGE
ACTIONS
DATASETS
PROJECTS
RECIPES
JOBS
WHAT'S NEW
UNDO
REDO
FILTER
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
UNNEST
PIVOT
GROUP
JOIN
UNION
MORE
RECIPE
Viewing 13 columns 1,000 rows
SAMPLE
GRID
SCHEMA
PROFILE
SOURCE
object name
object number
title
Rename
Sort
Format
Clean
Extract
Remove or fill missing values
Remove or replace invalid values
Remove duplicate values
Split column
Create a flag column
Word tokenization
Categorical mapping
Move column
Hide column
Remove
Special characters
Numbers
Letters
Custom values
Punctuation
Leading and trailing white space
Consecutive white space
Leading and trailing quotation marks
Add
Add prefix (leading characters)
Add suffix (trailing characters)
Leading and trailing single quotation marks
Leading and trailing double
Painting
Coffeepot
Candlestick
Furniture hardware
Woven piece
Drawing
Painting, miniature
Drawing
Coffee cup and saucer
Drawing
Side Chair
Cachepot
Vase
Plate
Dinner plate
Pitcher
Dish
20.109.5
36.22.20
Column details
object name
The statistics below are only on the sample data.
Column statistics
Recommendations
Data quality
VALID VALUES
1000 100%
MISSING VALUES
0 0%
Value distribution
Distinct 254 Unique 145 Total 1,000


Graph
Details
History
Legend:
Start
Trigger
Job
Crawler
Incomplete
Error
Deleting
Remove
Action
WaitToTriggerJob
GetNewSalesData
DataSynced
JoinSalesInventoryAnditem...
WatchForJoinJobCompletion
CrawlSalesInventoryProduc...
WaitForCrawlerCompletion
GenerateReportForCFO
UpdateItemTable
UpdateInventoryTable

Workflows > ReportGenerationPipeline > Run: wr_2aa2fb069b8090652f0537a0cf57e0d16b8c901dc0c3a34fe14af67a87121b0e > Graph

Stop run

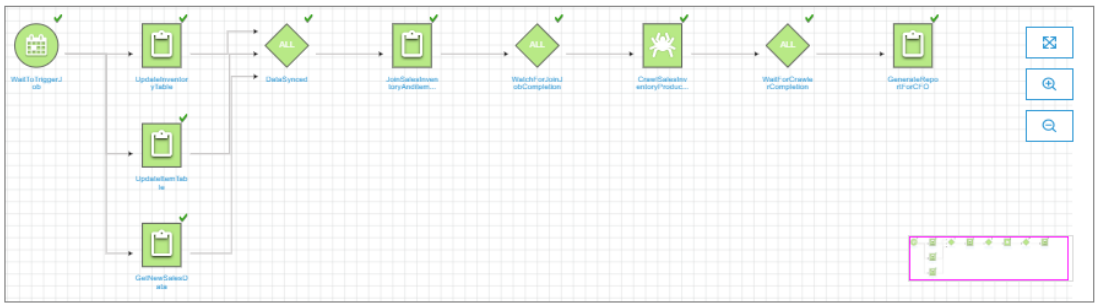
Workflow name	ReportGenerationPipeline
Previous run ID	-
Run ID	wr_2aa2fb069b8090652f0537a0cf57e0d16b8c901dc0c3a34fe14af67a87121b0e
Run status	Completed
Error message	-
Start time	Wed, 05 May 2021 00:00:04 GMT
End time	Wed, 05 May 2021 00:20:27 GMT
Execution time	20 Minutes
Run properties	-

Graph

Select the graph nodes to resume and then choose Resume run.

Resume run

Legend: Succeeded | Running | Stopped | Failed | Timeout | Error | Warning | Resume | Not started



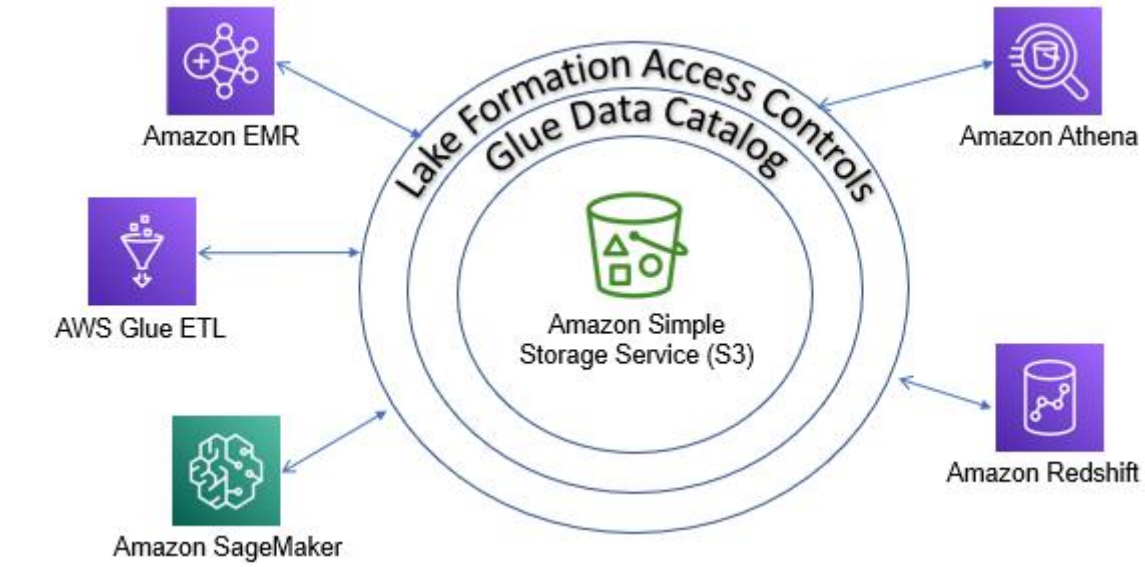
AWS Lake Formation > Blueprints > Use a blueprint

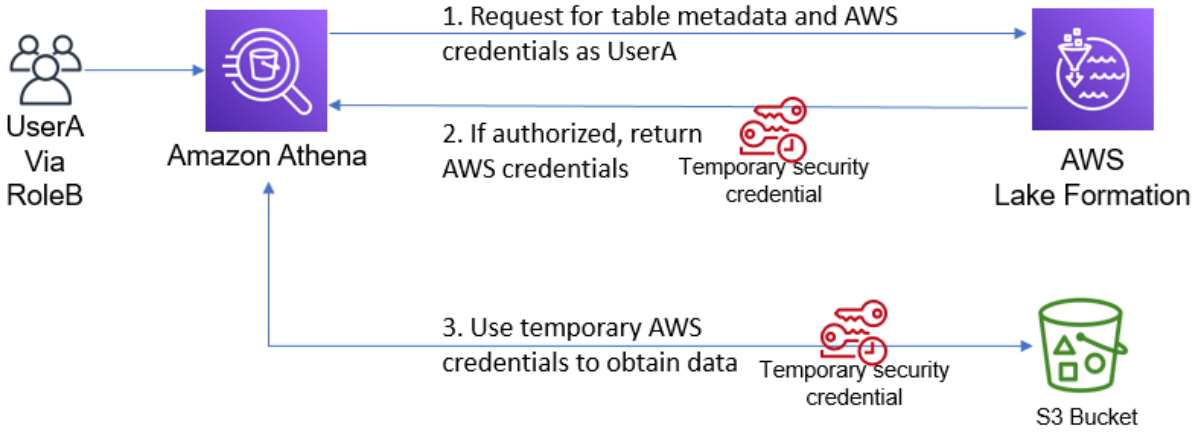
Use a blueprint

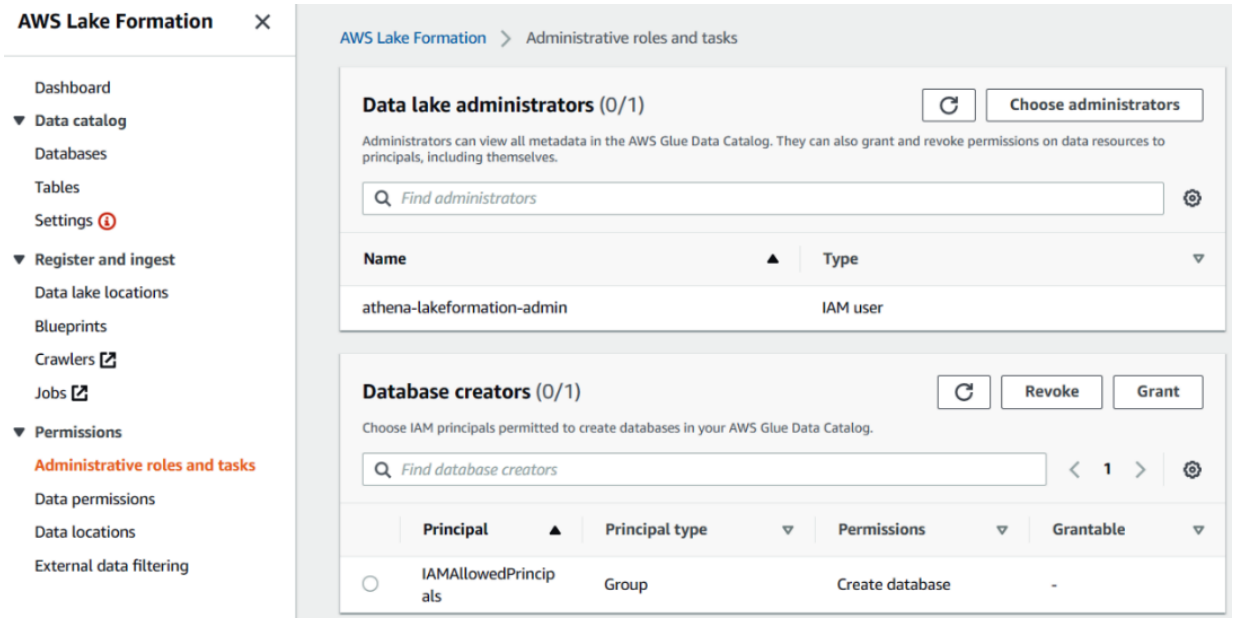
Blueprint type

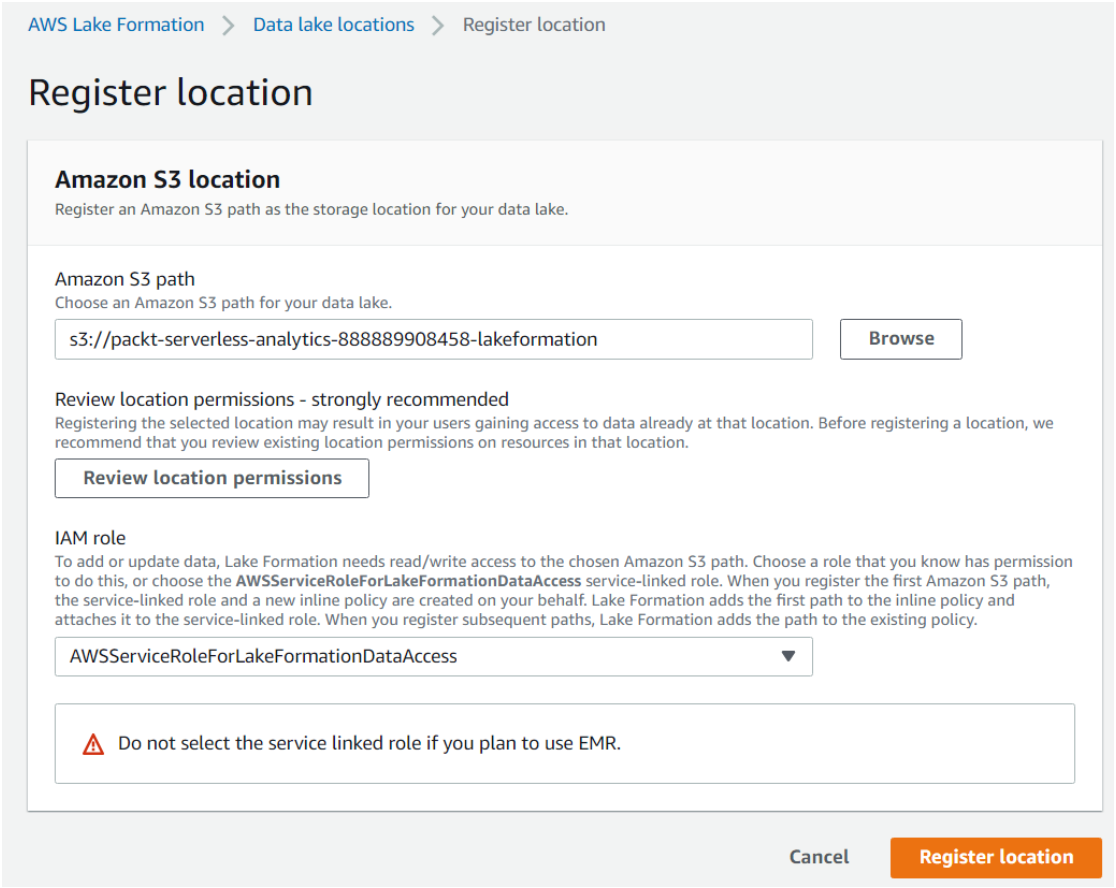
Configure a blueprint to create a workflow.

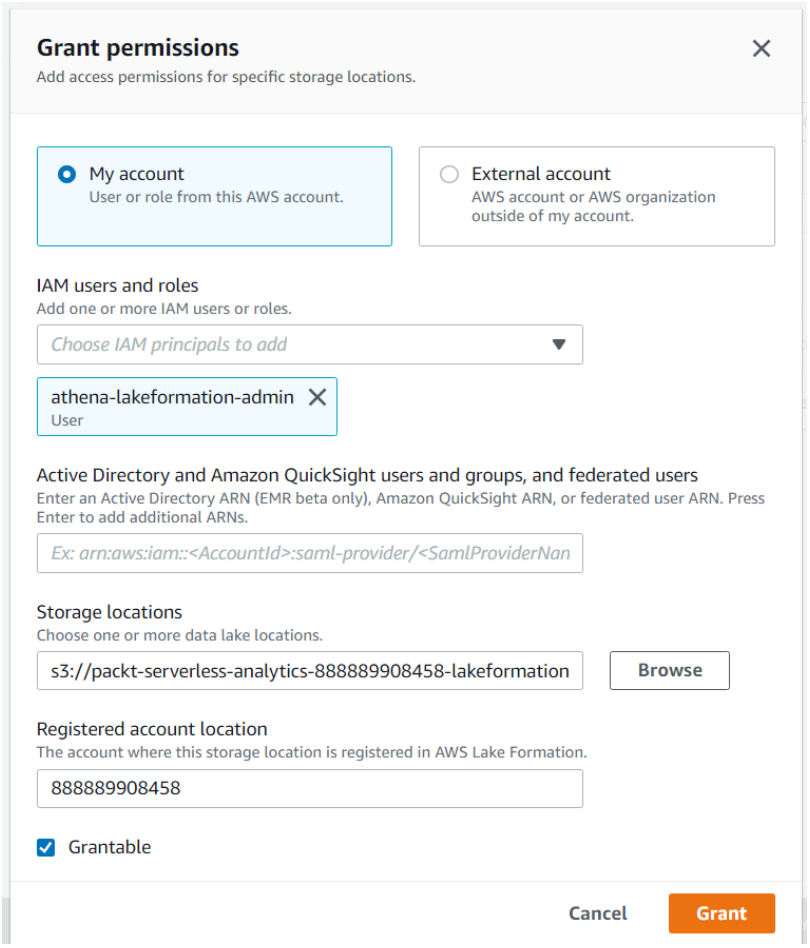
- (●) **Database snapshot**
  Bulk load data to your data lake from MySQL, PostgreSQL, Oracle, and Microsoft SQL Server databases.
- ( ) **Incremental database**
  Load new data to your data lake from MySQL, PostgreSQL, Oracle, and SQL Server databases.
- ( ) **AWS CloudTrail**
  Bulk load data from AWS CloudTrail sources.
- ( ) **Classic Load Balancer logs**
  Load data from Classic Load Balancer logs.
- ( ) **Application Load Balancer logs**
  Load data from Application Load Balancer logs.

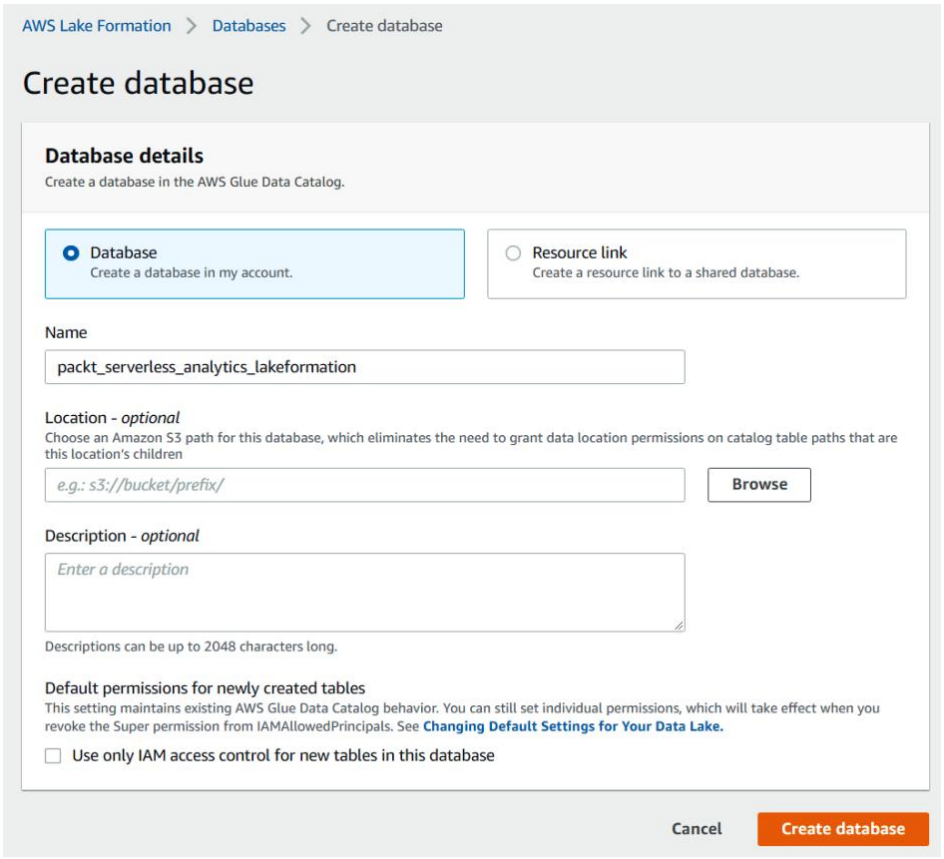

Amazon EMR
Amazon Athena
Lake Formation Access Controls
Glue Data Catalog
AWS Glue ETL
Amazon Simple
Storage Service (S3)
Amazon Redshift
Amazon SageMaker

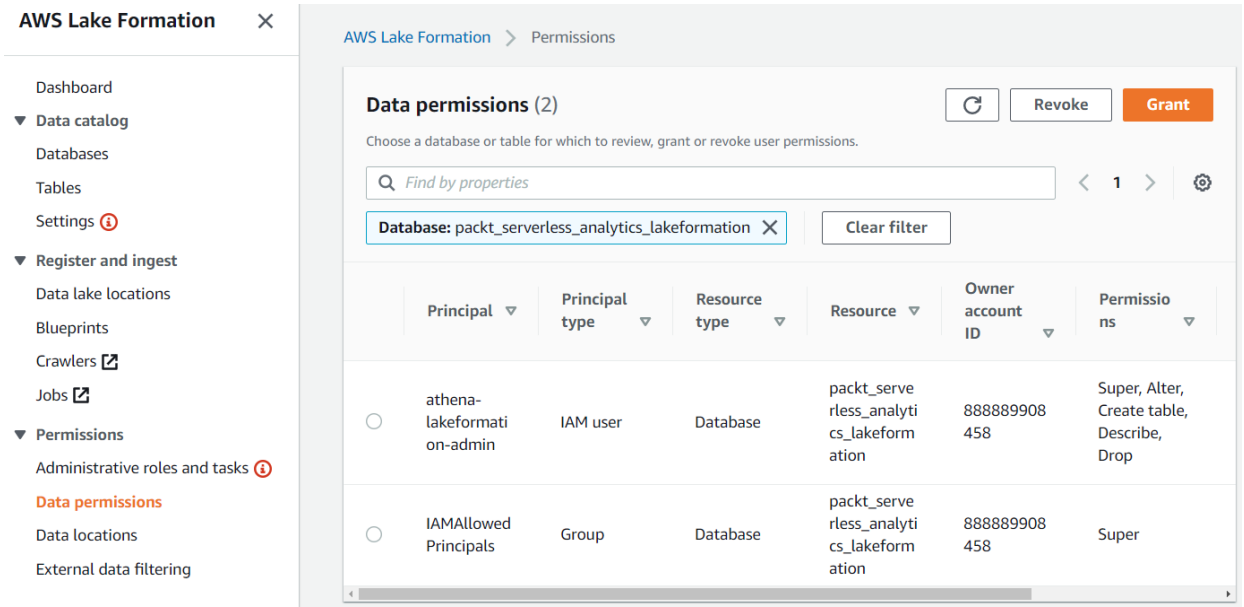

UserA
Via
RoleB
Amazon Athena
1. Request for table metadata and AWS credentials as UserA
2. If authorized, return AWS credentials
Temporary security credential
AWS
Lake Formation
3. Use temporary AWS credentials to obtain data
Temporary security credential
S3 Bucket

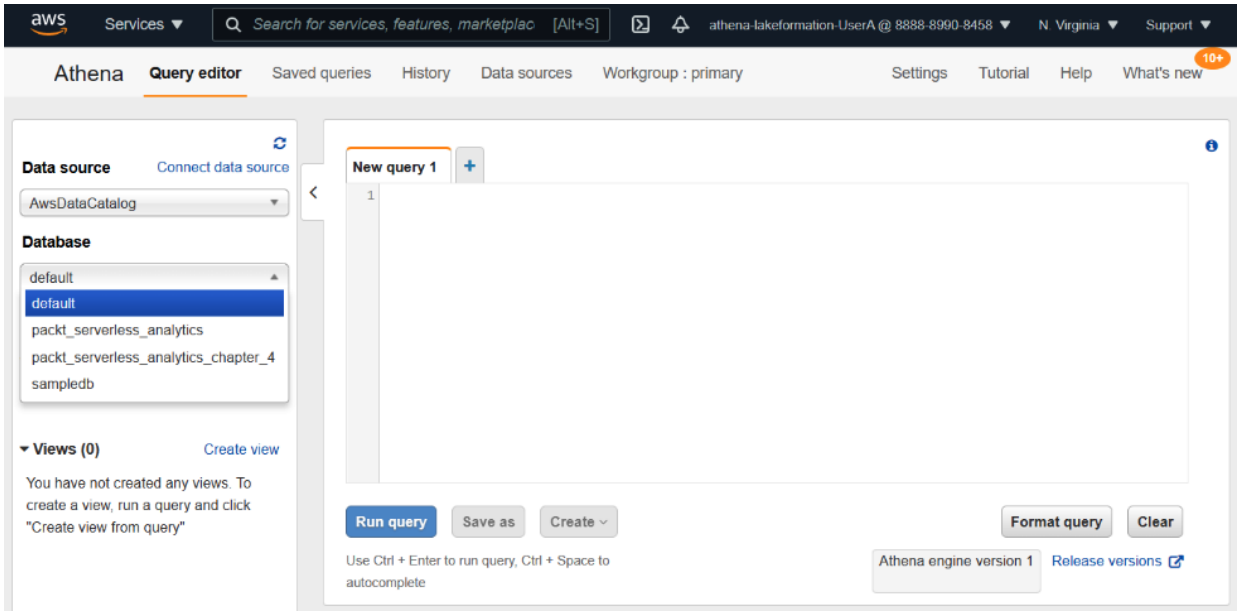

AWS Lake Formation
Dashboard
Data catalog
Databases
Tables
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
Data permissions
Data locations
External data filtering
AWS Lake Formation > Administrative roles and tasks
Data lake administrators (0/1)
Choose administrators
Administrators can view all metadata in the AWS Glue Data Catalog. They can also grant and revoke permissions on data resources to principals, including themselves.
Find administrators
Name
Type
athena-lakeformation-admin
IAM user
Database creators (0/1)
Revoke
Grant
Choose IAM principals permitted to create databases in your AWS Glue Data Catalog.
Find database creators
1
Principal
Principal type
Permissions
Grantable
IAMAllowedPrincipals
Group
Create database
-


AWS Lake Formation
Data lake locations
Register location
Register location
Amazon S3 location
Register an Amazon S3 path as the storage location for your data lake.
Amazon S3 path
Choose an Amazon S3 path for your data lake.
s3://packt-serverless-analytics-888889908458-lakeformation
Browse
Review location permissions - strongly recommended
Registering the selected location may result in your users gaining access to data already at that location. Before registering a location, we recommend that you review existing location permissions on resources in that location.
Review location permissions
IAM role
To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the AWSServiceRoleForLakeFormationDataAccess service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.
AWSServiceRoleForLakeFormationDataAccess
Do not select the service linked role if you plan to use EMR.
Cancel
Register location


Grant permissions
Add access permissions for specific storage locations.
My account
User or role from this AWS account.
External account
AWS account or AWS organization outside of my account.
IAM users and roles
Add one or more IAM users or roles.
Choose IAM principals to add
athena-lakeformation-admin
User
Active Directory and Amazon QuickSight users and groups, and federated users
Enter an Active Directory ARN (EMR beta only), Amazon QuickSight ARN, or federated user ARN. Press Enter to add additional ARNs.
Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderNan
Storage locations
Choose one or more data lake locations.
s3://packt-serverless-analytics-888889908458-lakeformation
Browse
Registered account location
The account where this storage location is registered in AWS Lake Formation.
888889908458
Grantable
Cancel
Grant


AWS Lake Formation > Databases > Create database
Create database
Database details
Create a database in the AWS Glue Data Catalog.
Database
Create a database in my account.
Resource link
Create a resource link to a shared database.
Name
packt_serverless_analytics_lakeformation
Location - optional
Choose an Amazon S3 path for this database, which eliminates the need to grant data location permissions on catalog table paths that are this location's children
e.g.: s3://bucket/prefix/
Browse
Description - optional
Enter a description
Descriptions can be up to 2048 characters long.
Default permissions for newly created tables
This setting maintains existing AWS Glue Data Catalog behavior. You can still set individual permissions, which will take effect when you revoke the Super permission from IAMAllowedPrincipals. See Changing Default Settings for Your Data Lake.
Use only IAM access control for new tables in this database
Cancel
Create database


AWS Lake Formation
Dashboard
Data catalog
Databases
Tables
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
Data permissions
Data locations
External data filtering
AWS Lake Formation > Permissions
Data permissions (2)
Revoke
Grant
Choose a database or table for which to review, grant or revoke user permissions.
Find by properties
1
Database: packt_serverless_analytics_lakeformation
Clear filter
Principal
Principal type
Resource type
Resource
Owner account ID
Permissions
athena-lakeformation-admin
IAM user
Database
packt_serverless_analytics_lakeformation
888889908458
Super, Alter, Create table, Describe, Drop
IAMAllowed Principals
Group
Database
packt_serverless_analytics_lakeformation
888889908458
Super


aws
Services
Search for services, features, marketplac [Alt+S]
athena-lakeformation-UserA @ 8888-8990-8458
N. Virginia
Support
Athena
Query editor
Saved queries
History
Data sources
Workgroup : primary
Settings
Tutorial
Help
What's new
10+
Data source
Connect data source
AwsDataCatalog
Database
default
default
packt_serverless_analytics
packt_serverless_analytics_chapter_4
sampledb
Views (0)
Create view
You have not created any views. To create a view, run a query and click "Create view from query"
New query 1
1
Run query
Save as
Create
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Athena engine version 1
Release versions

Grant permissions: packt_serverless_analytics_lakeformation

Choose the access permissions to grant.

(•) My account
User or role from this AWS account.

( ) External account
AWS account or AWS organization outside of my account.

IAM users and roles

Add one or more IAM users or roles.

Choose IAM principals to add

athena-lakeformation-UserA
User

SAML and Amazon QuickSight users and groups

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

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

Database permissions

Choose the specific access permissions to grant.

[ ] Create table [ ] Alter [ ] Drop [x] Describe

[ ] Super

This permission is the union of the individual permissions above and supersedes them. See here

Grantable permissions

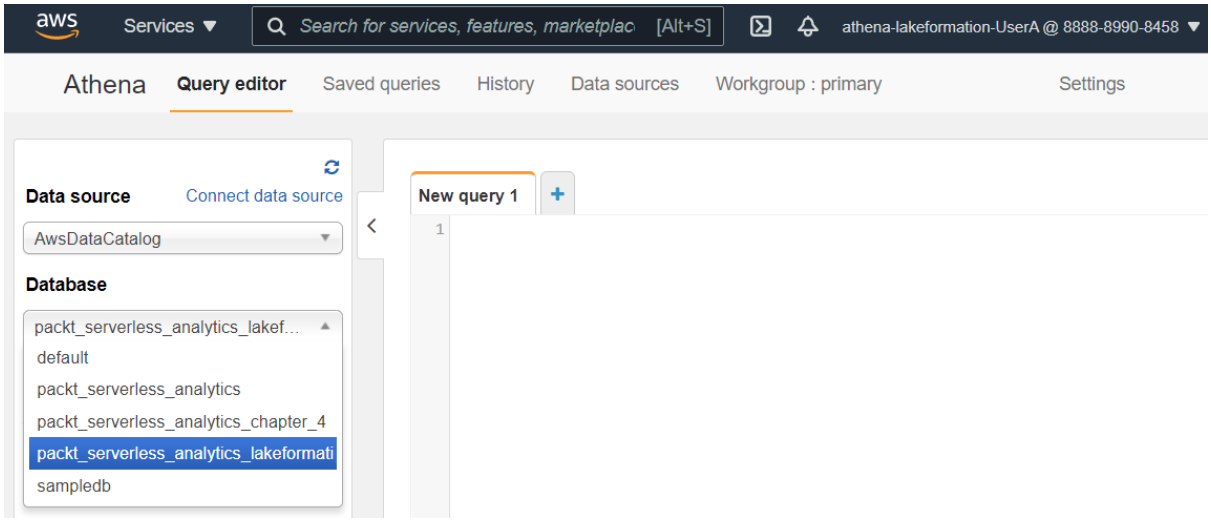
Choose the permissions that may be granted to others.

[ ] Create table [ ] Alter [ ] Drop [ ] Describe

[ ] Super

This permission allows the principal to grant any of the above permissions and supersedes those grantable permissions.

Cancel Grant



Grant permissions: nyc_taxi

Choose the access permissions to grant.

My account
User or role from this AWS account.

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

IAM users and roles

Add one or more IAM users or roles.

Choose IAM principals to add

athena-lakeformation-UserA
User

SAML and Amazon QuickSight users and groups

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

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

Columns - *optional*

Choose filter type

Exclude columns

Exclude columns - *optional*

Grant permissions to access all but the selected columns.

Choose one or more columns

tip_amount
double

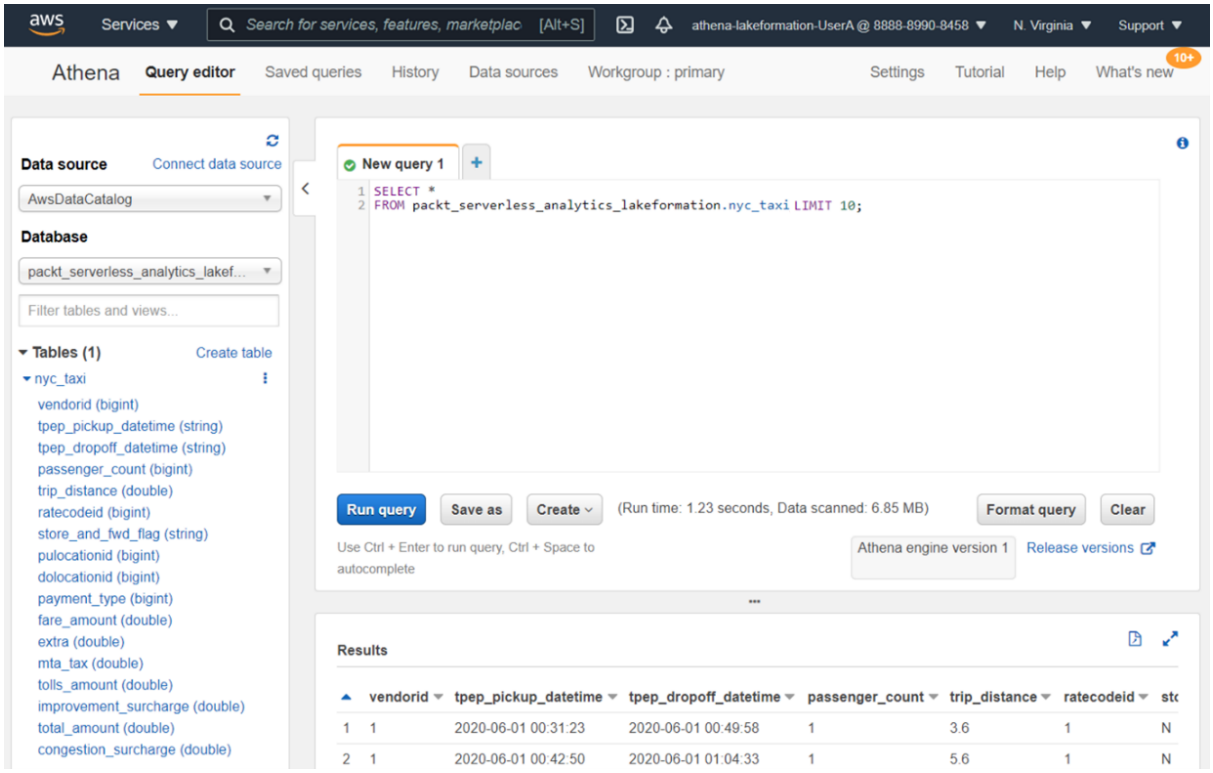
Table permissions

Choose the specific access permissions to grant.

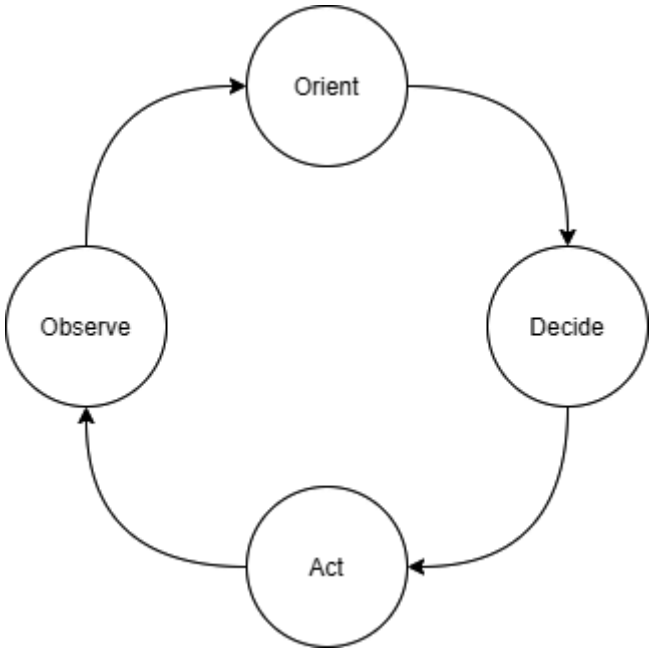
- [ ] Alter
- [ ] Insert
- [ ] Drop
- [ ] Delete
- [x] Select
- [ ] Describe

- [ ] Super

This permission is the union of the individual permissions above and supersedes them. See here


aws
Services
Search for services, features, marketplac [Alt+S]
athena-lakeformation-UserA @ 8888-8990-8458
N. Virginia
Support
Athena
Query editor
Saved queries
History
Data sources
Workgroup : primary
Settings
Tutorial
Help
What's new
10+
Data source
Connect data source
AwsDataCatalog
Database
packt_serverless_analytics_lakef...
Filter tables and views...
Tables (1)
Create table
nyc_taxi
vendorid (bigint)
tpep_pickup_datetime (string)
tpep_dropoff_datetime (string)
passenger_count (bigint)
trip_distance (double)
ratecodeid (bigint)
store_and_fwd_flag (string)
pulocationid (bigint)
dolocationid (bigint)
payment_type (bigint)
fare_amount (double)
extra (double)
mta_tax (double)
tolls_amount (double)
improvement_surcharge (double)
total_amount (double)
congestion_surcharge (double)
New query 1
1 SELECT *
2 FROM packt_serverless_analytics_lakeformation.nyc_taxi LIMIT 10;
Run query
Save as
Create
(Run time: 1.23 seconds, Data scanned: 6.85 MB)
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Athena engine version 1
Release versions
Results
vendorid
tpep_pickup_datetime
tpep_dropoff_datetime
passenger_count
trip_distance
ratecodeid
1 1 2020-06-01 00:31:23 2020-06-01 00:49:58 1 3.6 1 N
2 1 2020-06-01 00:42:50 2020-06-01 01:04:33 1 5.6 1 N

Chapter 7: Ad Hoc Analytics



QuickSight region

Select a region.

US East (N. Virginia)

QuickSight account name

packt_serverless_analytics

You will need this for you and others to sign in.

Notification email address

amazon.athena.book@gmail.com

For QuickSight to send important notifications.

Enable invitation by email

☑ Allow inviting new users by email.This setting cannot be changed after sign-up is complete.

☐ Enable autodiscovery of data and users in your Amazon Redshift, Amazon RDS, and AWS IAM services.

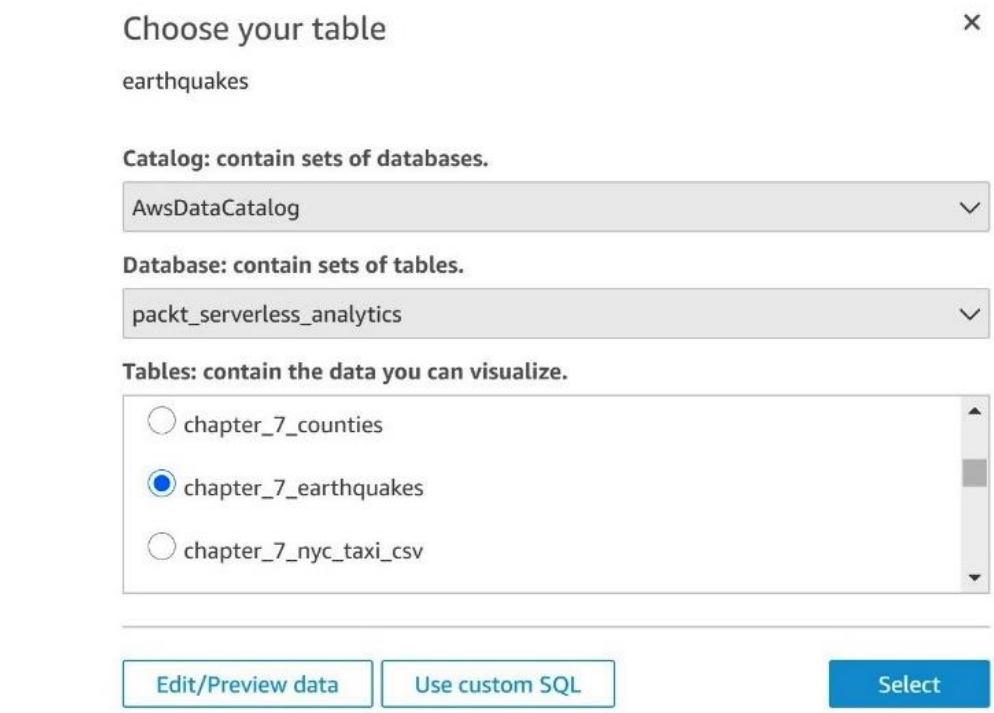
☑ Amazon Athena
Enables QuickSight access to Amazon Athena databases

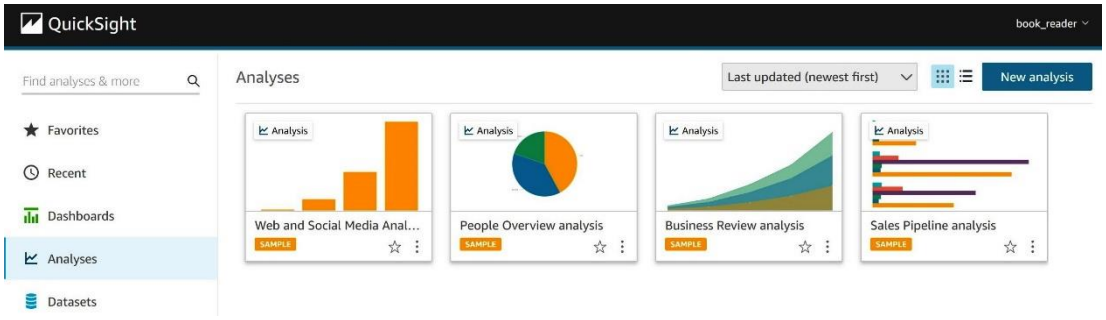
Please ensure the right Amazon S3 buckets are also enabled for QuickSight.

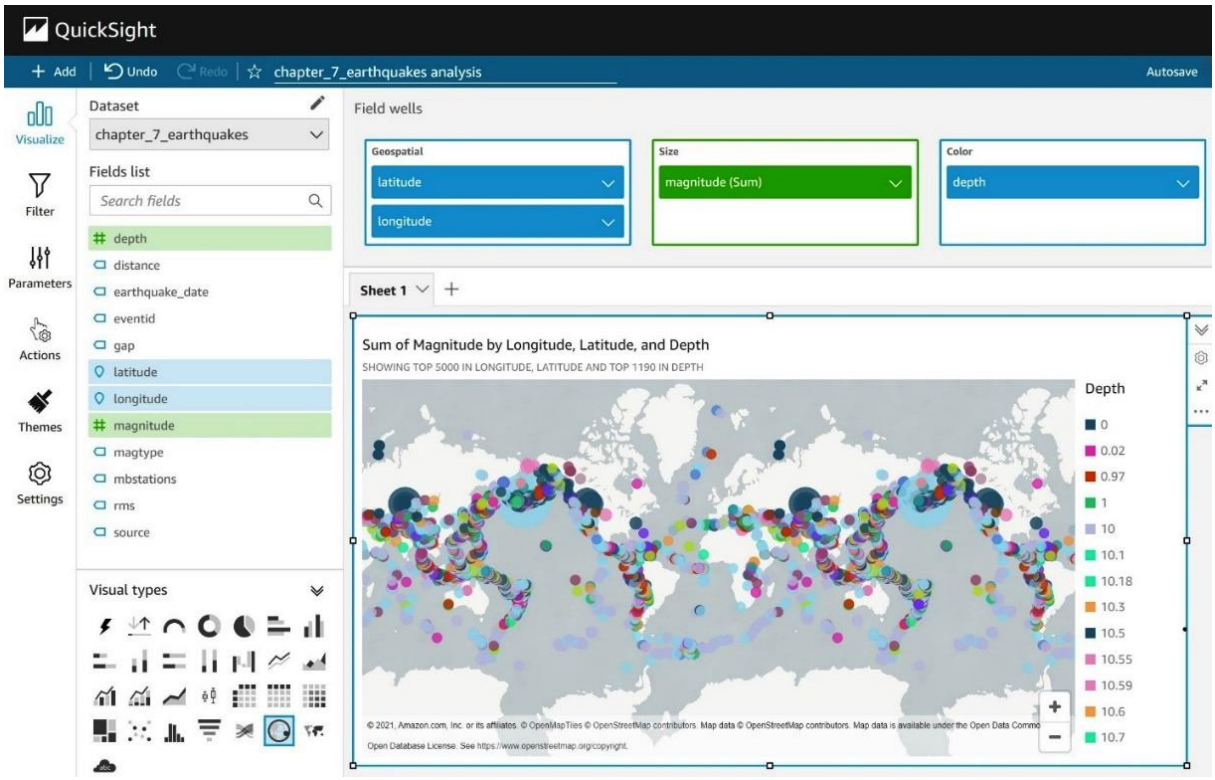
☑ Amazon S3 (1 buckets selected) Choose S3 bucket
Enables QuickSight to auto-discover your Amazon S3 buckets

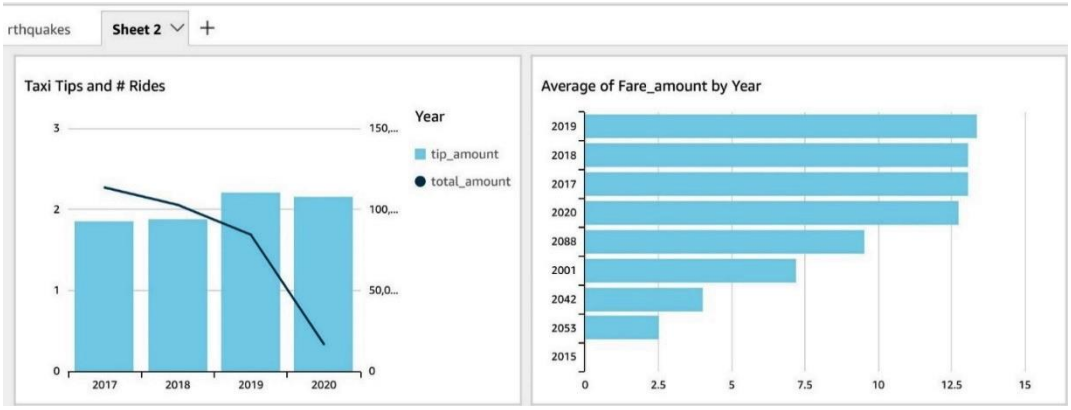
☐ Amazon S3 Storage Analytics
Enables QuickSight to visualize your S3 Storage Analytics data

☐ AWS IoT Analytics
Enables QuickSight to visualize your IoT Analytics data


Choose your table
earthquakes
Catalog: contain sets of databases.
AwsDataCatalog
Database: contain sets of tables.
packt_serverless_analytics
Tables: contain the data you can visualize.
chapter_7_counties
chapter_7_earthquakes
chapter_7_nyc_taxi_csv
Edit/Preview data
Use custom SQL
Select


QuickSight
book_reader
Find analyses & more
Favorites
Recent
Dashboards
Analyses
Datasets
Analyses
Last updated (newest first)
New analysis
Web and Social Media Anal...
People Overview analysis
Business Review analysis
Sales Pipeline analysis
SAMPLE


QuickSight
Add
Undo
Redo
chapter_7_earthquakes analysis
Autosave
Visualize
Filter
Parameters
Actions
Themes
Settings
Dataset
chapter_7_earthquakes
Fields list
Search fields
depth
distance
earthquake_date
eventid
gap
latitude
longitude
magnitude
magtype
mbstations
rms
source
Visual types
Field wells
Geospatial
latitude
longitude
Size
magnitude (Sum)
Color
depth
Sheet 1
Sum of Magnitude by Longitude, Latitude, and Depth
SHOWING TOP 5000 IN LONGITUDE, LATITUDE AND TOP 1190 IN DEPTH
Depth
0
0.02
0.97
1
10
10.1
10.18
10.3
10.5
10.55
10.59
10.6
10.7
© 2021, Amazon.com, Inc. or its affiliates. © OpenMapTiles © OpenStreetMap contributors. Map data © OpenStreetMap contributors. Map data is available under the Open Data Commo
Open Database License. See https://www.openstreetmap.org/copyright.



Create role

1 2 3 4

Select type of trusted entity

AWS service
EC2, Lambda and others

Another AWS account
Belonging to you or 3rd party

Web identity
Cognito or any OpenID provider

SAML 2.0 federation
Your corporate directory

Allows AWS services to perform actions on your behalf. Learn more

Choose a use case

Common use cases

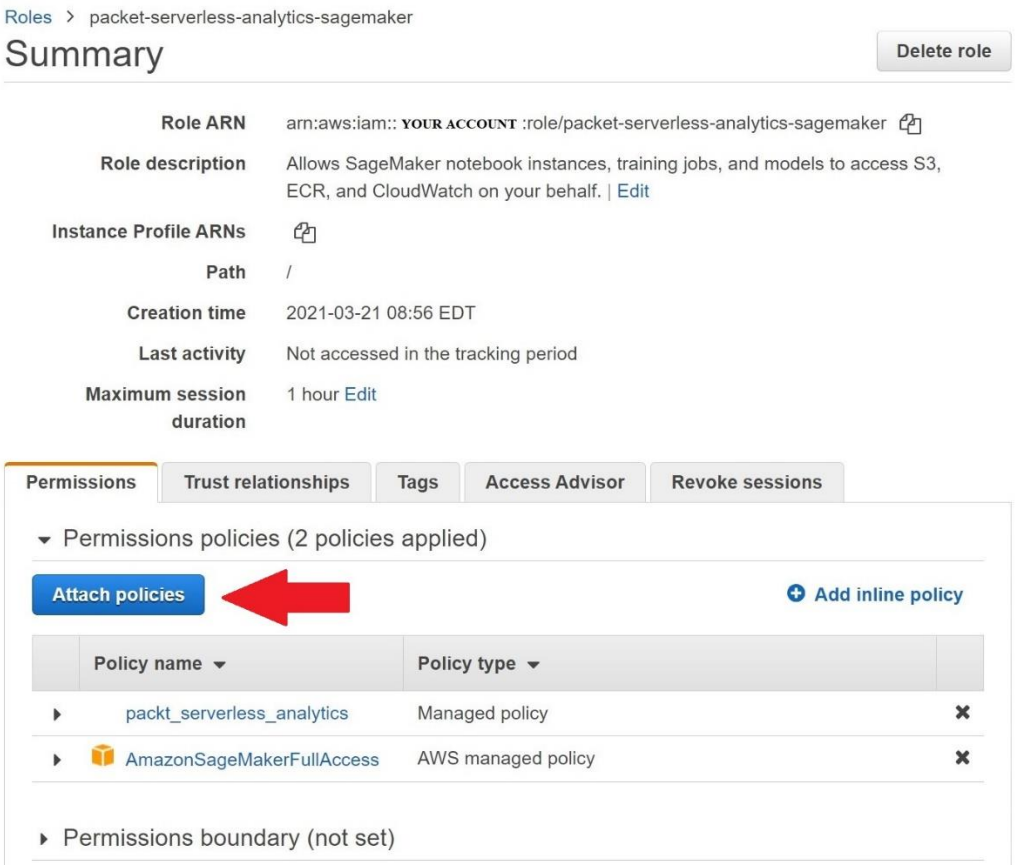
EC2
Allows EC2 instances to call AWS services on your behalf.

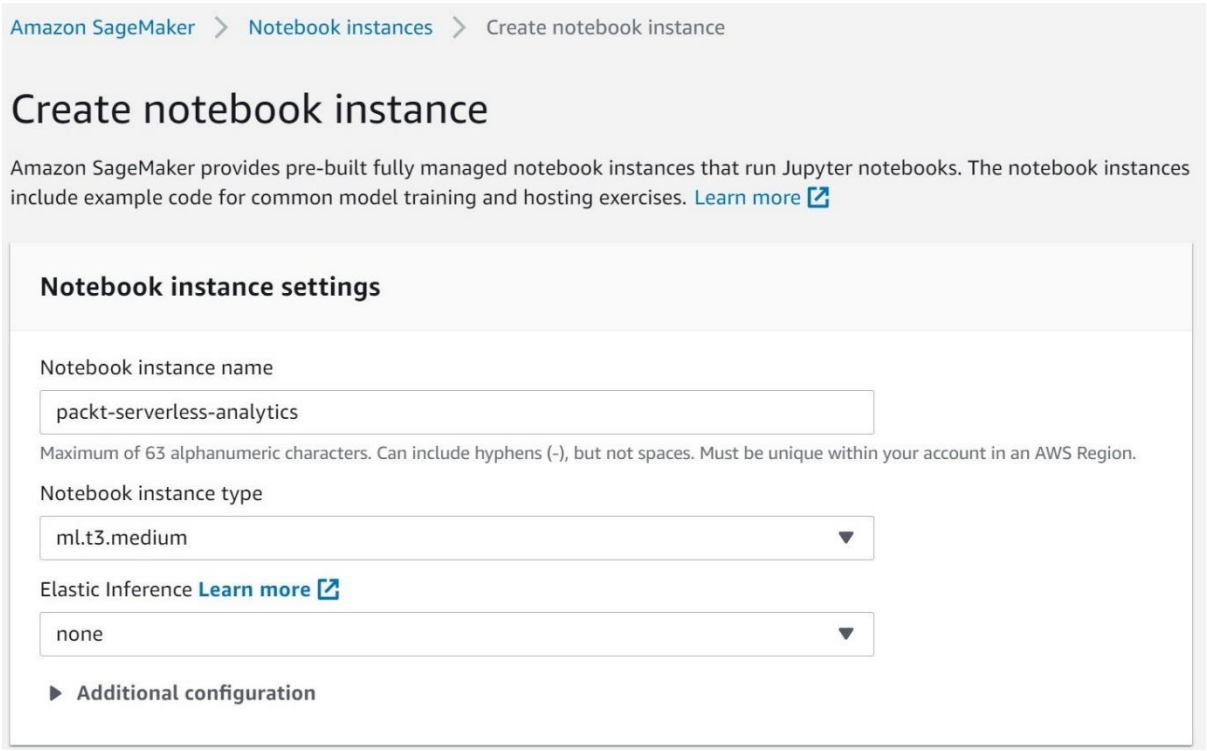
Lambda
Allows Lambda functions to call AWS services on your behalf.

Or select a service to view its use cases

API Gateway	CloudWatch Events	EKS	IoT Things Graph	Redshift
AWS Backup	CodeBuild	EMR	KMS	Rekognition
AWS Chatbot	CodeDeploy	ElastiCache	Kinesis	RoboMaker
AWS Marketplace	CodeGuru	Elastic Beanstalk	Lake Formation	S3
AWS Support	CodeStar Notifications	Elastic Container Registry	Lambda	SMS
Amplify	Comprehend	Elastic Container Service	Lex	SNS
AppStream 2.0	Config	Elastic Transcoder	License Manager	SWF
AppSync	Connect	ElasticLoadBalancing	MQ	**SageMaker**

Cancel **Next: Permissions**


Roles > packet-serverless-analytics-sagemaker
Summary
Delete role
Role ARN
arn:aws:iam:: YOUR ACCOUNT :role/packet-serverless-analytics-sagemaker
Role description
Allows SageMaker notebook instances, training jobs, and models to access S3, ECR, and CloudWatch on your behalf. | Edit
Instance Profile ARNs
Path
/
Creation time
2021-03-21 08:56 EDT
Last activity
Not accessed in the tracking period
Maximum session duration
1 hour Edit
Permissions
Trust relationships
Tags
Access Advisor
Revoke sessions
Permissions policies (2 policies applied)
Attach policies
Add inline policy
Policy name
Policy type
packt_serverless_analytics
Managed policy
AmazonSageMakerFullAccess
AWS managed policy
Permissions boundary (not set)


Amazon SageMaker > Notebook instances > Create notebook instance
Create notebook instance
Amazon SageMaker provides pre-built fully managed notebook instances that run Jupyter notebooks. The notebook instances include example code for common model training and hosting exercises. Learn more
Notebook instance settings
Notebook instance name
packt-serverless-analytics
Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.
Notebook instance type
ml.t3.medium
Elastic Inference Learn more
none
Additional configuration

Permissions and encryption

IAM role

Notebook instances require permissions to call other services including SageMaker and S3. Choose a role or let us create a role with the **AmazonSageMakerFullAccess** IAM policy attached.

packet-serverless-analytics-sagemaker

Root access - *optional*

- (•) Enable - Give users root access to the notebook
- ( ) Disable - Don't give users root access to the notebook
  Lifecycle configurations always have root access

Encryption key - *optional*

Encrypt your notebook data. Choose an existing KMS key or enter a key's ARN.

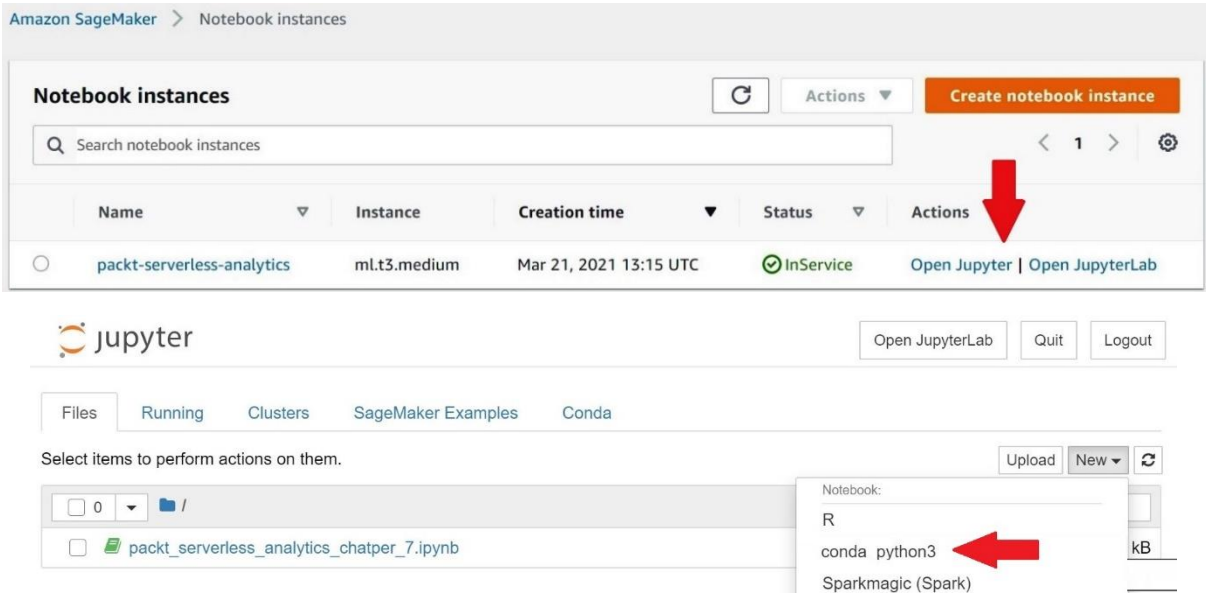
No Custom Encryption

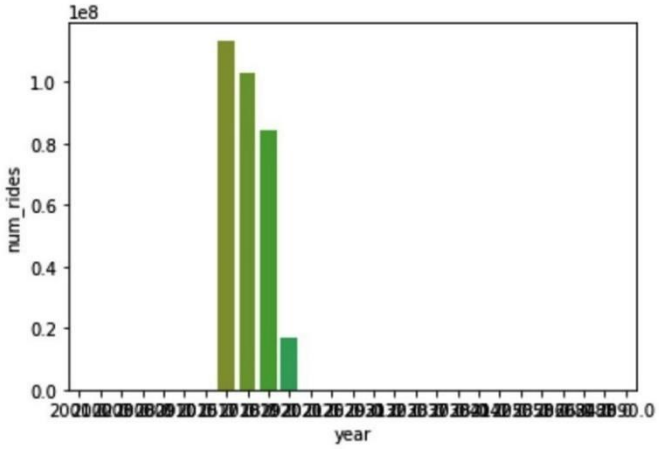
▶ **Network** - *optional*

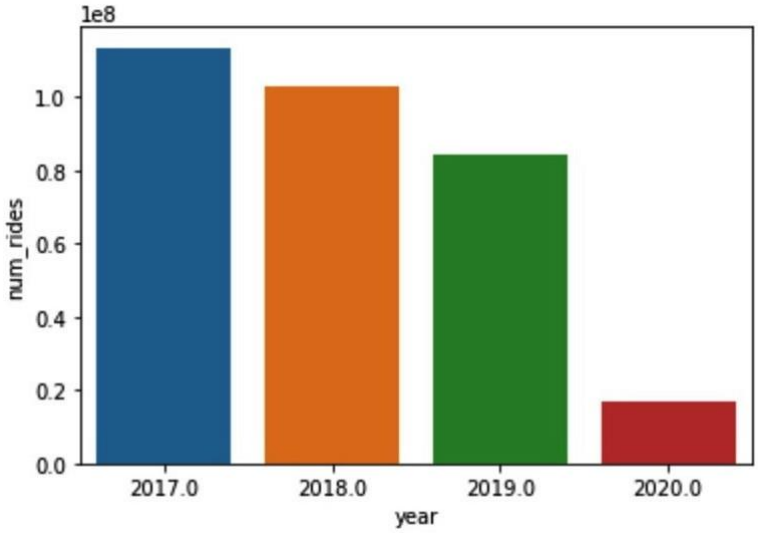
▶ **Git repositories** - *optional*

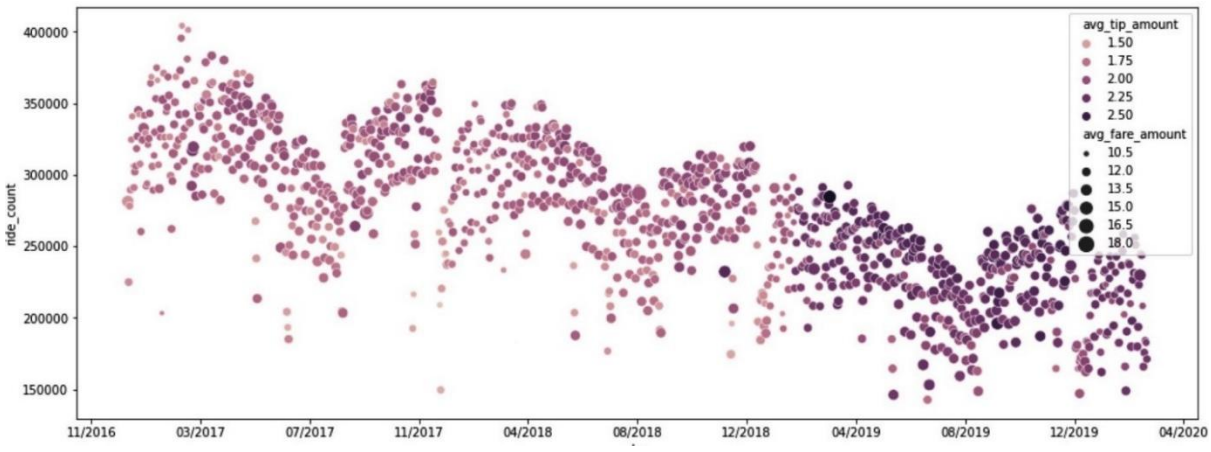
▶ **Tags** - *optional*

Cancel | Create notebook instance




1e8
num_rides
year


1e8
num_rides
2017.0
2018.0
2019.0
2020.0
year


ride_count
11/2016
03/2017
07/2017
11/2017
04/2018
08/2018
12/2018
04/2019
08/2019
12/2019
04/2020
avg_tip_amount
avg_fare_amount

	hour_val	duration	trip_distance	fare_amount	tip_amount	cnt
hour_val	1.000000	0.005754	-0.343937	-0.063703	0.269899	0.691085
duration	0.005754	1.000000	-0.011370	0.012235	0.019973	0.017925
trip_distance	-0.343937	-0.011370	1.000000	0.270629	0.327298	-0.721152
fare_amount	-0.063703	0.012235	0.270629	1.000000	0.210351	-0.150394
tip_amount	0.269899	0.019973	0.327298	0.210351	1.000000	0.190958
cnt	0.691085	0.017925	-0.721152	-0.150394	0.190958	1.000000

year	num_rides	zscore
2017	1.13E+08	3.274275
2018	1.03E+08	2.93152
2019	84397884	2.341594
2020	16847996	0.176508
2038	4	-0.3635
2058	3	-0.3635
2088	2	-0.3635
2042	1	-0.3635

Chapter 8: Querying Unstructured and Semi-Structured Data

Results

	customer_id	first_name	last_name	email	addresses
1	1	Rori	Struss	rstruss0@somedomain.com	[{address=20532 Debra Place, city=Boston, state=MA,
2	2	Maribeth	Myers	mmyers1@somedomain.com	[{address=4800 Montana Terrace, city=Anniston, state=
3	3	Pearle	Merrell	pmerrell2@somedomain.com	[{address=855 Upham Junction, city=Minneapolis, state=

Results

	State	Count
1	TX	11
2	CA	10
3	FL	7
4	NY	7
5	VA	6

Results

	customer_id	extrainfo
1	1	{"is_pinnacle_customer":"true","pinnacle_id":"12423"}
2	5	{"is_pinnacle_customer":"true","pinnacle_id":"543433"}
3	11	{"is_pinnacle_customer":"true","pinnacle_id":"544333"}
4	15	{"is_pinnacle_customer":"true","pinnacle_id":"667645"}
5	23	{"is_pinnacle_customer":"true","pinnacle_id":"2342322"}

Results

	pinnacle_id
1	12423
2	543433
3	544333
4	667645
5	2342322
6	234221

Results

	timestamp	item_id	customer_id	price	shipping_price	discount_code
1	2020-05-24T02:27:34Z	4	50	19.36	1.43	
2	2020-05-24T08:02:05Z	5	43	13.39	1.51	
3	2020-05-24T22:49:15Z	1	72	11.58	2.17	
4	2020-05-25T19:05:42Z	5	4	16.85	1.67	54162-269
5	2020-05-26T10:49:05Z	3	82	16.44	1.53	

Results

	start_date	end_date	marketing_id	description
1	2021-01-30 11:00:00Z	2021-01-30 18:00:00Z	193223	Extra adverstisement on search engine.
2	2021-06-19 00:00:00Z	2021-06-19 23:59:59Z	543222	A test marketing compaign, in order to observe im
3	2020-12-14 00:00:00Z	2020-12-14 23:59:59Z	432543	Pinnacle Day.
4	2020-07-01 00:00:00Z	2020-07-01 23:59:59Z	537734	Advertise on Canada Day.
5	2020-12-25 00:00:00Z	2020-12-25 23:59:59Z	796456	Advertise on Christmas Eve.

Results

	sales_date	has_marketing_campaign	number_of_sales	states
1	2021-01-30 00:00:00.000 UTC	true	9	{TX=1, AZ=1, GA=1, VA=2, CO=2, CA=2}
2	2020-12-04 00:00:00.000 UTC	true	8	{IL=1, IN=1, OH=1, MI=1, CA=2, DC=1}
3	2020-09-07 00:00:00.000 UTC	true	7	{NY=2, ID=2, CA=2}
4	2020-08-03 00:00:00.000 UTC	false	7	{MA=2, TX=2, FL=1, GA=1}
5	2020-07-08 00:00:00.000 UTC	false	7	{IN=1, NY=3, VA=1, NJ=1}

Results

	inventory_id	item_name	available_count
1	"1"	"A simple widget"	"5"
2	"2"	"A more advanced widget	"10"
3	"3"	"The most advanced widget	"1"
4	"4"	"A premium widget"	"0"
5	"5"	"A gold plated widget"	"9"

Results

	inventory_id	item_name	available_count
1	1	A simple widget	5
2	2	A more advanced widget	10
3	3	The most advanced widget	1
4	4	A premium widget	0
5	5	A gold plated widget	9

Results

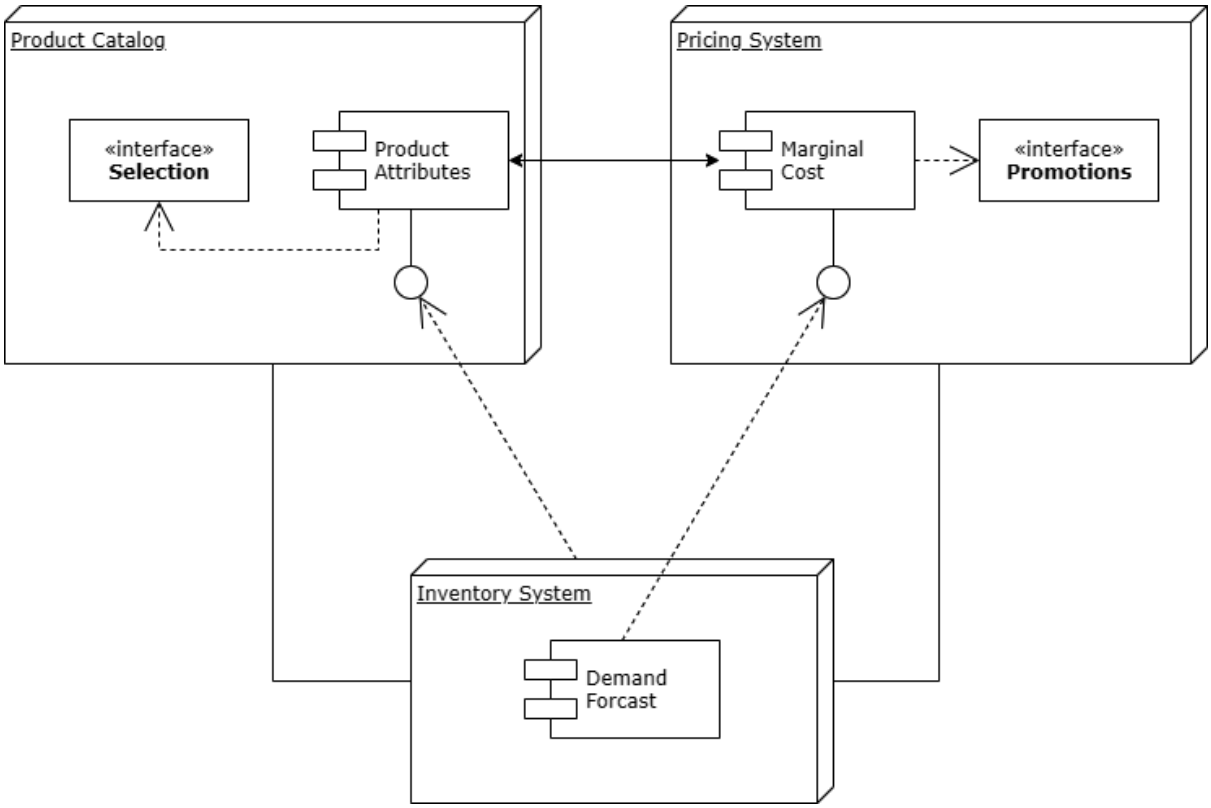
	log_line	$PATH
1	Caused by: ExitCodeException exitCode=143:	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
2	Caused by: ExitCodeException exitCode=143:	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
3	Caused by: ExitCodeException exitCode=143:	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
4	Caused by: ExitCodeException exitCode=143:	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
5	Caused by: java.lang.InterruptedException	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
6	Caused by: java.lang.InterruptedException	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
7	Caused by: java.lang.InterruptedException	s3://aws-logs-XXXXXXXXXX-us-east-1/j-106EEVKIB88X9/i-0ea08c8003c460588/hadoop-yarn/nodemanager-ip-10-0-0-227.log.gz
8	HTTP ERROR: 404	s3://aws-logs-XXXXXXXXXX-us-east-1/j-10WFFA6AT6SM8/i-030875b499532e0b4/daemons/instance-state/instance-state.log-202
9	HTTP ERROR: 404	s3://aws-logs-XXXXXXXXXX-us-east-1/j-10WFFA6AT6SM8/i-030875b499532e0b4/daemons/instance-state/instance-state.log-202

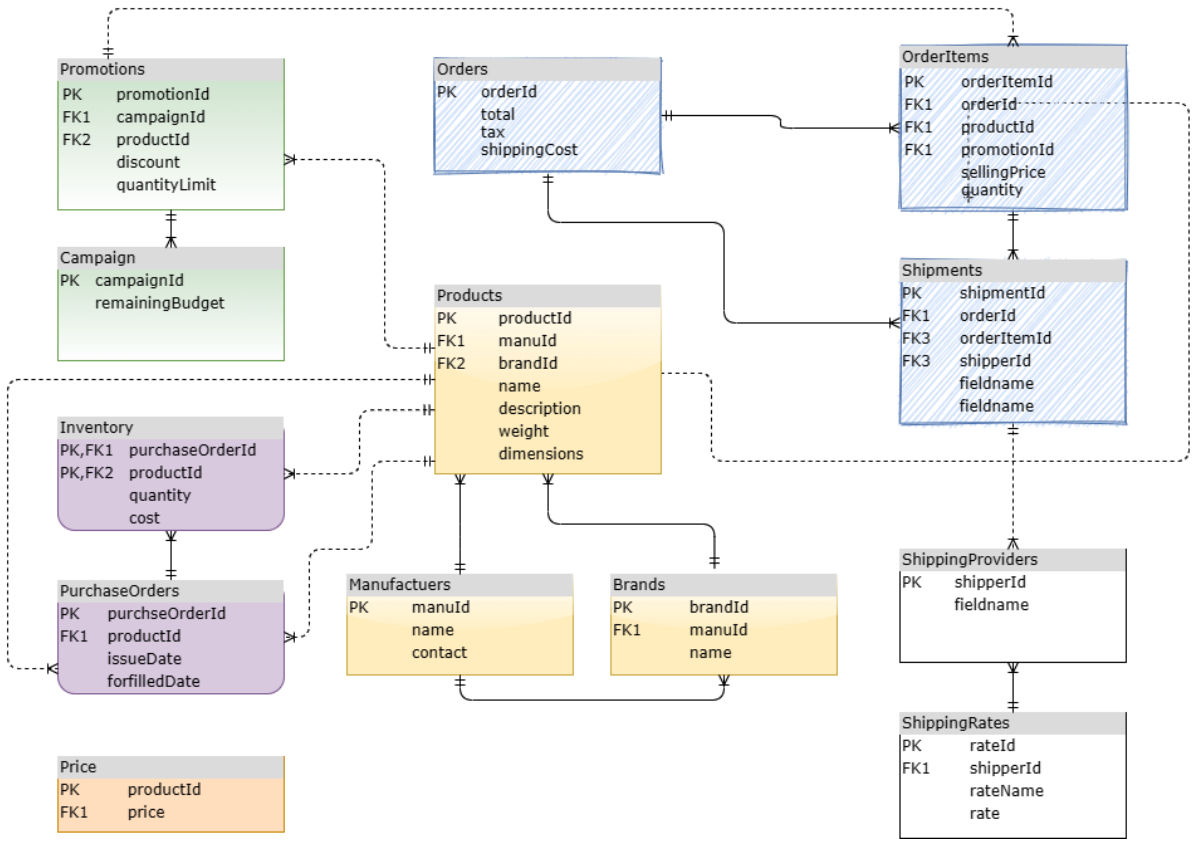
Results

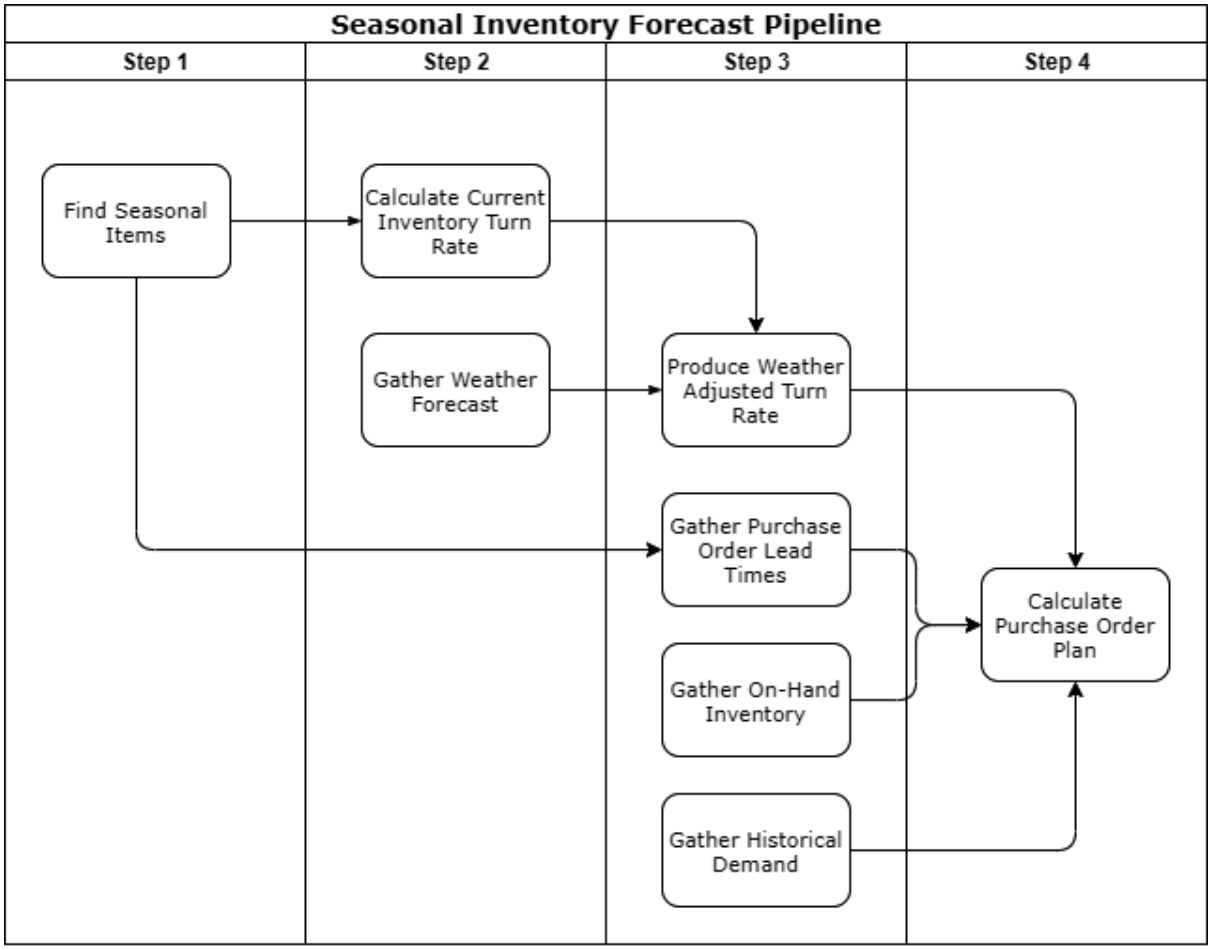
	time_epoch	source_addr	page_visited	referrer
1	1619373564	176.53.241.205	http://www.acmestore.com/product/1	https://www.yahoo.com
2	1620345099	52.96.144.79	http://www.acmestore.com/product/3	https://www.acmestore.com/search?s=awesome
3	1609494659	74.208.254.89	http://www.acmestore.com/product/5	https://www.duckduckgo.com
4	1600553532	159.68.111.113	http://www.acmestore.com/product/1	https://www.duckduckgo.com
5	1616796863	87.233.147.184	http://www.acmestore.com/product/1	https://www.acmestore.com/search?s=awesome

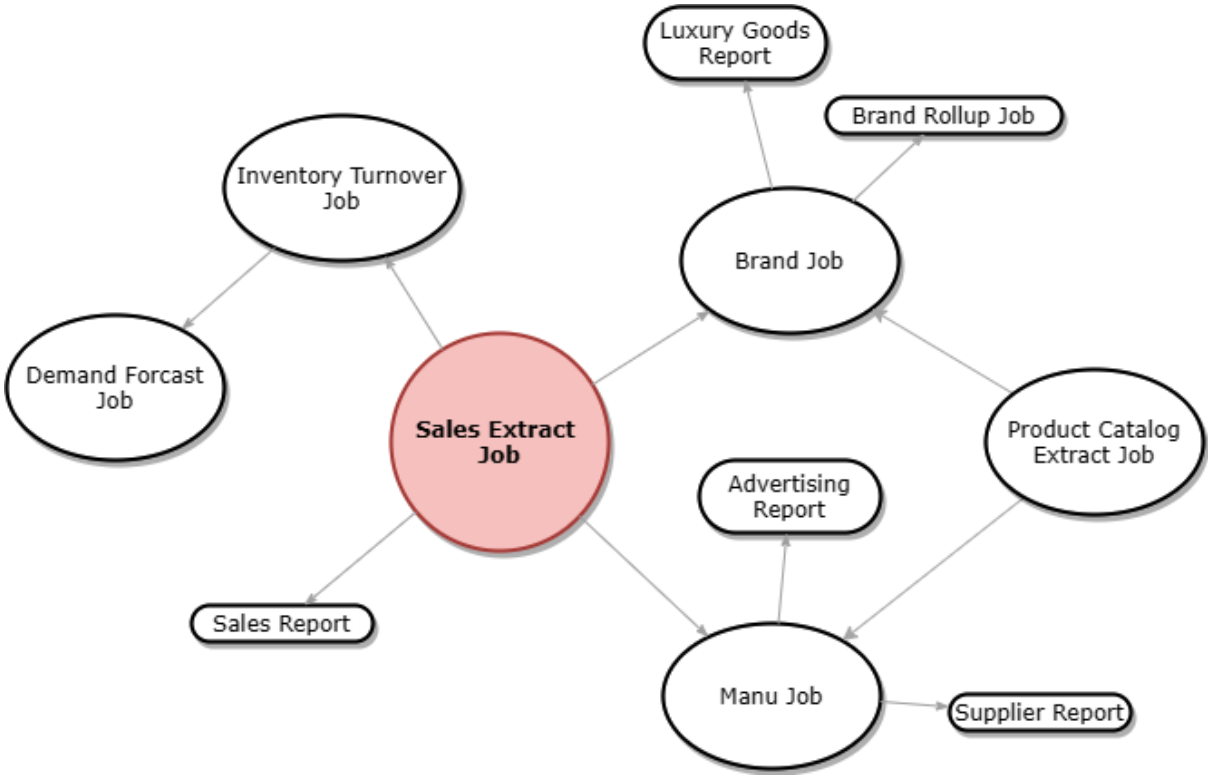
Data Source	Data Format	Description
Customers	JSON	Contains information about customers, such as their names and addresses
Sales	CSV	Contains transactions of customers purchasing widgets
Marketing	TSV	Contains information about a marketing campaign
Inventory	CSV	Contains the number of widgets left in our inventory
Website clicks	Text	Contains information about how customers are using the website

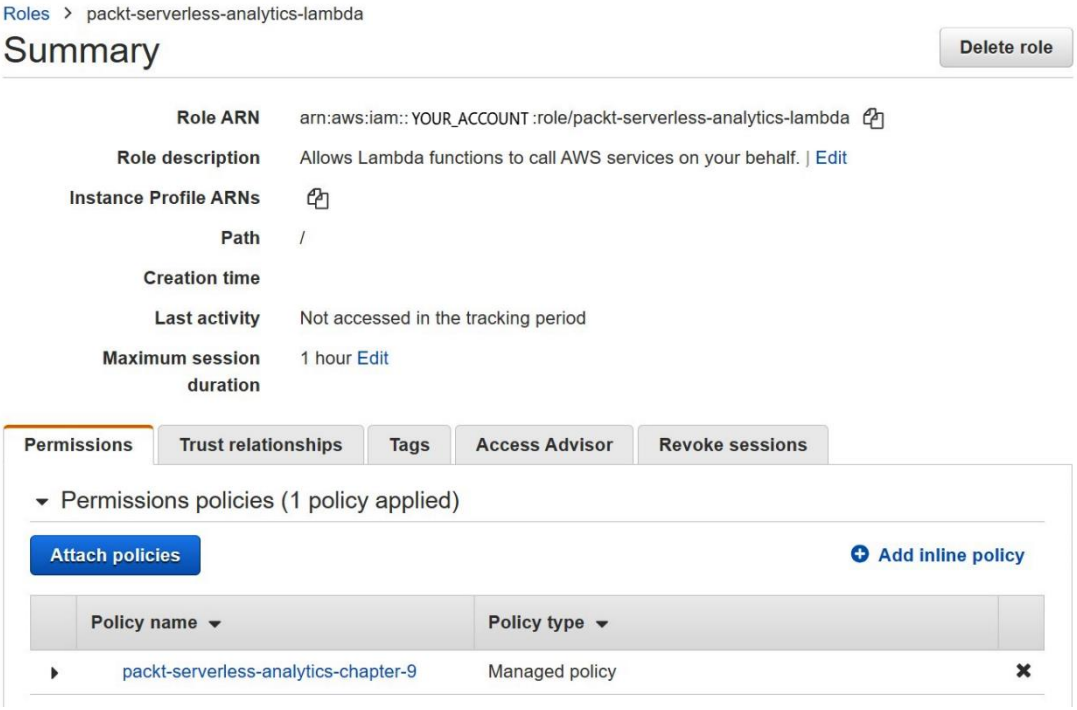
Chapter 9: Serverless ETL Pipelines

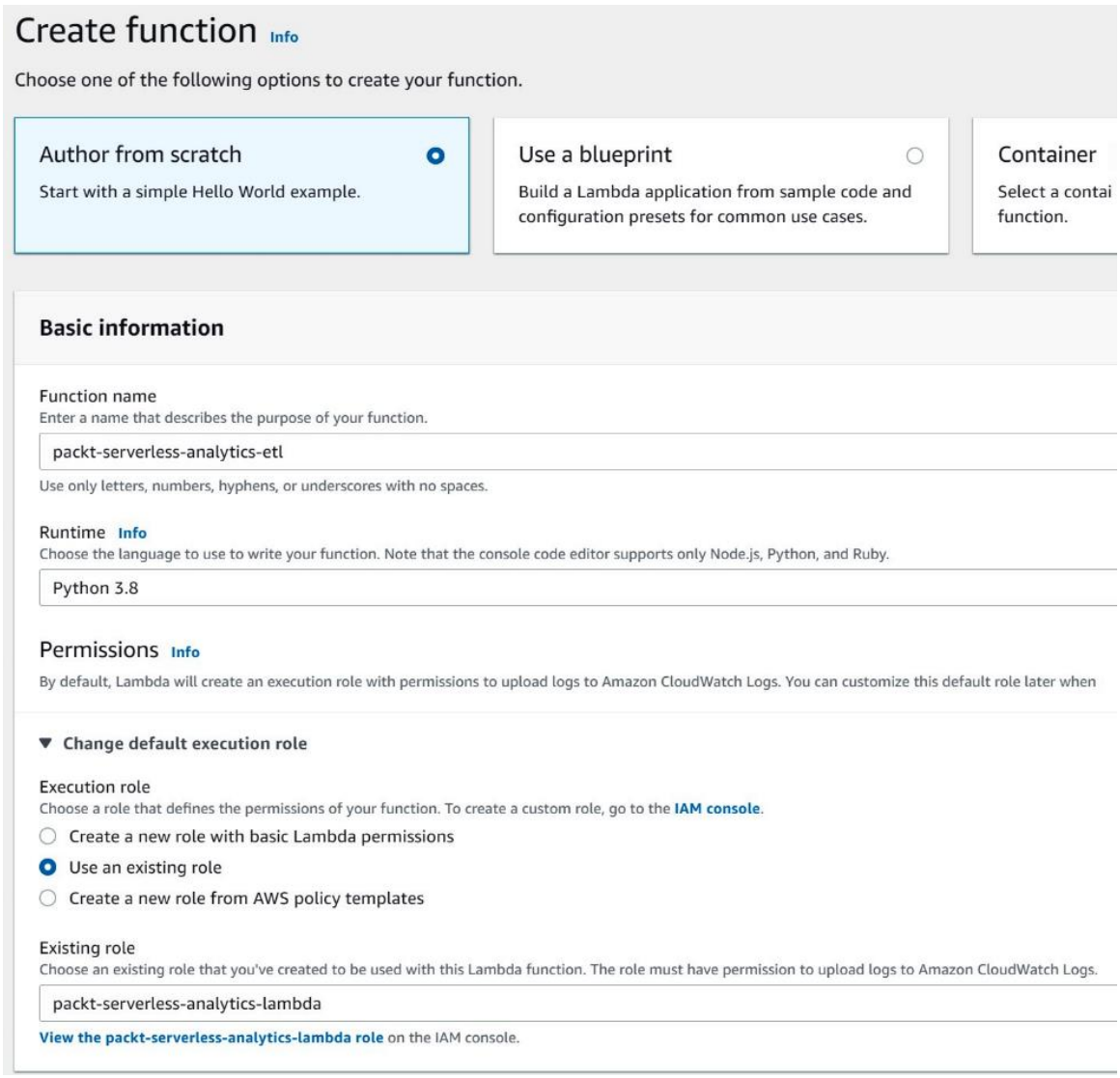


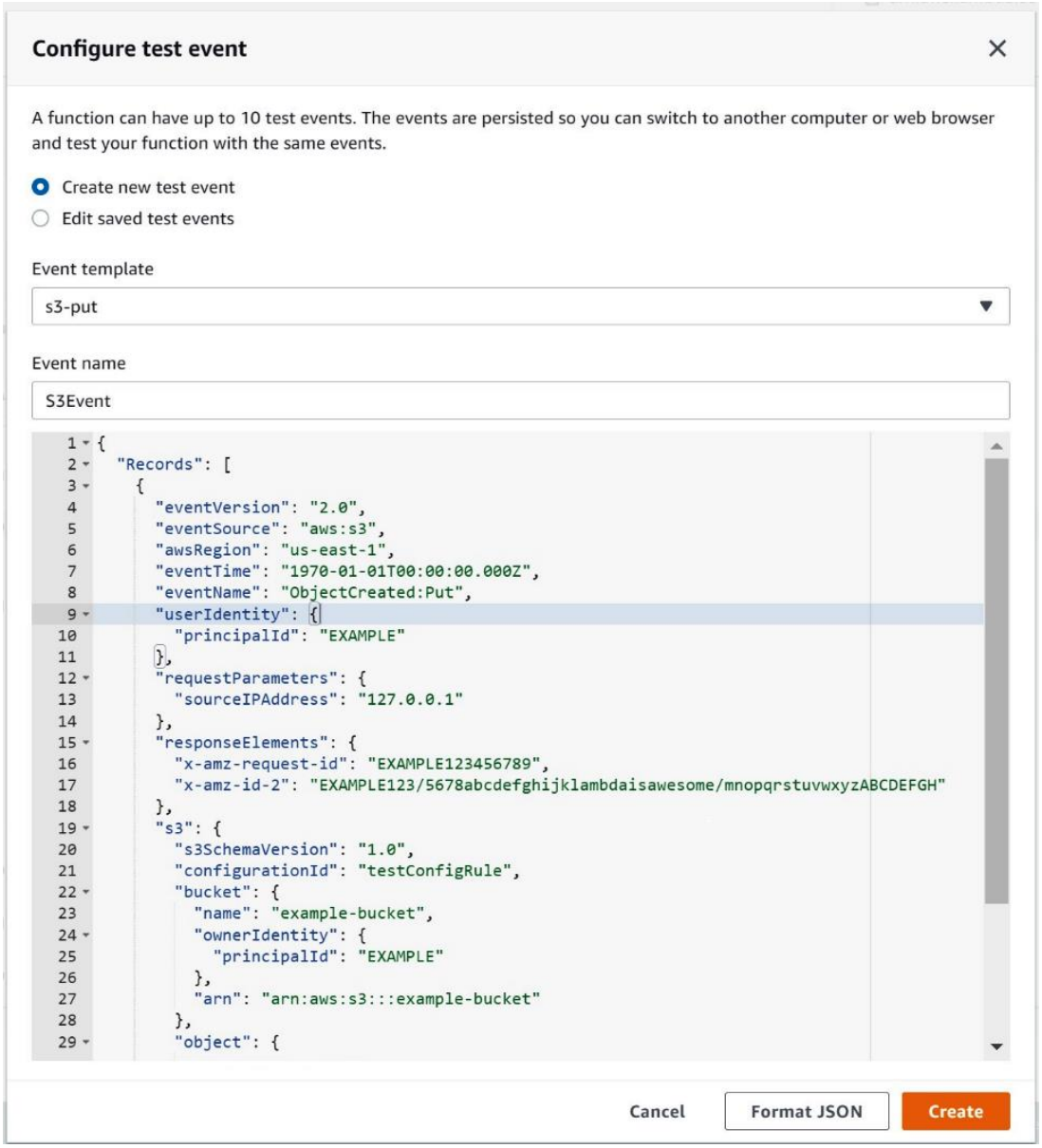



Seasonal Inventory Forecast Pipeline
Step 1
Step 2
Step 3
Step 4
Find Seasonal Items
Calculate Current Inventory Turn Rate
Gather Weather Forecast
Produce Weather Adjusted Turn Rate
Gather Purchase Order Lead Times
Gather On-Hand Inventory
Gather Historical Demand
Calculate Purchase Order Plan


Luxury Goods Report
Brand Rollup Job
Inventory Turnover Job
Brand Job
Demand Forcast Job
Sales Extract Job
Product Catalog Extract Job
Advertising Report
Sales Report
Manu Job
Supplier Report


Roles > packt-serverless-analytics-lambda
Summary
Delete role
Role ARN
arn:aws:iam:: YOUR_ACCOUNT :role/packt-serverless-analytics-lambda
Role description
Allows Lambda functions to call AWS services on your behalf. | Edit
Instance Profile ARNs
Path
/
Creation time
Last activity
Not accessed in the tracking period
Maximum session duration
1 hour Edit
Permissions
Trust relationships
Tags
Access Advisor
Revoke sessions
Permissions policies (1 policy applied)
Attach policies
Add inline policy
Policy name
Policy type
packt-serverless-analytics-chapter-9
Managed policy


Create function Info
Choose one of the following options to create your function.
Author from scratch
Start with a simple Hello World example.
Use a blueprint
Build a Lambda application from sample code and configuration presets for common use cases.
Container
Select a contai
function.
Basic information
Function name
Enter a name that describes the purpose of your function.
packt-serverless-analytics-etl
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 3.8
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
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.
packt-serverless-analytics-lambda
View the packt-serverless-analytics-lambda role on the IAM console.



Create event notification

The notification configuration identifies the events you want Amazon S3 to publish and the destinations where you want Amazon S3 to send the notifications. Learn more

General configuration

Event name

packet-serverless-analytics-chapter-9-new-data

Event name can contain up to 255 characters.

Prefix - *optional*

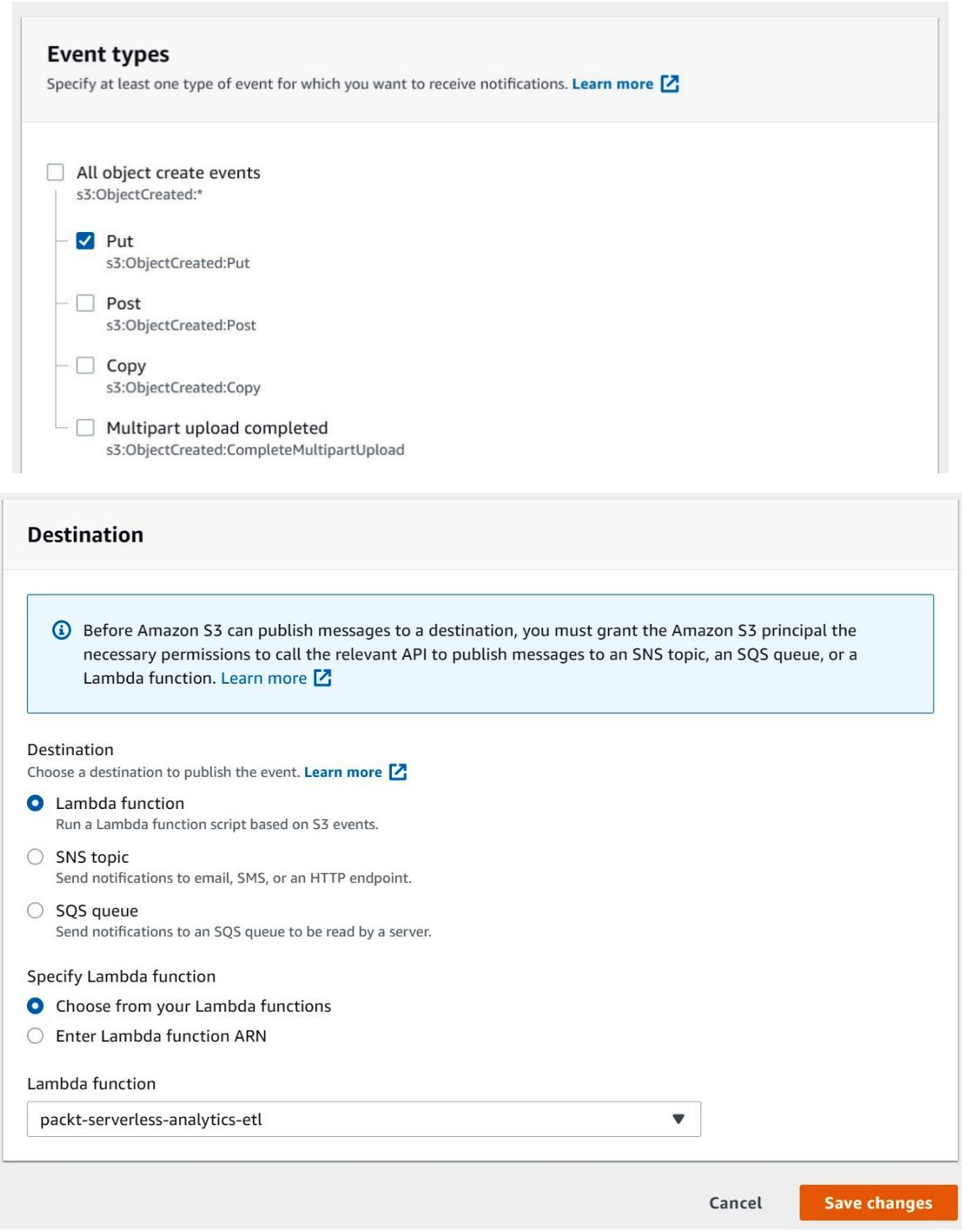
Limit the notifications to objects with key starting with specified characters.

chapter_9/import/

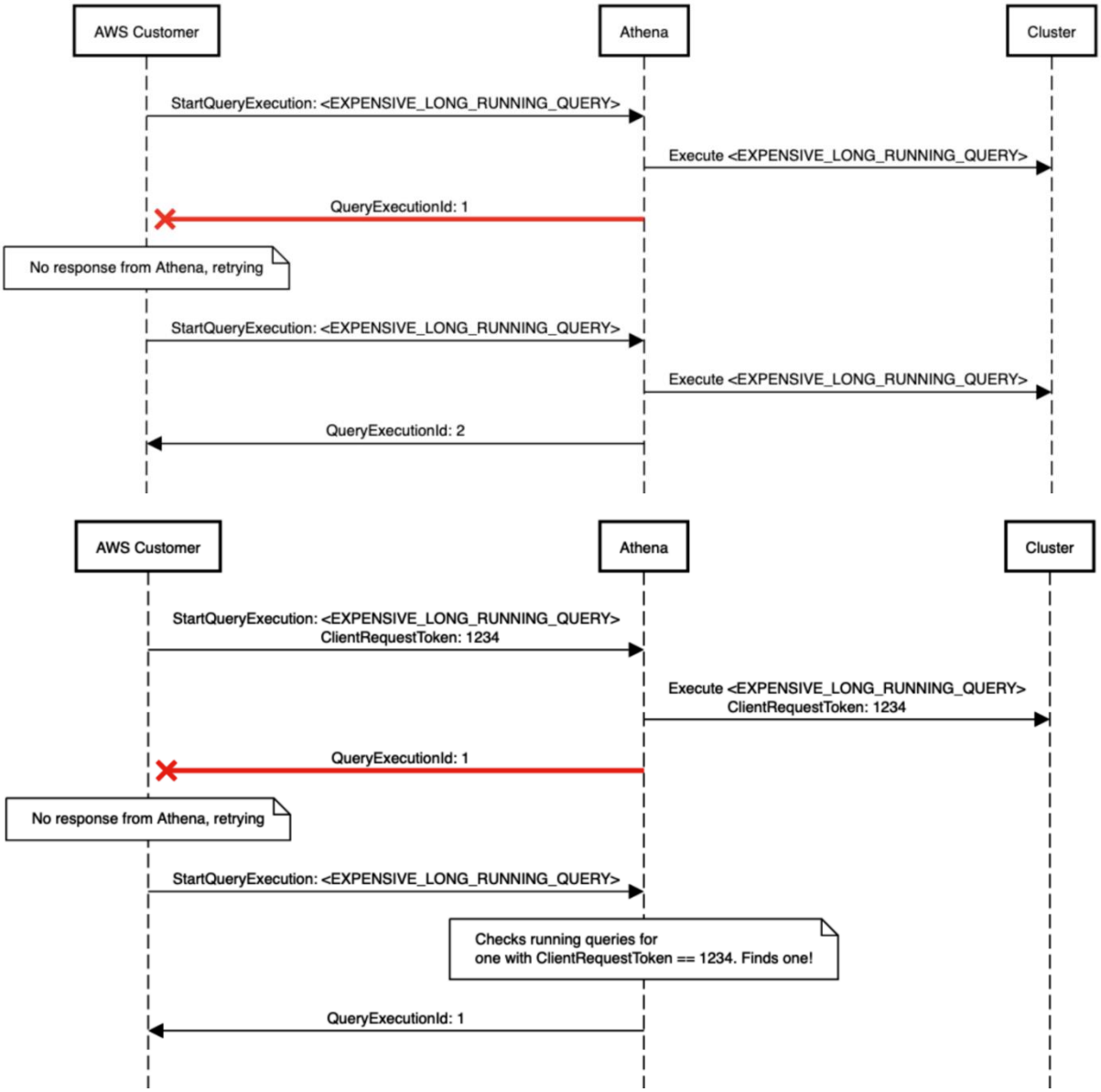
Suffix - *optional*

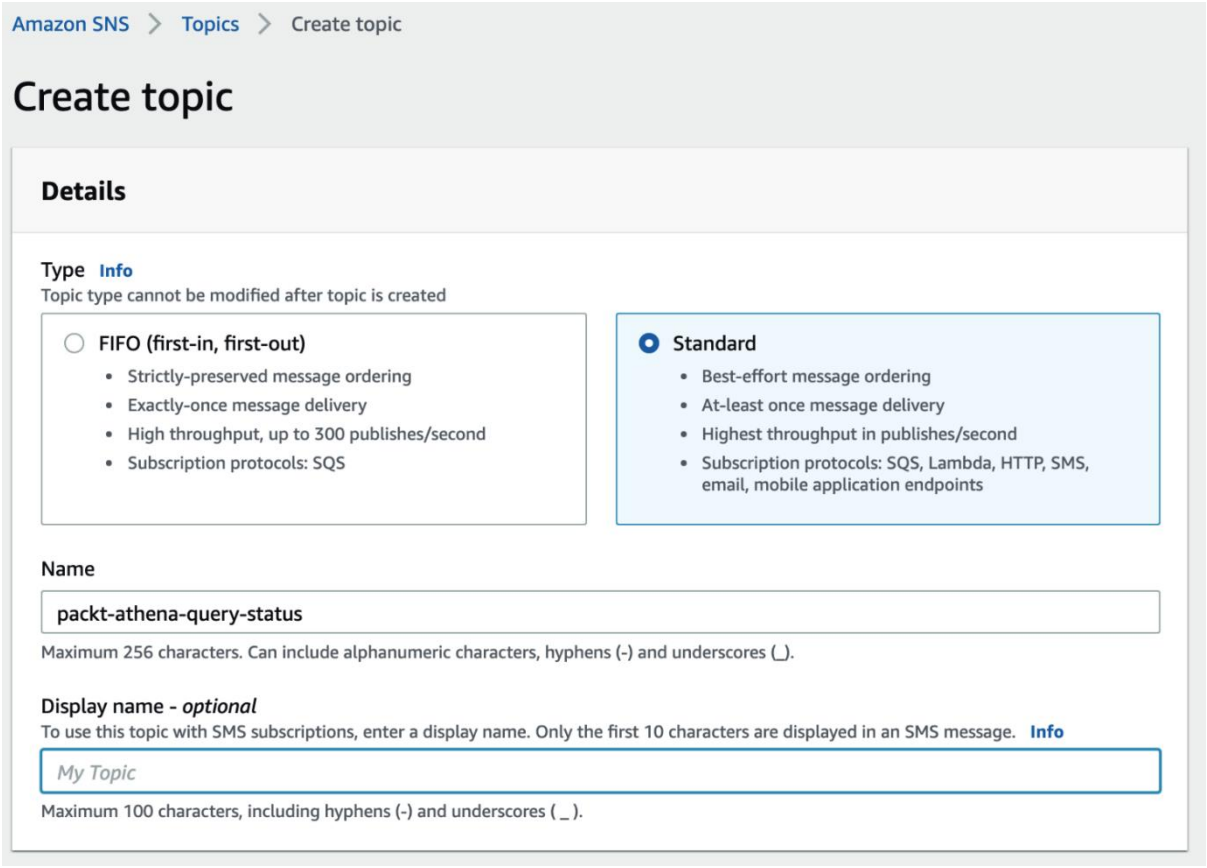
Limit the notifications to objects with key ending with specified characters.

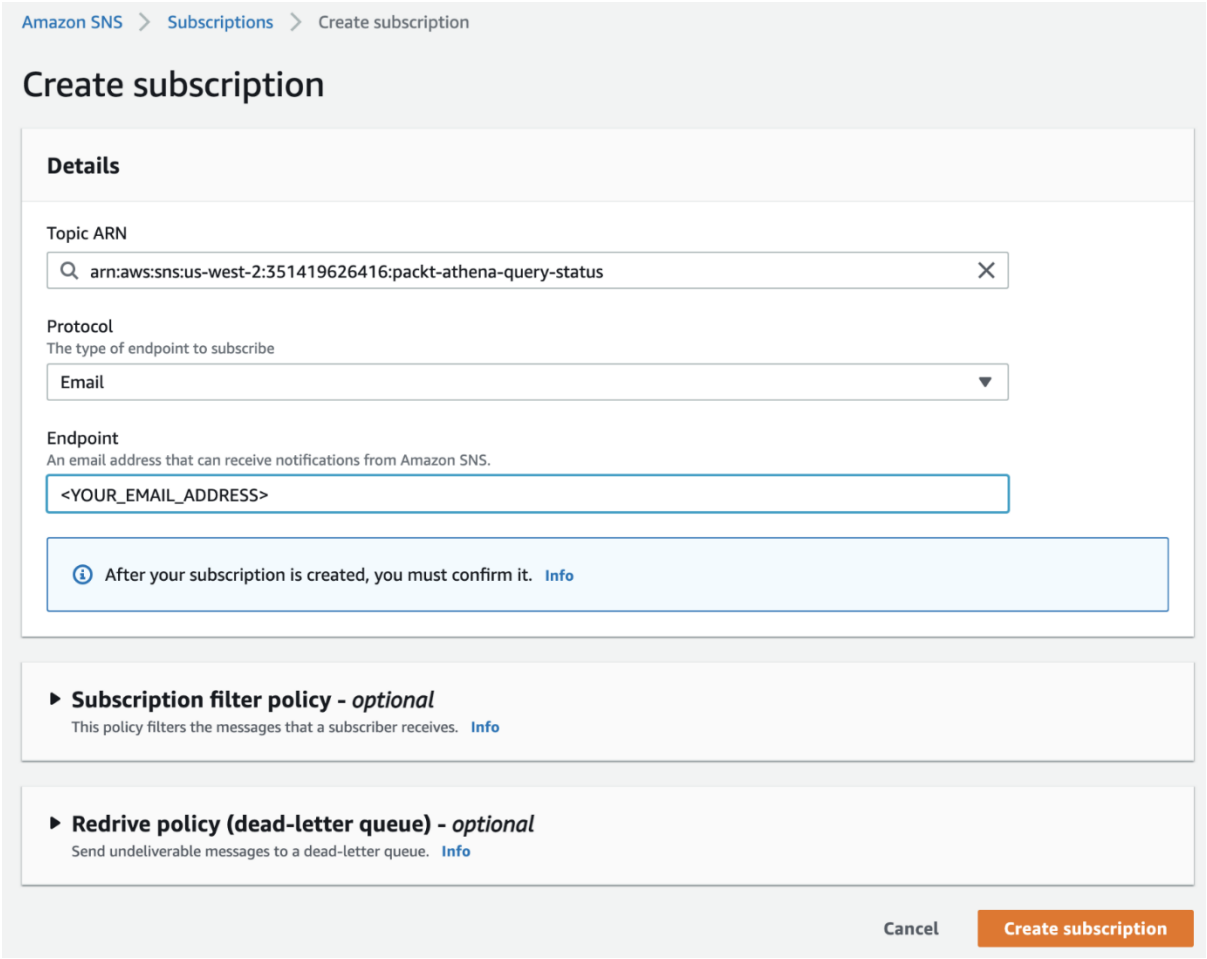
.csv


Event types
Specify at least one type of event for which you want to receive notifications. Learn more
All object create events
s3:ObjectCreated:*
Put
s3:ObjectCreated:Put
Post
s3:ObjectCreated:Post
Copy
s3:ObjectCreated:Copy
Multipart upload completed
s3:ObjectCreated:CompleteMultipartUpload
Destination
Before Amazon S3 can publish messages to a destination, you must grant the Amazon S3 principal the necessary permissions to call the relevant API to publish messages to an SNS topic, an SQS queue, or a Lambda function. Learn more
Destination
Choose a destination to publish the event. Learn more
Lambda function
Run a Lambda function script based on S3 events.
SNS topic
Send notifications to email, SMS, or an HTTP endpoint.
SQS queue
Send notifications to an SQS queue to be read by a server.
Specify Lambda function
Choose from your Lambda functions
Enter Lambda function ARN
Lambda function
packt-serverless-analytics-etl
Cancel
Save changes

Chapter 10: Building Applications with Amazon Athena




Amazon SNS > Topics > Create topic
Create topic
Details
Type Info
Topic type cannot be modified after topic is created
FIFO (first-in, first-out)
Strictly-preserved message ordering
Exactly-once message delivery
High throughput, up to 300 publishes/second
Subscription protocols: SQS
Standard
Best-effort message ordering
At-least once message delivery
Highest throughput in publishes/second
Subscription protocols: SQS, Lambda, HTTP, SMS, email, mobile application endpoints
Name
packt-athena-query-status
Maximum 256 characters. Can include alphanumeric characters, hyphens (-) and underscores (_).
Display name - optional
To use this topic with SMS subscriptions, enter a display name. Only the first 10 characters are displayed in an SMS message. Info
My Topic
Maximum 100 characters, including hyphens (-) and underscores (_).


Amazon SNS > Subscriptions > Create subscription
Create subscription
Details
Topic ARN
arn:aws:sns:us-west-2:351419626416:packt-athena-query-status
Protocol
The type of endpoint to subscribe
Email
Endpoint
An email address that can receive notifications from Amazon SNS.
<YOUR_EMAIL_ADDRESS>
After your subscription is created, you must confirm it. Info
Subscription filter policy - optional
This policy filters the messages that a subscriber receives. Info
Redrive policy (dead-letter queue) - optional
Send undeliverable messages to a dead-letter queue. Info
Cancel
Create subscription

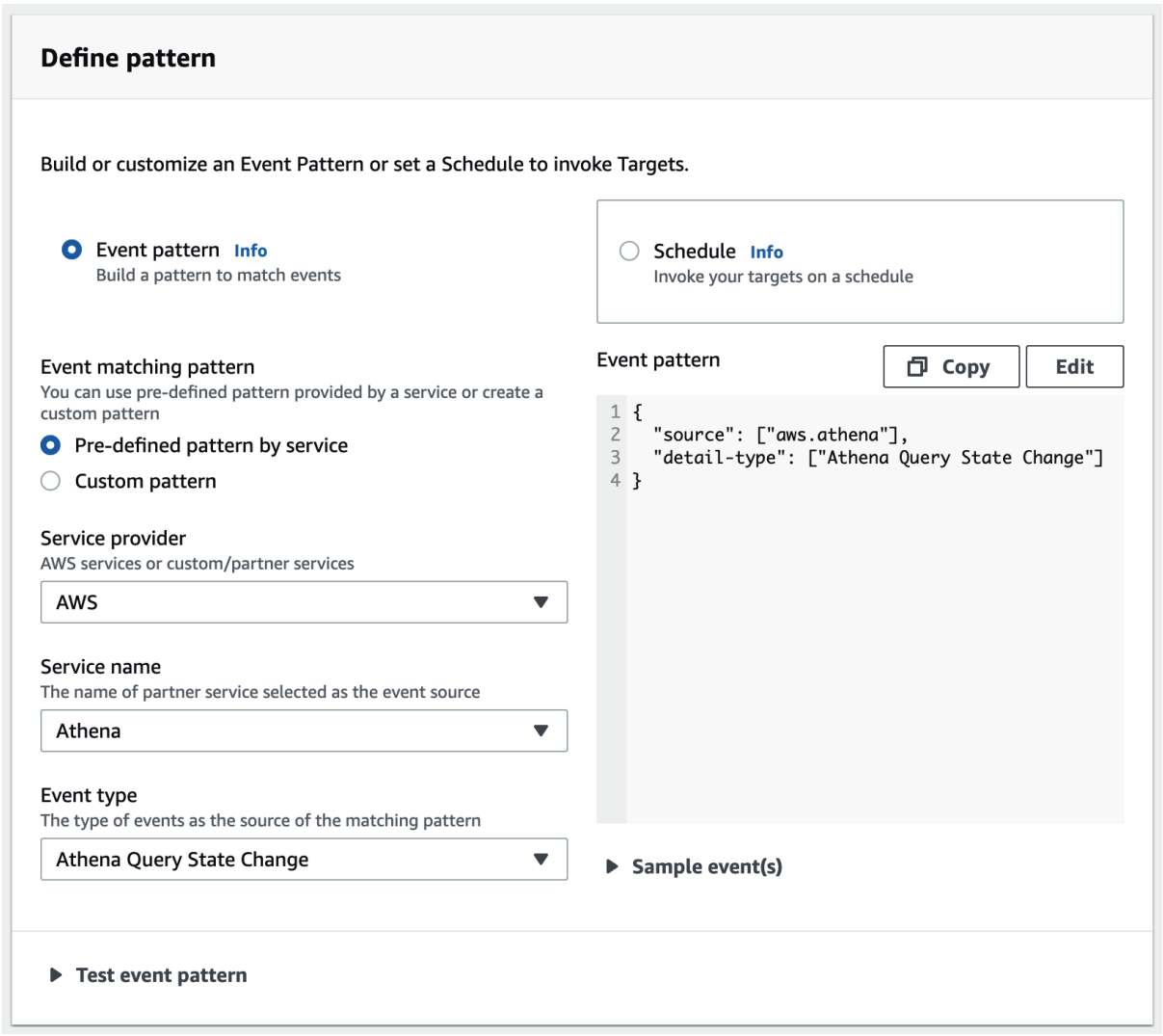
AWS Notification - Subscription Confirmation Inbox ×

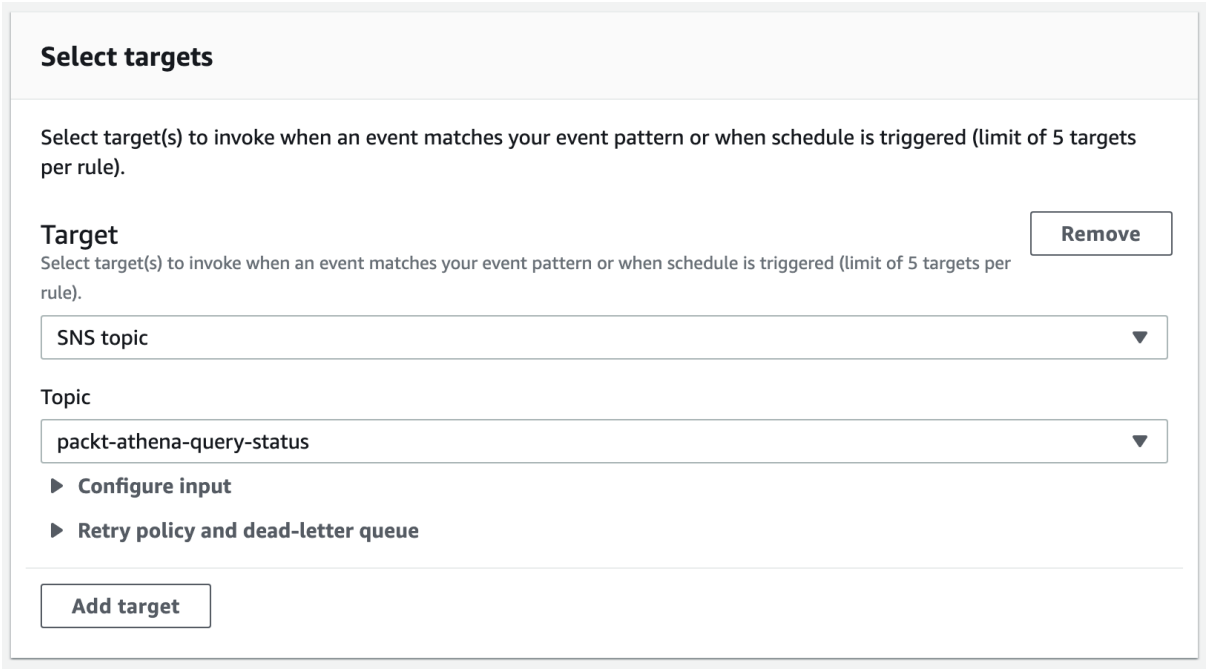
AWS Notifications <no-reply@sns.amazonaws.com>
to me

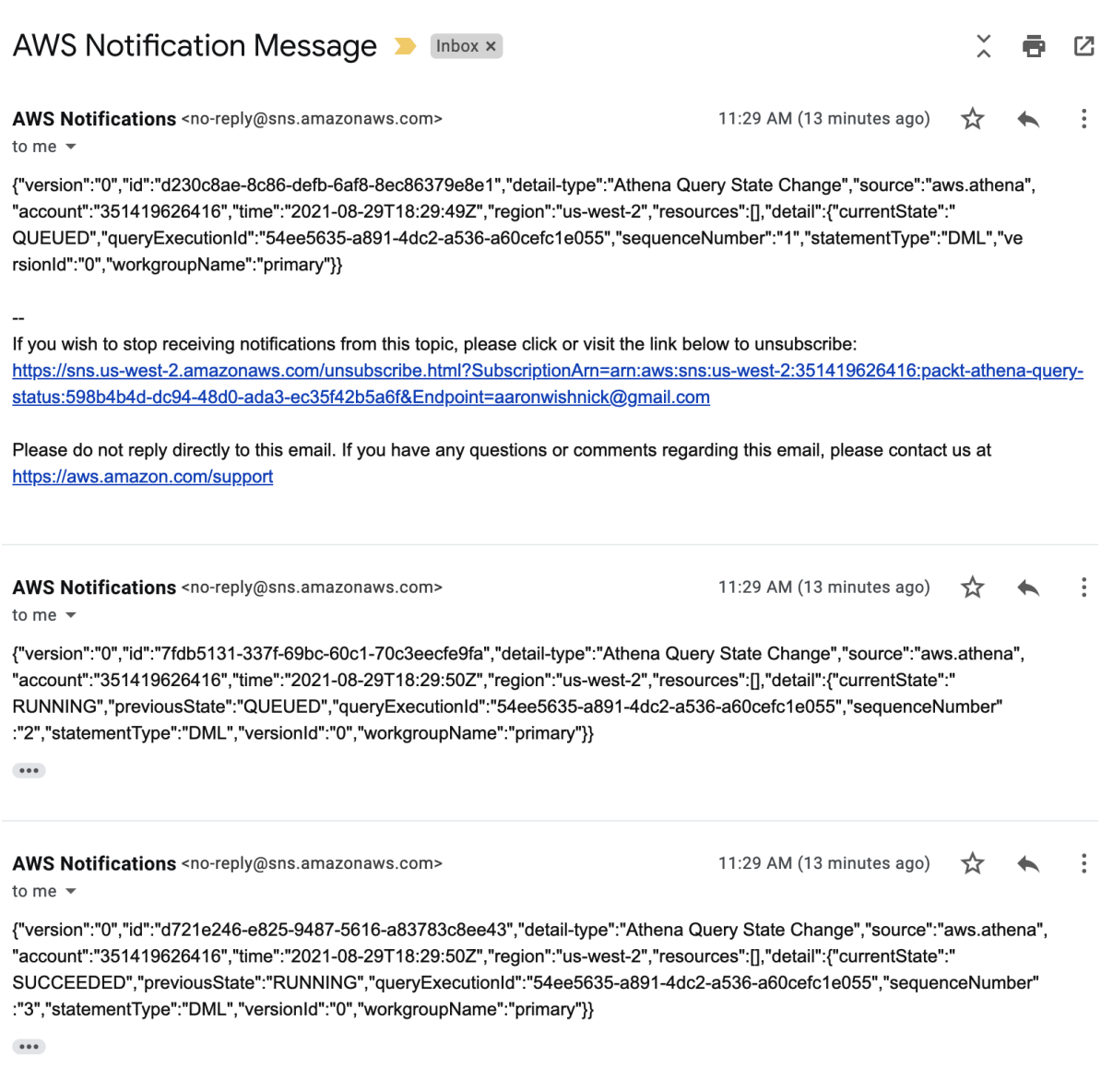
You have chosen to subscribe to the topic:
arn:aws:sns:us-west-2:351419626416:packt-athena-query-status

To confirm this subscription, click or visit the link below (If this was in error no action is necessary):
Confirm subscription

Please do not reply directly to this email. If you wish to remove yourself from receiving all future SNS subscription confirmation requests please send an email to sns-opt-out







AWS Notification Message Inbox ×

AWS Notifications <no-reply@sns.amazonaws.com> 11:29 AM (13 minutes ago)

to me

{"version":"0","id":"d230c8ae-8c86-defb-6af8-8ec86379e8e1","detail-type":"Athena Query State Change","source":"aws.athena", "account":"351419626416","time":"2021-08-29T18:29:49Z","region":"us-west-2","resources":[],"detail":{"currentState":" QUEUED","queryExecutionId":"54ee5635-a891-4dc2-a536-a60cefc1e055","sequenceNumber":"1","statementType":"DML","ve rsionId":"0","workgroupName":"primary"}}

--

If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
https://sns.us-west-2.amazonaws.com/unsubscribe.html?SubscriptionArn=arn:aws:sns:us-west-2:351419626416:packt-athena-query-status:598b4b4d-dc94-48d0-ada3-ec35f42b5a6f&Endpoint=aaronwishnick@gmail.com

Please do not reply directly to this email. If you have any questions or comments regarding this email, please contact us at
https://aws.amazon.com/support

AWS Notifications <no-reply@sns.amazonaws.com> 11:29 AM (13 minutes ago)

to me

{"version":"0","id":"7fdb5131-337f-69bc-60c1-70c3eecfe9fa","detail-type":"Athena Query State Change","source":"aws.athena", "account":"351419626416","time":"2021-08-29T18:29:50Z","region":"us-west-2","resources":[],"detail":{"currentState":" RUNNING","previousState":"QUEUED","queryExecutionId":"54ee5635-a891-4dc2-a536-a60cefc1e055","sequenceNumber" :"2","statementType":"DML","versionId":"0","workgroupName":"primary"}}

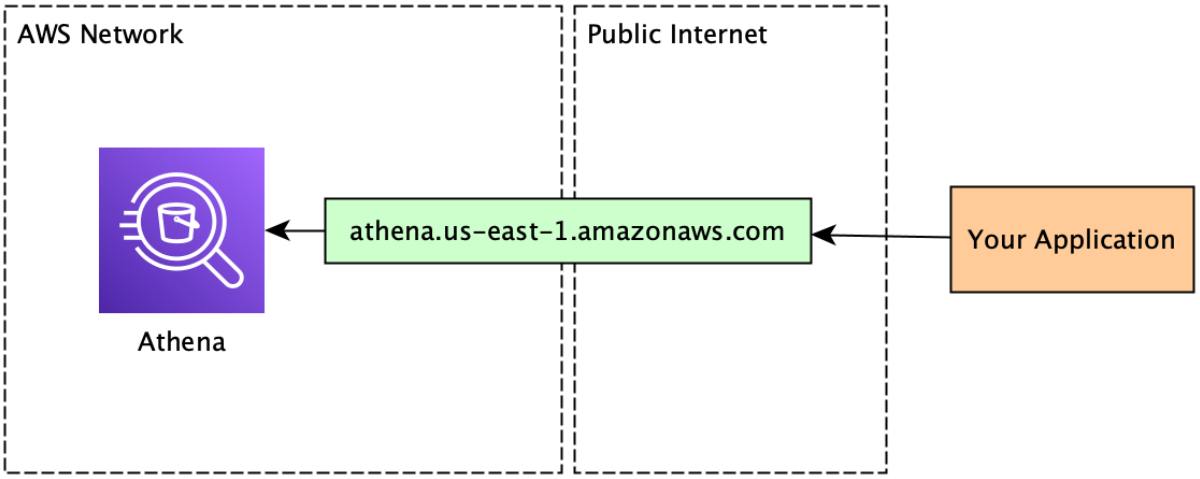
•••

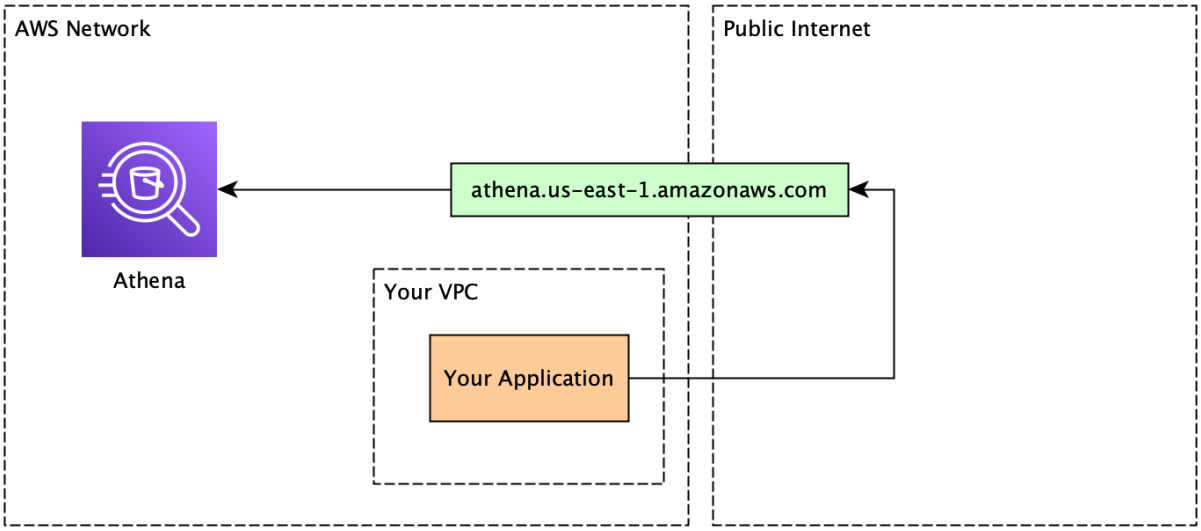
AWS Notifications <no-reply@sns.amazonaws.com> 11:29 AM (13 minutes ago)

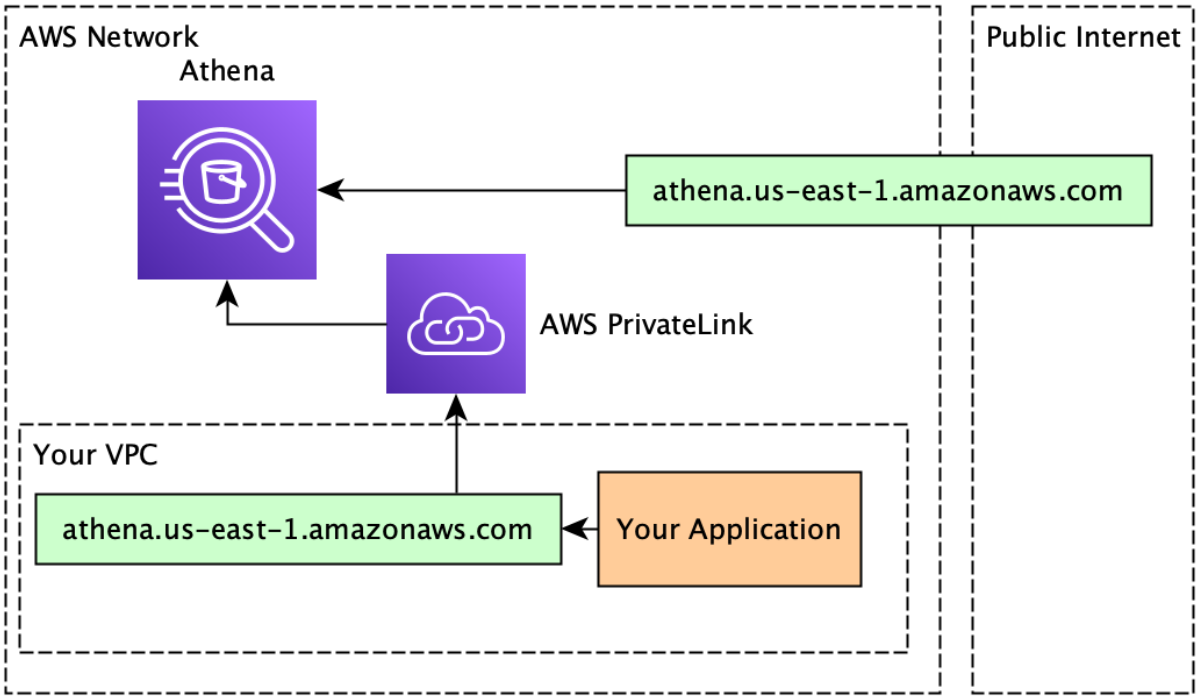
to me

{"version":"0","id":"d721e246-e825-9487-5616-a83783c8ee43","detail-type":"Athena Query State Change","source":"aws.athena", "account":"351419626416","time":"2021-08-29T18:29:50Z","region":"us-west-2","resources":[],"detail":{"currentState":" SUCCEEDED","previousState":"RUNNING","queryExecutionId":"54ee5635-a891-4dc2-a536-a60cefc1e055","sequenceNumber" :"3","statementType":"DML","versionId":"0","workgroupName":"primary"}}

•••


AWS Network
Public Internet
athena.us-east-1.amazonaws.com
Your Application
Athena


AWS Network
Public Internet
athena.us-east-1.amazonaws.com
Athena
Your VPC
Your Application


AWS Network
Athena
Public Internet
athena.us-east-1.amazonaws.com
AWS PrivateLink
Your VPC
athena.us-east-1.amazonaws.com
Your Application

Metrics

Metrics ☑ Publish query metrics to AWS CloudWatch

Choose trail attributes

General details

A trail created in the console is a multi-region trail. Learn more

Trail name

Enter a display name for your trail.

packt-athena-cloudtrail

3-128 characters. Only letters, numbers, periods, underscores, and dashes are allowed.

☐ Enable for all accounts in my organization

To review accounts in your organization, open AWS Organizations. See all accounts

Storage location Info

◉ Create new S3 bucket
Create a bucket to store logs for the trail.

○ Use existing S3 bucket
Choose an existing bucket to store logs for this trail.

Trail log bucket and folder

Enter a new S3 bucket name and folder (prefix) to store your logs. Bucket names must be globally unique.

aws-cloudtrail-logs-351419626416-e16e65e0

Logs will be stored in aws-cloudtrail-logs-351419626416-e16e65e0/AWSLogs/351419626416

Log file SSE-KMS encryption Info

☑ Enabled

Customer managed AWS KMS key

◉ New

○ Existing

AWS KMS alias

packt-athena-cloudtrail-cmk

KMS key and S3 bucket must be in the same region.

▼ Additional settings

Log file validation Info

☑ Enabled

SNS notification delivery Info

☐ Enabled

Choose log events

Events Info

Record API activity for individual resources, or for all current and future resources in AWS account. Additional charges apply

Event type

Choose the type of events that you want to log.

- [x] Management events
  Capture management operations performed on your AWS resources.
- [ ] Data events
  Log the resource operations performed on or within a resource.
- [ ] Insights events
  Identify unusual activity, errors, or user behavior in your account.

Management events Info

Management events show information about management operations performed on resources in your AWS account.

No additional charges apply to log management events on this trail because this is your first copy of management events.

API activity

Choose the activities you want to log.

- [x] Read
- [x] Write
- [ ] Exclude AWS KMS events
- [ ] Exclude Amazon RDS Data API events

Create a table in Amazon Athena

You can use Amazon Athena to analyze events that are stored in a trail's Amazon S3 bucket. Athena is an interactive query service that helps you analyze data in S3 buckets by using standard SQL. Athena charges for running queries. Learn more

Storage location

aws-cloudtrail-logs-351419626416-3d0c07d3

Choose an S3 bucket that contains CloudTrail log files

Athena table name

cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3

This name is auto-generated. You can rename it in Amazon Athena.

Copy

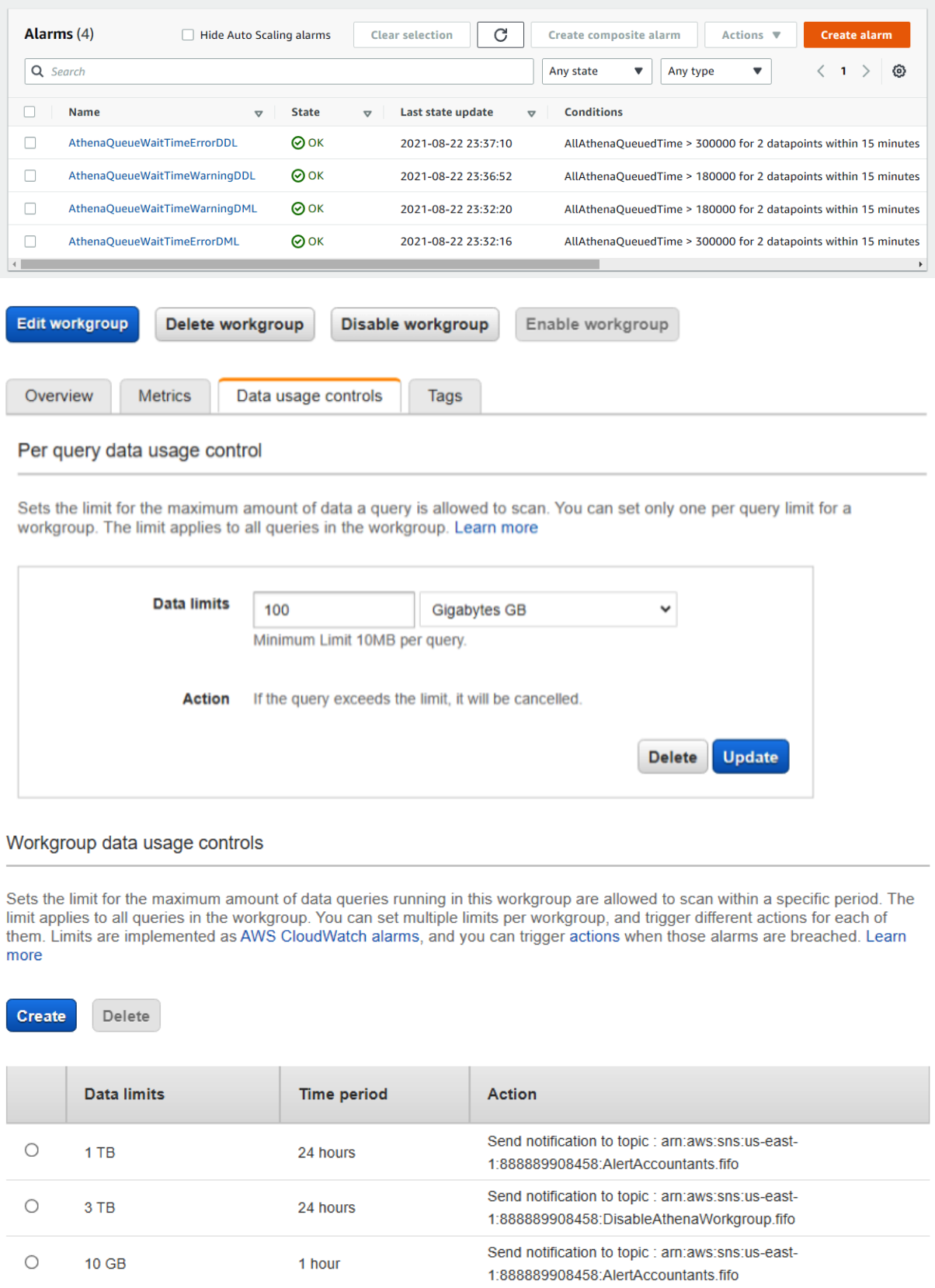
```
1 CREATE EXTERNAL TABLE
  cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3 (
2     eventVersion STRING,
```

Results

	queryString
1	"SELECT * FROM \"packt_serverless_analytics\".\"chapter_7_counties\" limit 10;"
2	"SELECT * FROM \"packt_serverless_analytics\".\"taxi_ridership_data\" limit 10;"
3	"CREATE EXTERNAL TABLE cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3 (\n event
4	"SELECT * FROM \"default\".\"cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3\" limit 10;
5	"SELECT * FROM \"default\".\"cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3\" limit 10;
6	"SELECT * FROM \"default\".\"cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3\" WHER
7	"SELECT * FROM \"default\".\"cloudtrail_logs_aws_cloudtrail_logs_351419626416_3d0c07d3\";"

Title	Publisher	Year of Publication
Serverless Analytics with Amazon Athena	Packt	2021

Chapter 11: Operational Excellence – Monitoring, Optimization, and Troubleshooting



Create workgroup data usage control

Sets the limit for the maximum amount of data queries running in this workgroup are allowed to scan within a specific period. The limit applies to all queries in the workgroup. You can set multiple limits per workgroup, and trigger different actions for each of them. Limits are implemented as AWS CloudWatch alarms, and you can trigger actions when those alarms are breached. Learn more

Data limits | Terabytes

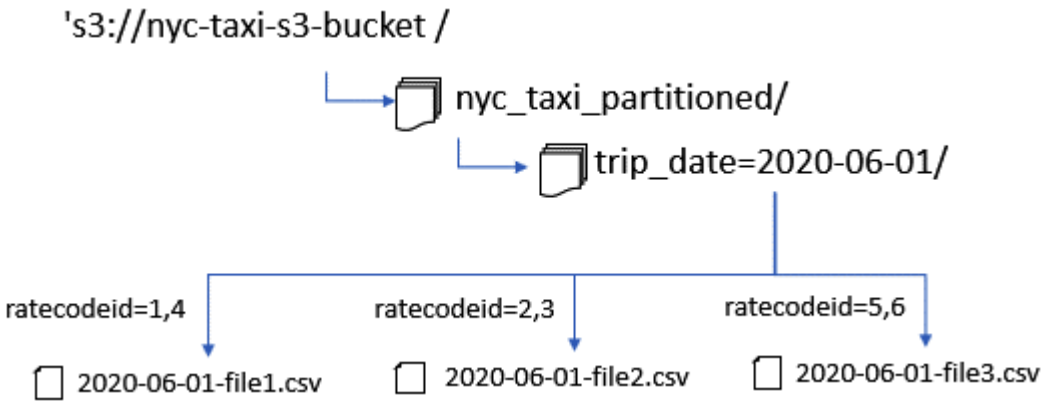
Time period | 1 day

Action | Send a notification to

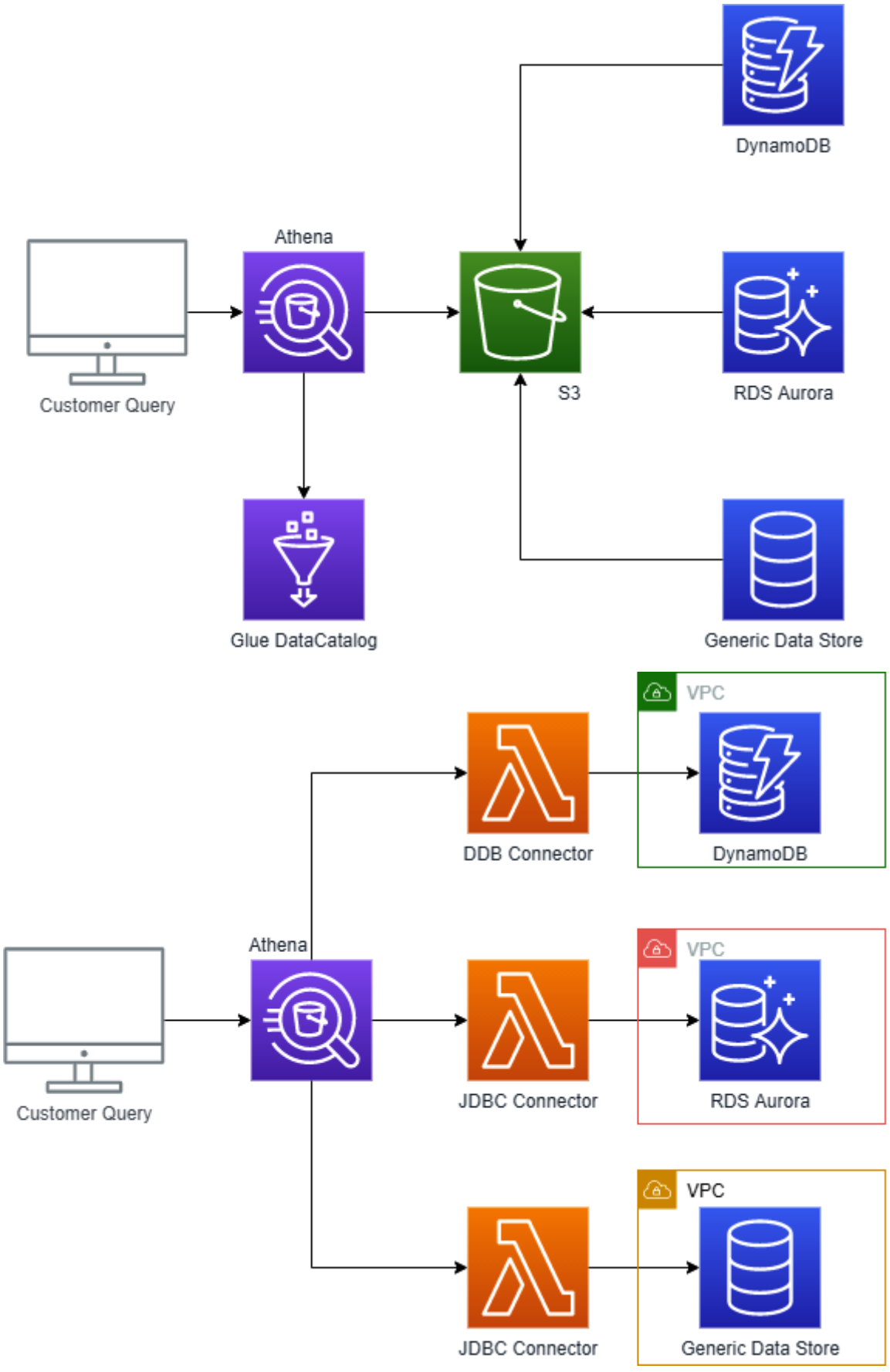
Choose an SNS topic | Create SNS topic

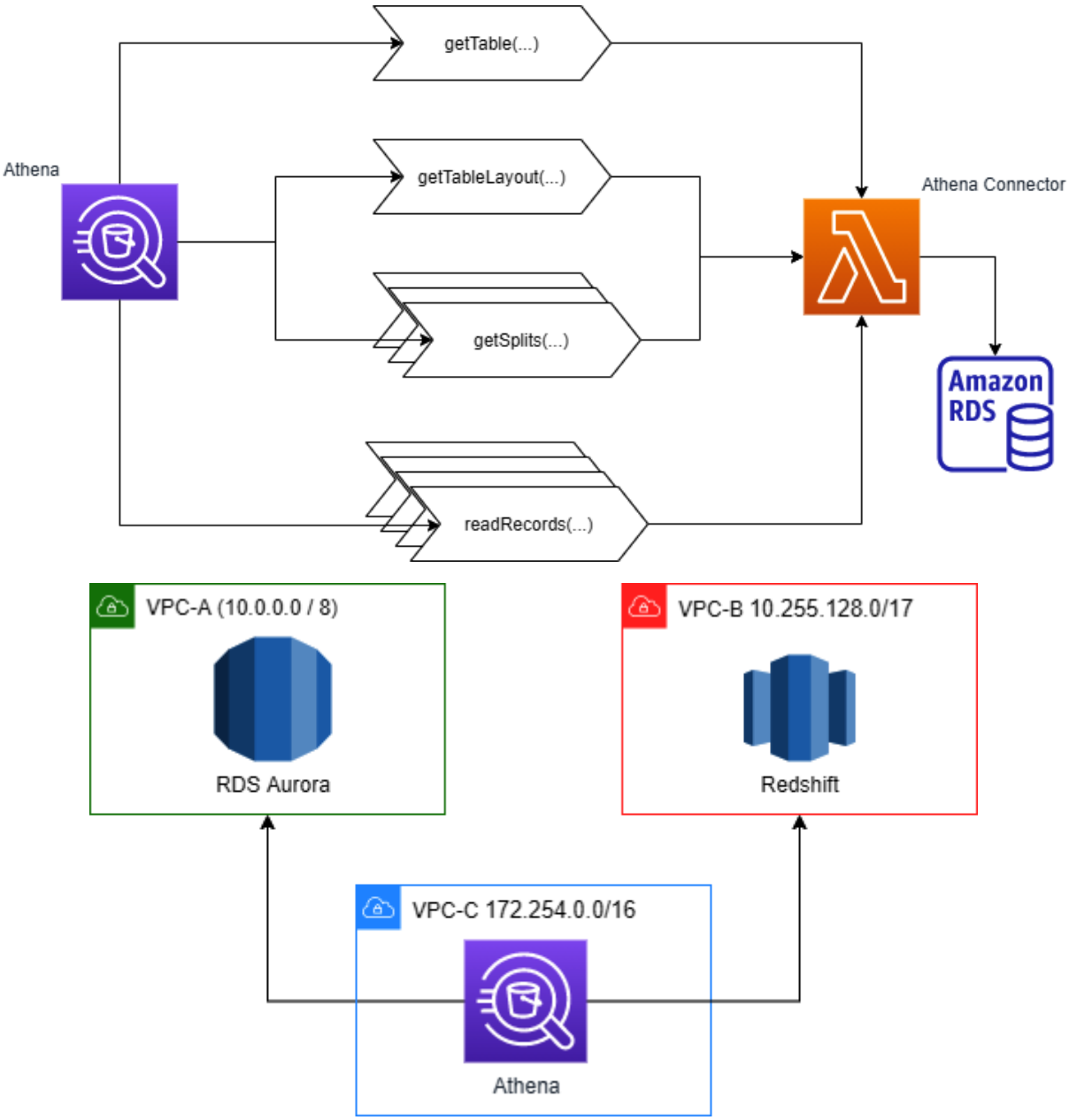
Enter an SNS topic ARN

Cancel | Create



Chapter 12: Athena Query Federation




Athena
getTable(...)
getTableLayout(...)
getSplits(...)
readRecords(...)
Athena Connector
Amazon RDS
VPC-A (10.0.0.0 / 8)
RDS Aurora
VPC-B 10.255.128.0/17
Redshift
VPC-C 172.254.0.0/16
Athena

AthenaAwsCmdbConnector — version 2021.21.1

Review, configure and deploy

Copy as SAM Resource

Application details

Author	Source code URL	Description	Report a vulnerability
Amazon Athena Federation AWS verified author	https://github.com/awslabs/aws-athena-query-federation	This connector enables Amazon Athena to communicate with various AWS Services, making your resource inventories accessible via SQL.	If you believe this application poses a security risk, please file a vulnerability report.

▶ **Template**

▶ **Permissions**

▶ **License**

Readme file

Amazon Athena AWS CMDB Connector

This connector enables Amazon Athena to communicate with various AWS Services, making your AWS Resource inventory accessible via SQL.

Athena Federated Queries are now enabled as GA in us-east-1, us-east-2, us-west-2, eu-west-1, ap-northeast-1, ap-south-1, us-west-1, ap-southeast-1, ap-southeast-2, eu-west-2, ap-northeast-2, eu-west-3, ca-central-1, sa-east-1, and eu-central-1. To use this feature, upgrade your engine version to Athena V2 in your workgroup settings. Check documentation here for more details: https://docs.aws.amazon.com/athena/latest/ug/engine-versions.html.

Usage

Parameters

The Athena AWS CMDB Connector provides several configuration options via Lambda environment variables. More detail on the available parameters can be found below.

1. **spill_bucket** - When the data returned by your Lambda function exceeds Lambda's limits, this is the bucket that the data will be

Application settings

Application name
The stack name of this application created via AWS CloudFormation
packt-serverless-analytics-AthenaAwsCmdbConnector

SpillBucket
The name of the bucket where this function can spill data.
YOUR_S3_BUCKET_HERE

▼ **ConnectorConfig**

AthenaCatalogName
The name you will give to this catalog in Athena. It will also be used as the function name. This name must satisfy the pattern ^[a-z0-9-_]{1,64}$
packt_serverless_analytics_cmdb

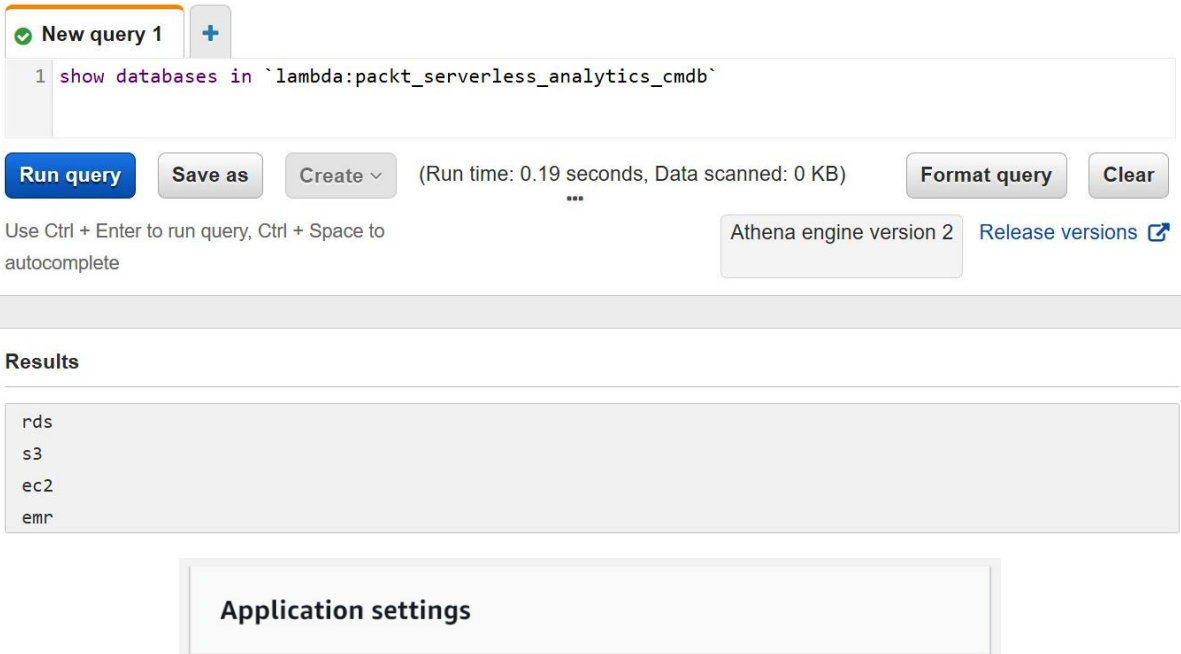
DisableSpillEncryption
WARNING: If set to 'true' encryption for spilled data is disabled.
false

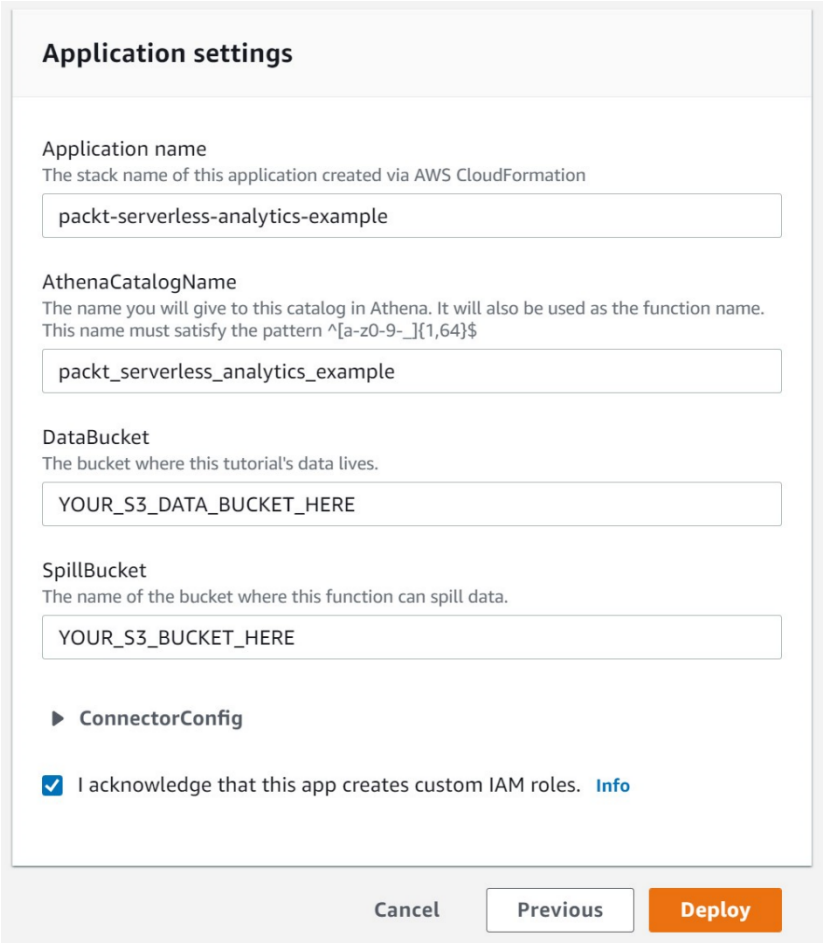
LambdaMemory
Lambda memory in MB (min 128 - 3008 max).
3008

LambdaTimeout
Maximum Lambda invocation runtime in seconds. (min 1 - 900 max)
900

SpillPrefix
The prefix within SpillBucket where this function can spill data.
athena-spill

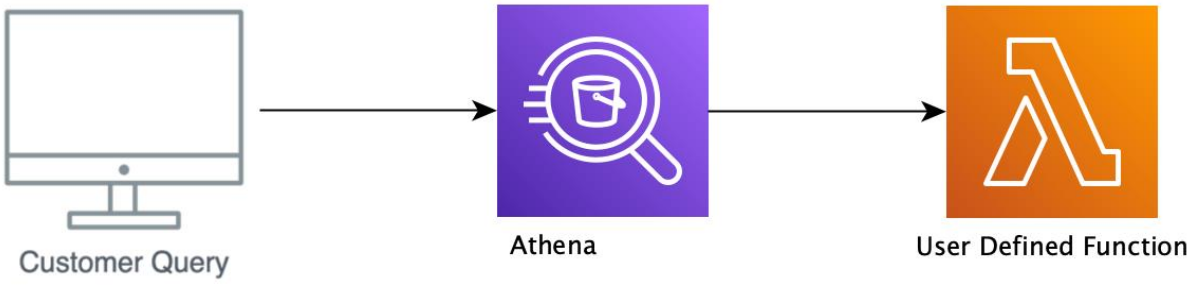
☑ I acknowledge that this app creates custom IAM roles. Info

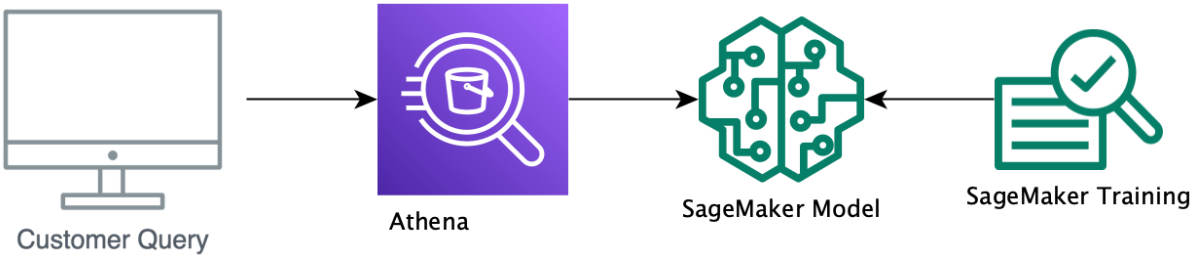

New query 1
1 show databases in `lambda:packt_serverless_analytics_cmdb`
Run query
Save as
Create
(Run time: 0.19 seconds, Data scanned: 0 KB)
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Athena engine version 2
Release versions
Results
rds
s3
ec2
emr


Application settings
Application name
The stack name of this application created via AWS CloudFormation
packt-serverless-analytics-example
AthenaCatalogName
The name you will give to this catalog in Athena. It will also be used as the function name. This name must satisfy the pattern ^[a-z0-9-_]{1,64}$
packt_serverless_analytics_example
DataBucket
The bucket where this tutorial's data lives.
YOUR_S3_DATA_BUCKET_HERE
SpillBucket
The name of the bucket where this function can spill data.
YOUR_S3_BUCKET_HERE
ConnectorConfig
I acknowledge that this app creates custom IAM roles. Info
Cancel
Previous
Deploy

Column Name	Data Type	Description
bucket_name	varchar	The name of the bucket that this object is in. You must specify a specific bucket in the where clause to query this table as that is how the connector will configure its S3 client. It will not list all the objects in all buckets.
key	varchar	The key of the S3 object.
e_tag	varchar	The ETag of the S3 object.
bytes	bigint	The size of the S3 object in bytes.
storage_class	varchar	The S3 storage class of the S3 object.
last_modified	timestamp	The last time the S3 object was modified.
owner_name	varchar	The owner name of the S3 object.
owner_id	varchar	The ID of the S3 object's owner.

Chapter 13: Athena UDFs and ML





Create role

1 2 3 4

Select type of trusted entity

AWS service
EC2, Lambda and others

Another AWS account
Belonging to you or 3rd party

Web identity
Cognito or any OpenID provider

SAML 2.0 federation
Your corporate directory

Allows AWS services to perform actions on your behalf. Learn more

Choose a use case

Common use cases

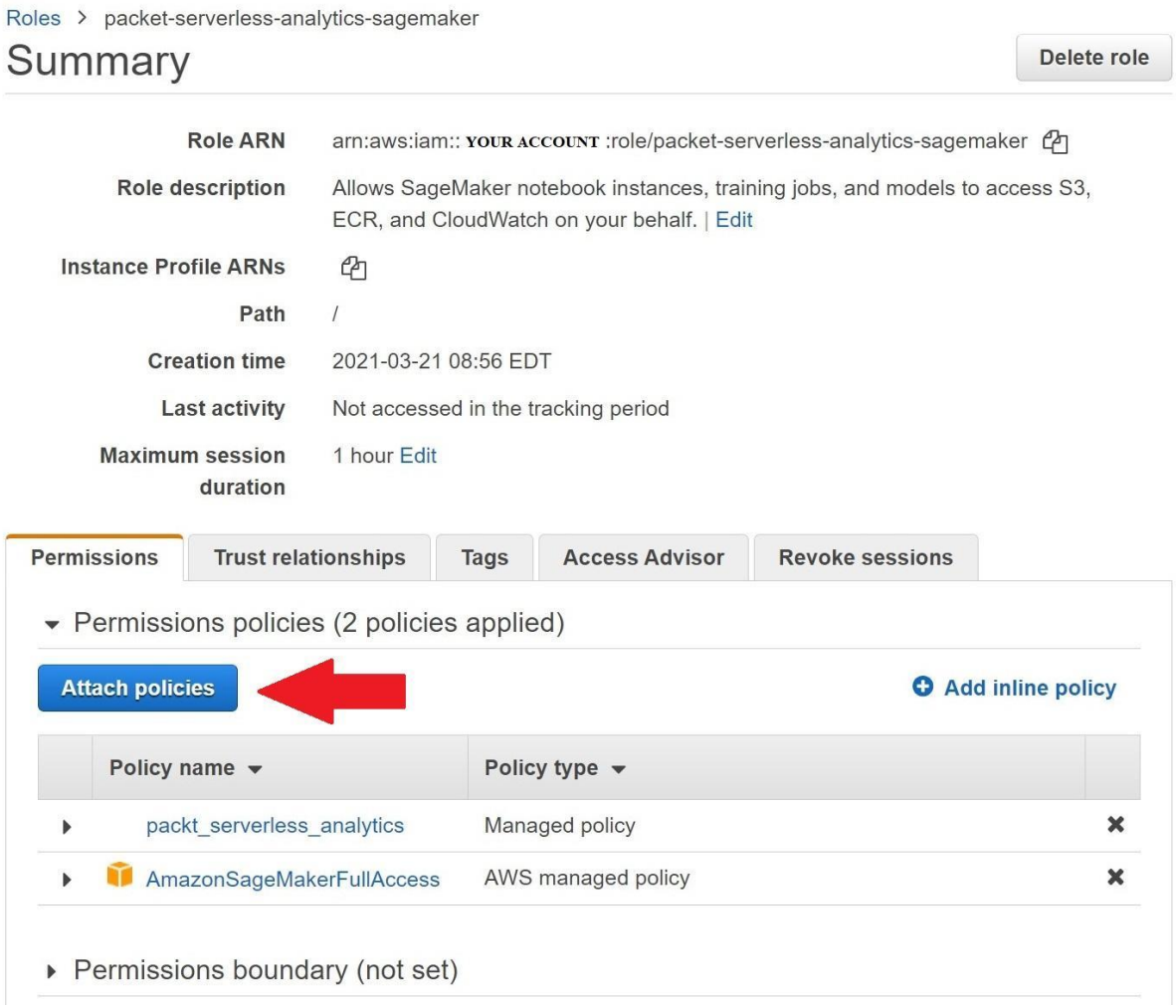
EC2
Allows EC2 instances to call AWS services on your behalf.

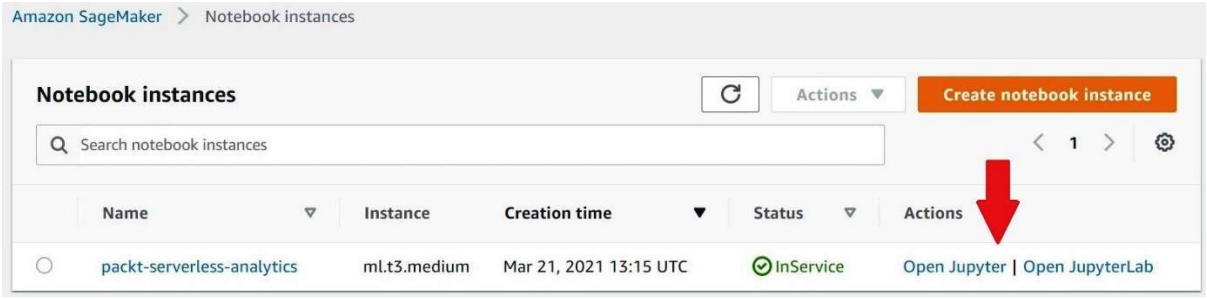
Lambda
Allows Lambda functions to call AWS services on your behalf.

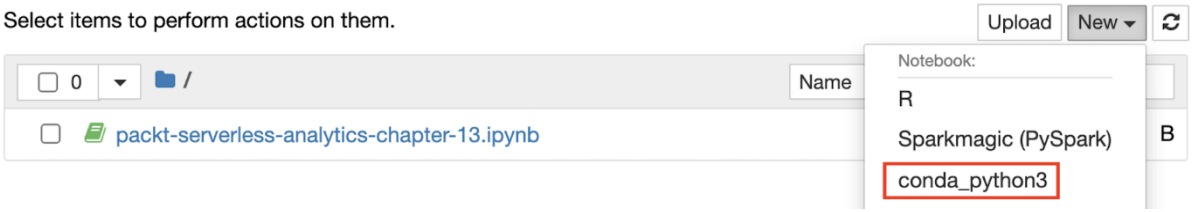
Or select a service to view its use cases

API Gateway	CloudWatch Events	EKS	IoT Things Graph	Redshift
AWS Backup	CodeBuild	EMR	KMS	Rekognition
AWS Chatbot	CodeDeploy	ElastiCache	Kinesis	RoboMaker
AWS Marketplace	CodeGuru	Elastic Beanstalk	Lake Formation	S3
AWS Support	CodeStar Notifications	Elastic Container Registry	Lambda	SMS
Amplify	Comprehend	Elastic Container Service	Lex	SNS
AppStream 2.0	Config	Elastic Transcoder	License Manager	SWF
AppSync	Connect	ElasticLoadBalancing	MQ	**SageMaker**

Cancel Next: Permissions





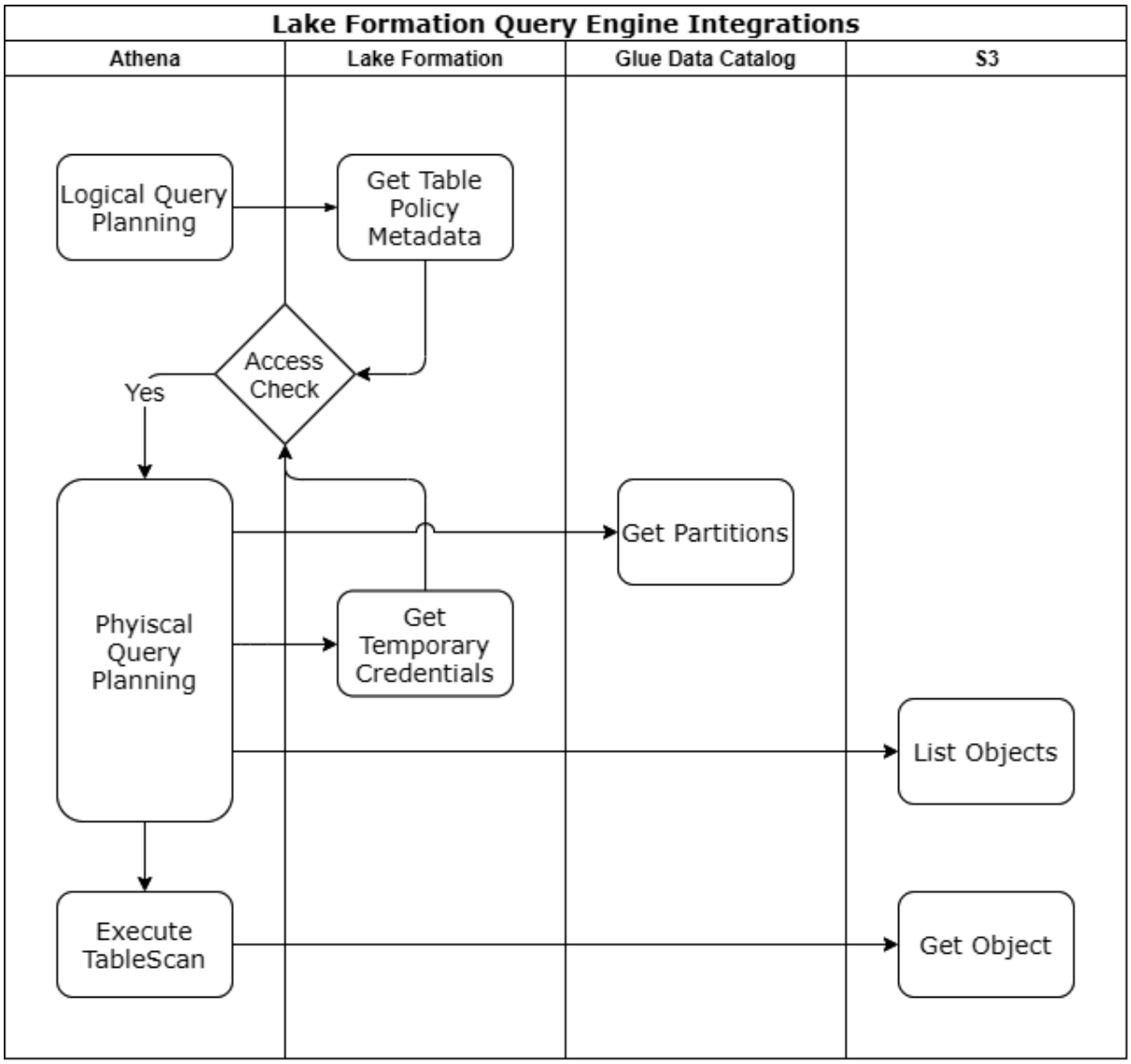


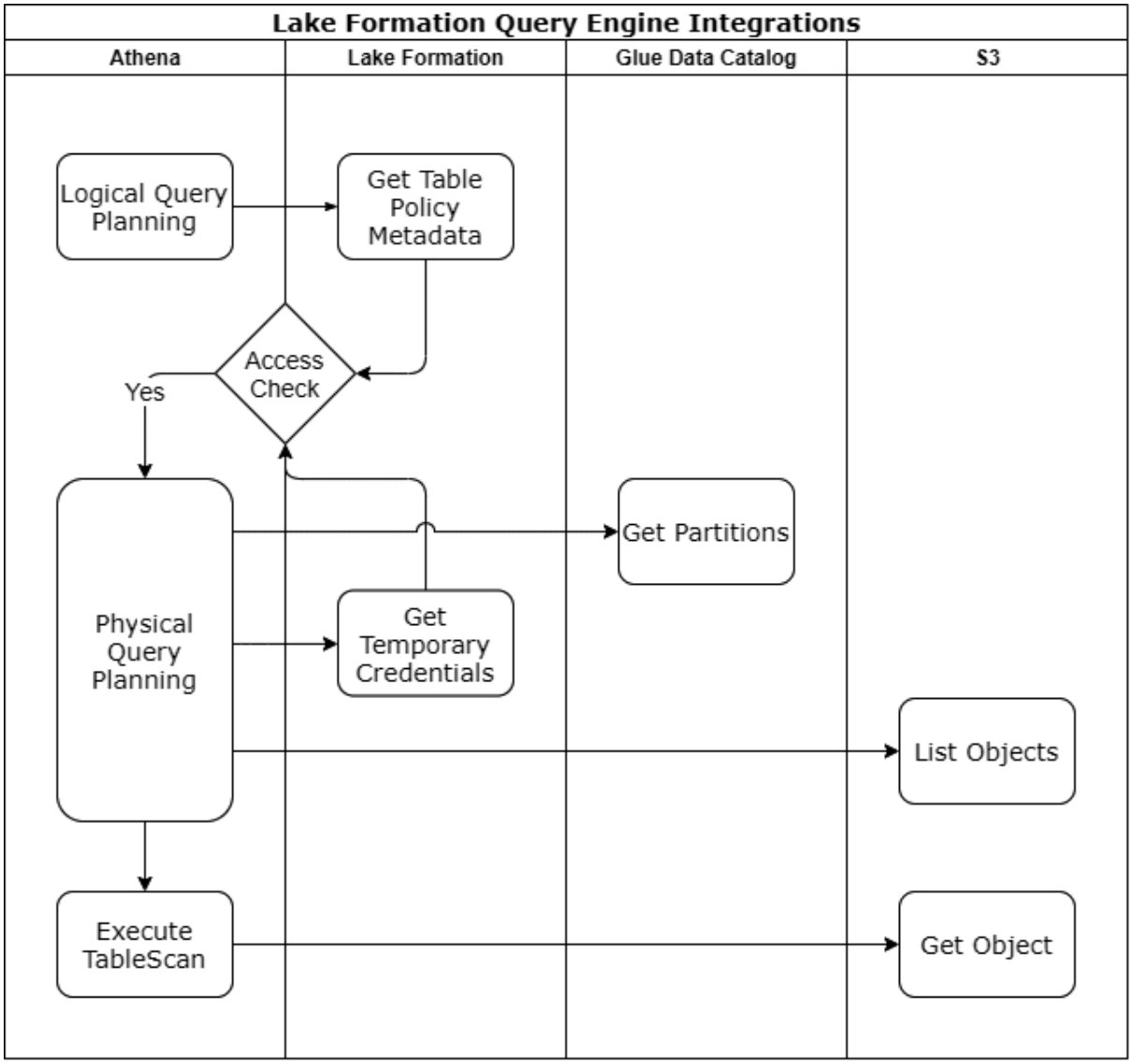

```
Training job name: randomcutforest-2021-08-22-03-07-26-016
-------------!
Endpoint name (used by Athena): randomcutforest-2021-08-22-03-10-43-029
```

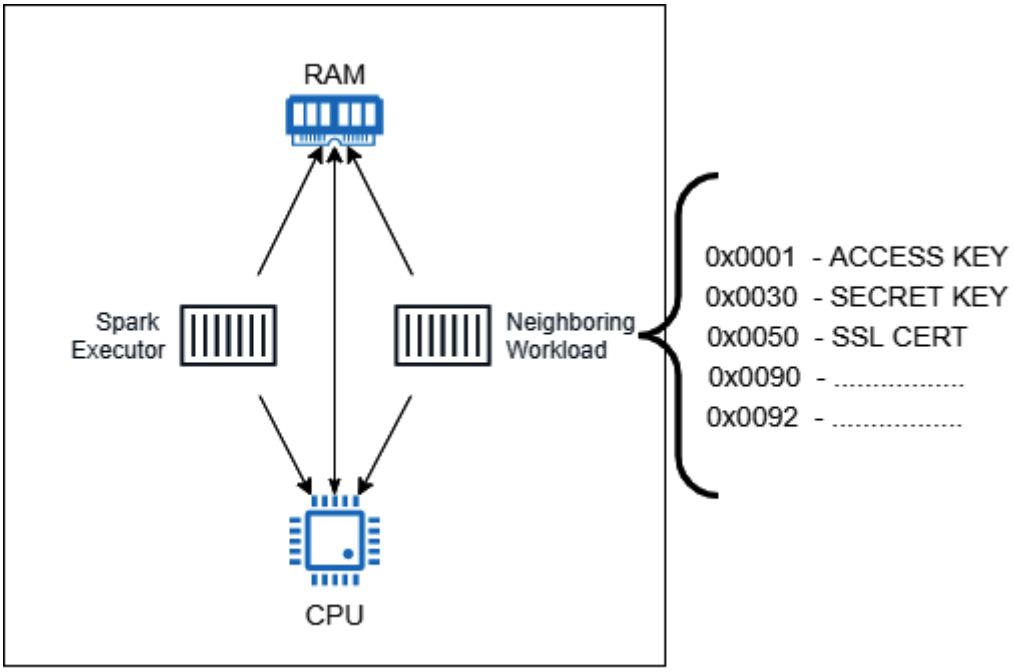
Year	Month	Day	Encrypted Payload	Decrypted Payload
2017	11	1	0UTIXoWnKqtQe8y+BSHNmdEXmWfQalRQH60pobsgwws=	SecretText-1755604178
2017	11	1	i9AoMmLI6JidPjw/SFXduBB6HUmE8aXQLMhekhIfE1U=	SecretText-747575690
2017	11	1	HWsLCXAnGFXnnjD8Nc1RbO0+5JzrhnCB/feJ/EzSxto=	SecretText-1720603622
2017	11	1	lqL0mxeOeEesRY7EU95Fi6QEW92nj2mh8xyex69j+8A=	SecretText-1167647466
2017	11	1	C57VAyZ6Y0C+xKA2Lv6fOcIP0x6Px8BlEVBGSc74C4I=	SecretText-1854103174

Time	Number of Rides	Score
11/2/14 1:00	39,197	4.59422851305562
11/2/14 1:30	35,212	4.11130198098128
9/6/14 23:00	30,373	3.1411160633939024
9/6/14 22:30	30,313	3.1235797654023556
1/1/15 1:00	30,236	3.1022678031766535

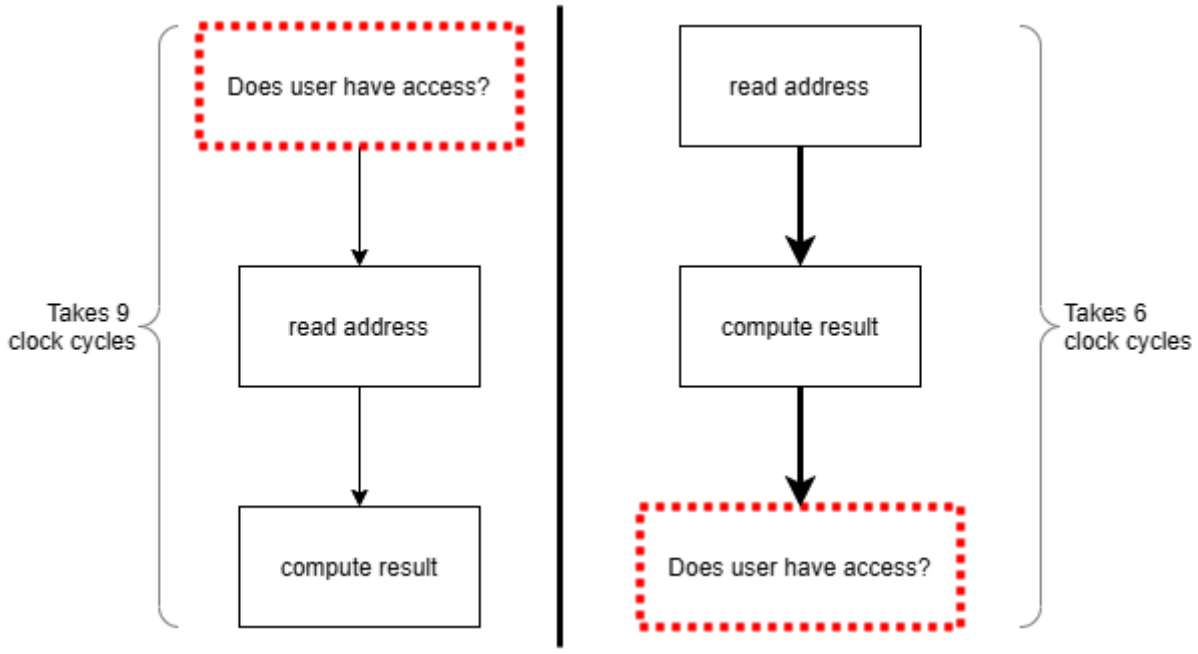
Chapter 14: Lake Formation – Advanced Topics

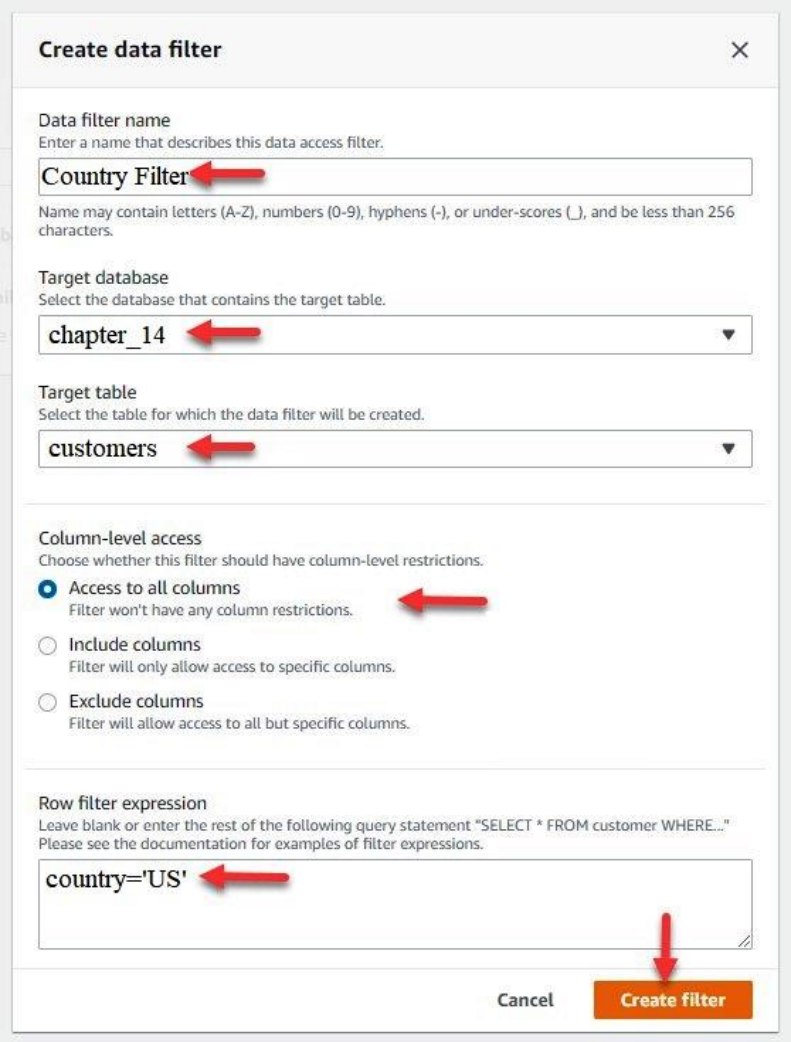



Lake Formation Query Engine Integrations
Athena
Lake Formation
Glue Data Catalog
S3
Logical Query Planning
Get Table Policy Metadata
Access Check
Yes
Physical Query Planning
Get Temporary Credentials
Get Partitions
List Objects
Execute TableScan
Get Object



Typical Analytics Node (e.g. Spark)




Create data filter
Data filter name
Enter a name that describes this data access filter.
Country Filter
Name may contain letters (A-Z), numbers (0-9), hyphens (-), or under-scores (_), and be less than 256 characters.
Target database
Select the database that contains the target table.
chapter_14
Target table
Select the table for which the data filter will be created.
customers
Column-level access
Choose whether this filter should have column-level restrictions.
Access to all columns
Filter won't have any column restrictions.
Include columns
Filter will only allow access to specific columns.
Exclude columns
Filter will allow access to all but specific columns.
Row filter expression
Leave blank or enter the rest of the following query statement "SELECT * FROM customer WHERE..."
Please see the documentation for examples of filter expressions.
country='US'
Cancel
Create filter