AWS Glue and Athena

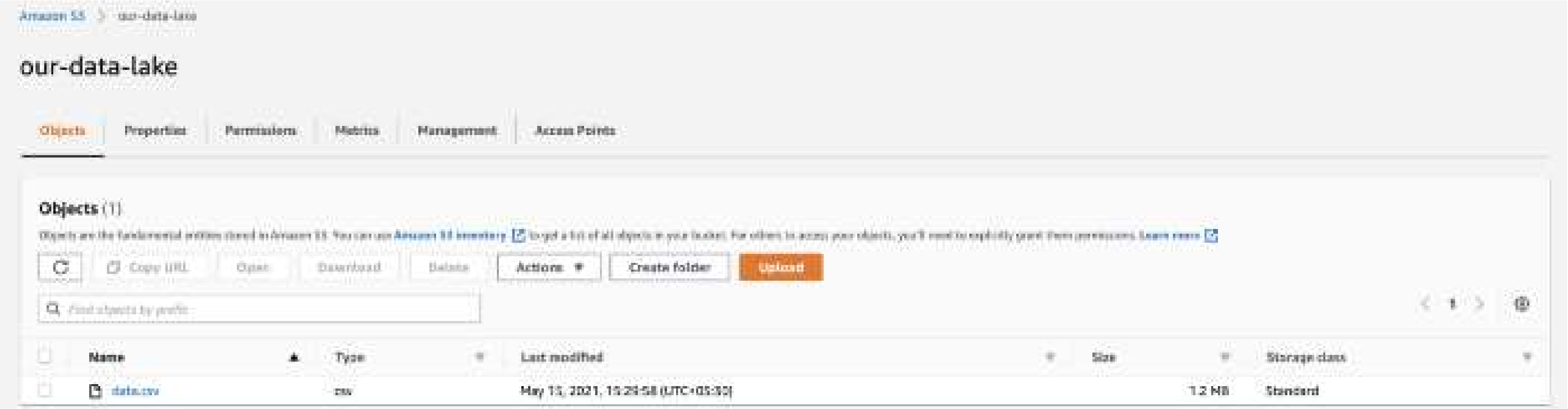
ETL Workflow Guide using Glue Studio with S3, and Athena

What is AWS Athena?

- A query service that makes it easy to analyze data in Amazon S3
- Serverless
- Easy to Use
- Supports the following data formats
    - Parquet (Optimized Row Columnar)
    - JSON
    - Avro
    - CSV
    - ORC (Optimized Row Columnar)
- Pay per query executed

Data

- Create se bucket named: <your-name>_data_lake
- Upload the CSV file to the S3 bucket
    - https://www.stats.govt.nz/assets/Uploads/Annual-enterprise-survey/Annual-enterprise-survey-2019-financial-year-provisional/Download-data/annual-enterprise-survey-2019-financial-year-provisional-size-bands-csv.csv

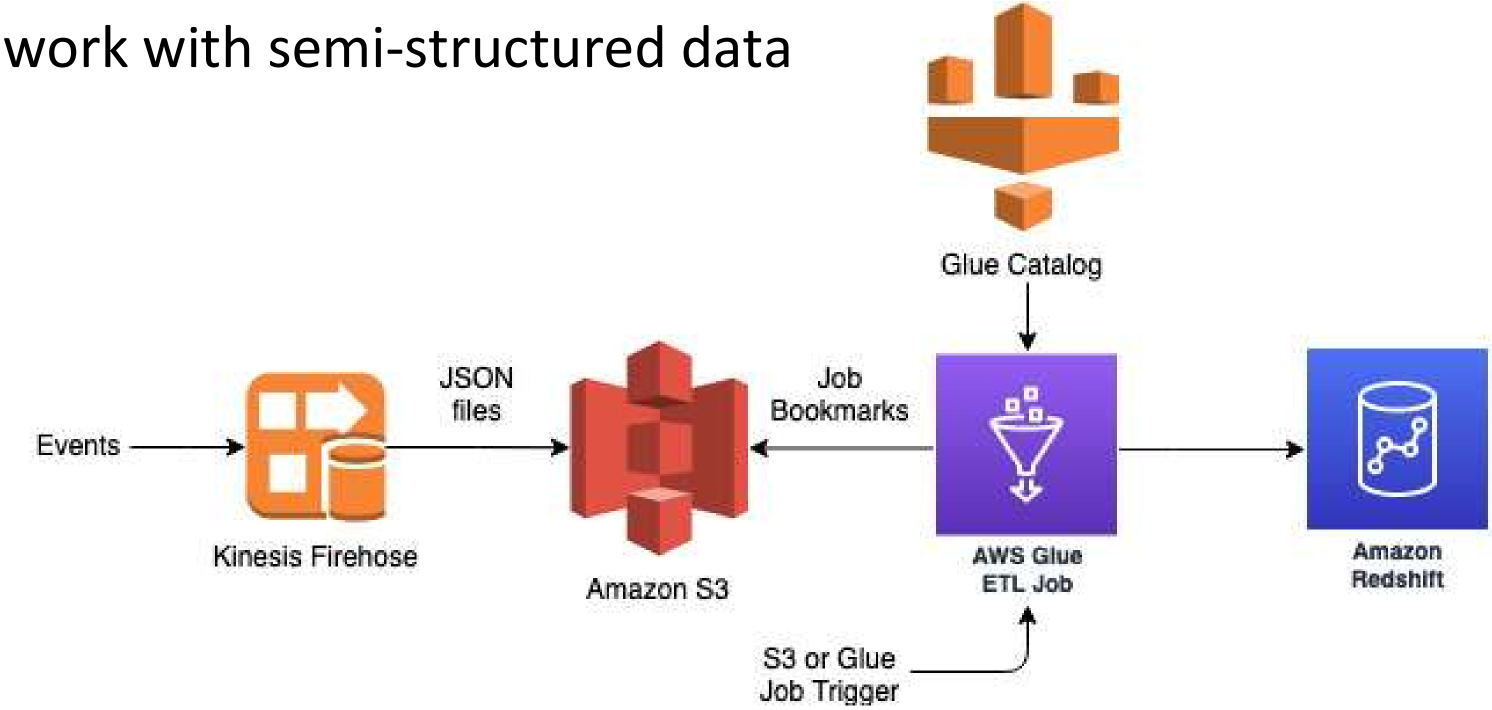


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	data.csv	csv	May 15, 2021, 15:29:58 (UTC+05:30)	1.2 MB	Standard

What is Glue?

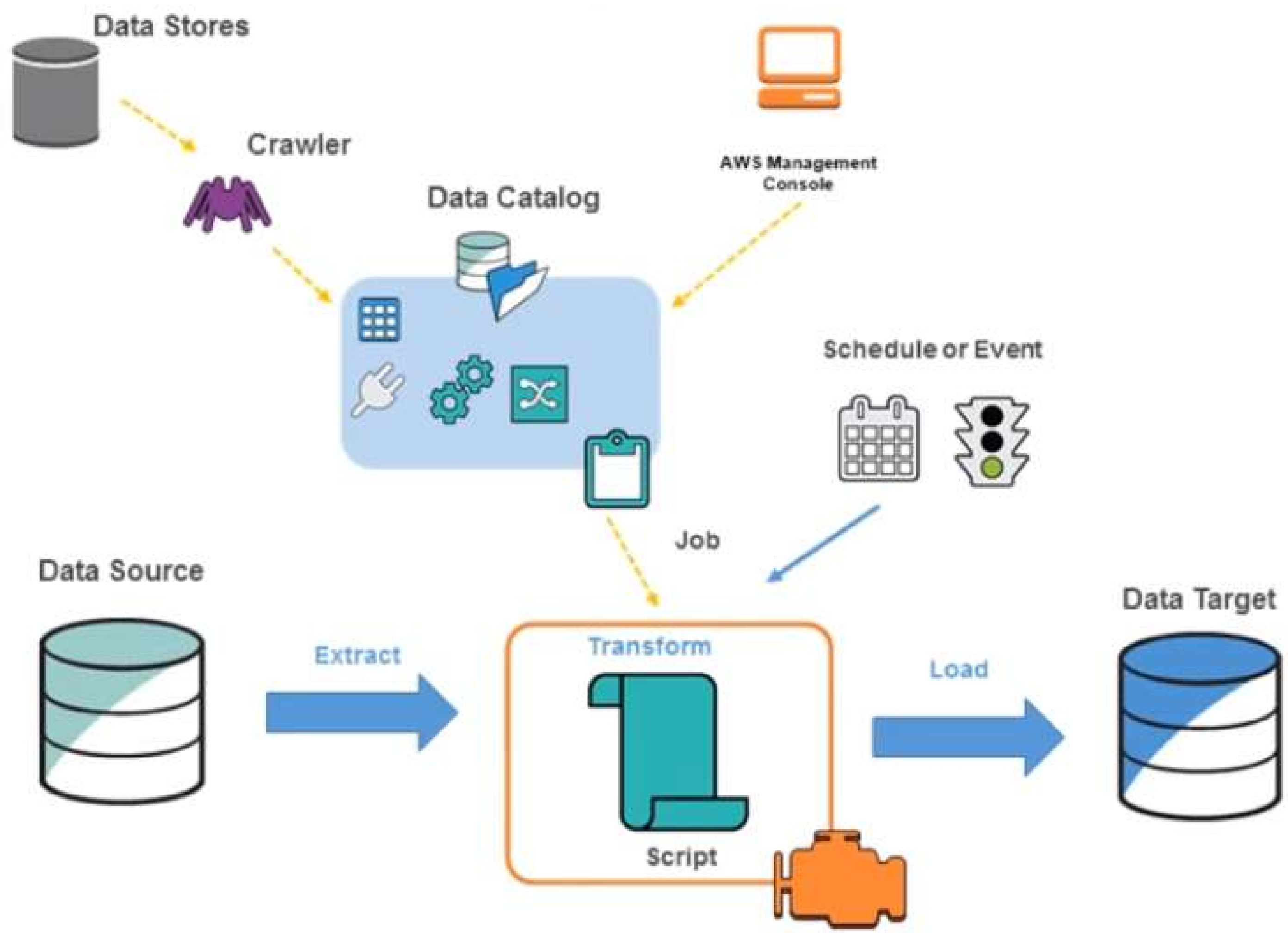
- A serverless, fully-managed, and cloud-optimized ETL service
- Runs on a Apache spark (https://spark.apache.org) environment
- Can be used when we need to clean, organise and format data
- Also designed to work with semi-structured data



Glue Benefits

- Server-less
- Has crawlers which identifies your data , suggest schemas and stores into a central catalog table
- Glue ETL engine automatically generates Python/Scala code
- We can build ETL pipelines
- With the support of Glue catalog, we can directly query S3 data using Athena
- Integrated with wide range of AWS services which helps us to build event-driven ETL pipelines
- API and AWS CLI support for all Glue operations.
- Scales resources as needed to run your jobs
- Retry tasks and handle errors automatically
- Less hassle and cost effective

Glue Architecture



Glue Components

- Crawlers
  - To populate tables in Glue Catalog
- Classifier
  - Checks if the given file is in pre-defined format or not
  - There are a couple of defined formats in Glue like: CSV, JSON, XML
  - We can also create or custom our own classifier
- Glue data catalog
  - Central repository of the metadata for our all data assets
  - Stores the table definition, location and many different attributes.
- Job authoring
  - Generates ETL code in Python and Scala
- Job execution/ Scheduler
  - Handles dependency, monitoring and alerting of the jobs

AWS Glue Use Cases

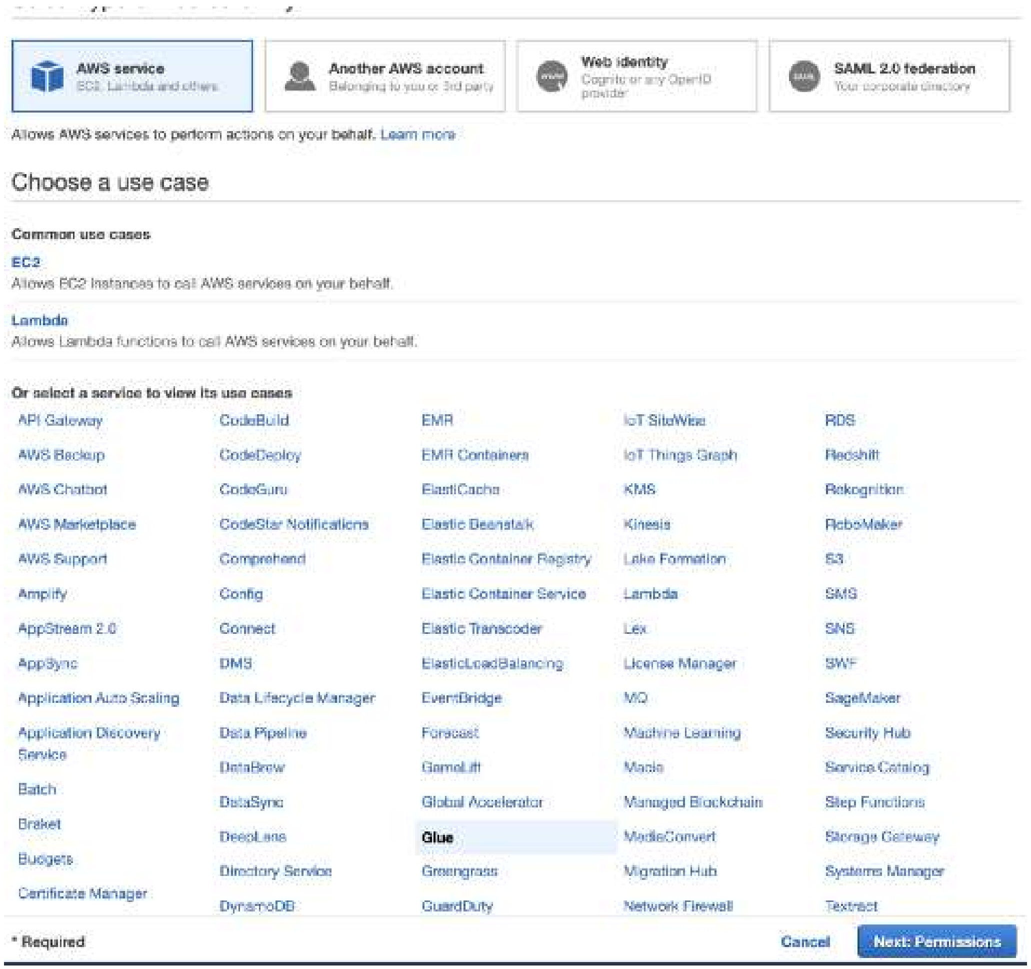
- To build data warehouse to organize, cleanse, validate and format data
- To run server-less queries against your Amazon S3 data lake
- To create event driven ETL pipelines
- To understand data assets

Crawler we will use

- There are multiple ways to connect to our data store
  - Crawler is the most popular method among ETL engineers
  - This Crawler will crawl the data from my S3, and based on available data, it will create a table schema.
- AWS Glue can be used to load a csv file from an S3 bucket into Glue, and then run SQL queries on this data in Athena.

Create an IAM role

- Create a role to give permission to different logged-in users.
- Go to https://console.aws.amazon.com/iam/
- Click on "Roles"
- Click "Create Role" button
- Choose Service - "Glue" and click on next.



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
AWS Backup
AWS Chatbot
AWS Marketplace
AWS Support
Amplify
AppStream 2.0
AppSync
Application Auto Scaling
Application Discovery Service
Batch
Braket
Budgets
Certificate Manager

CodeBuild
CodeDeploy
CodeGuru
CodeStar Notifications
Comprehend
Config
Connect
DMS
Data Lifecycle Manager
Data Pipeline
DataBrew
DataSync
DeepLens
Directory Service
DynamoDB

EMR
EMR Containers
ElastiCache
Elastic Beanstalk
Elastic Container Registry
Elastic Container Service
Elastic Transcoder
ElasticLoadBalancing
EventBridge
Forecast
GameLift
Global Accelerator
Glue
Greengrass
GuardDuty

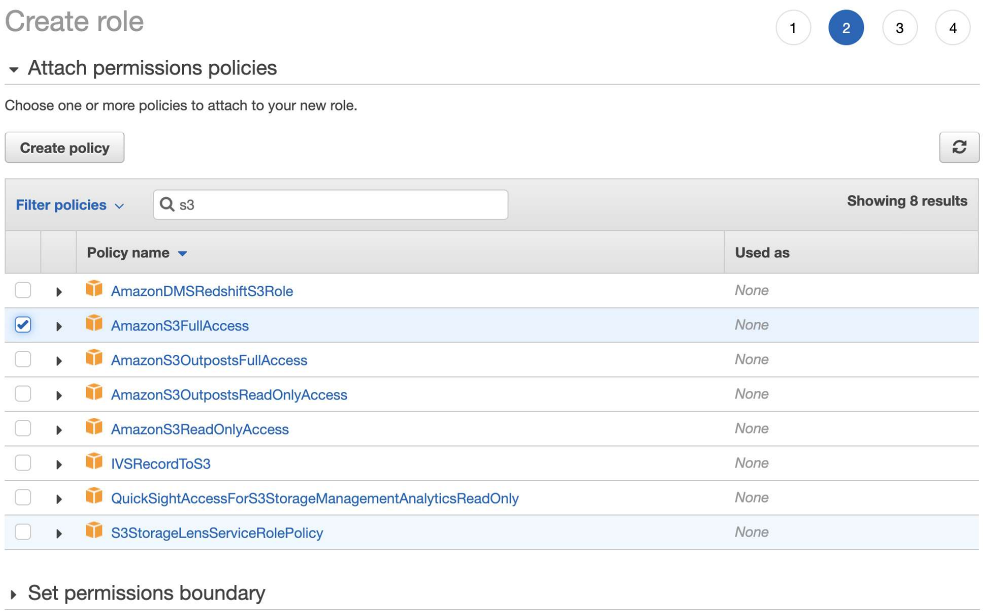
IoT SiteWise
IoT Things Graph
KMS
Kinesis
Lake Formation
Lambda
Lex
License Manager
MQ
Machine Learning
Macie
Managed Blockchain
MediaConvert
Migration Hub
Network Firewall

RDS
Redshift
Rekognition
RoboMaker
S3
SMS
SNS
SWF
SageMaker
Security Hub
Service Catalog
Step Functions
Storage Gateway
Systems Manager
Textract

* Required

Create an IAM role

- Then give permission to Glue and S3.



Create role

Attach permissions policies

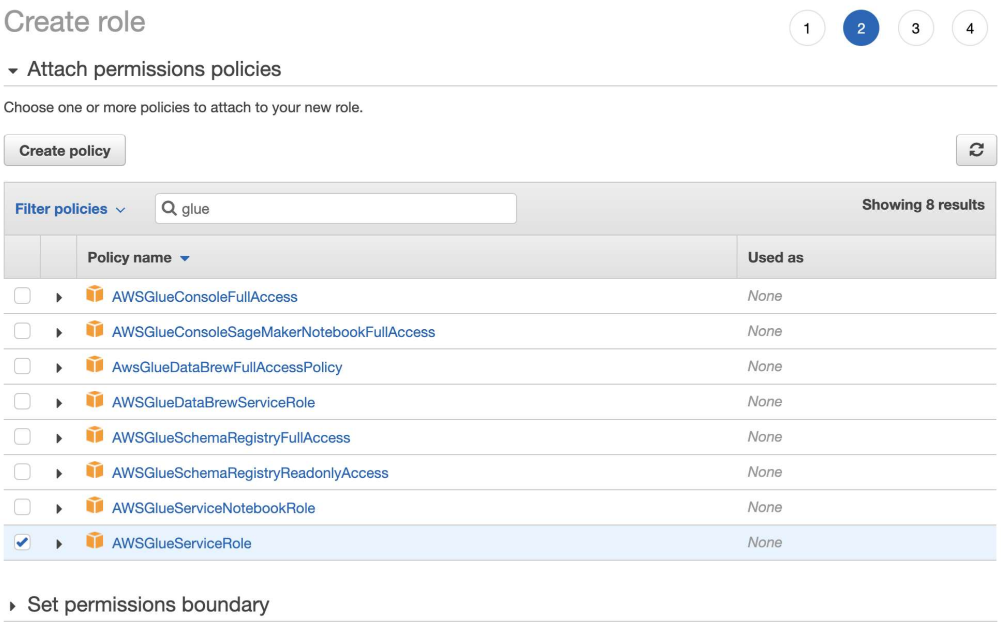
Choose one or more policies to attach to your new role.

s3

Showing 8 results

		Policy name	Used as
□	▸	AmazonDMSRedshiftS3Role	*None*
☑	▸	AmazonS3FullAccess	*None*
□	▸	AmazonS3OutpostsFullAccess	*None*
□	▸	AmazonS3OutpostsReadOnlyAccess	*None*
□	▸	AmazonS3ReadOnlyAccess	*None*
□	▸	IVSRecordToS3	*None*
□	▸	QuickSightAccessForS3StorageManagementAnalyticsReadOnly	*None*
□	▸	S3StorageLensServiceRolePolicy	*None*

Set permissions boundary



Create role

Attach permissions policies

Choose one or more policies to attach to your new role.

glue

Showing 8 results

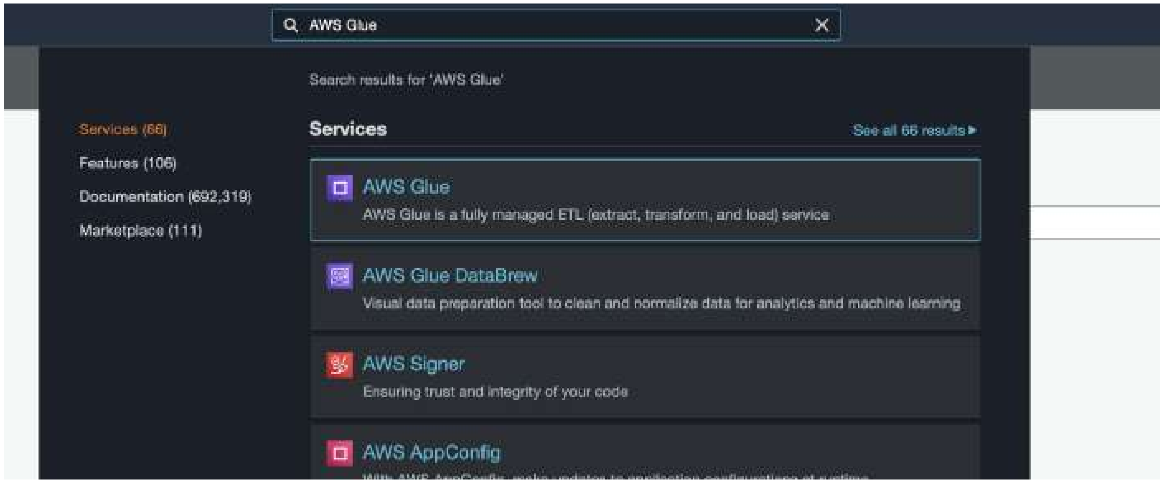
		Policy name	Used as
□	▸	AWSGlueConsoleFullAccess	*None*
□	▸	AWSGlueConsoleSageMakerNotebookFullAccess	*None*
□	▸	AwsGlueDataBrewFullAccessPolicy	*None*
□	▸	AWSGlueDataBrewServiceRole	*None*
□	▸	AWSGlueSchemaRegistryFullAccess	*None*
□	▸	AWSGlueSchemaRegistryReadonlyAccess	*None*
□	▸	AWSGlueServiceNotebookRole	*None*
☑	▸	AWSGlueServiceRole	*None*

Set permissions boundary

- Enter a role name (eg, AWSGlueServiceRole) and give some description about the role then
- Click on the “Create role” button and that's it a new role is created.

Configure AWS Glue Operation

- We are using AWS Glue to organize, cleanse, validate, and format data that is stored in S3.
- Search for "AWS Glue" in the AWS console and click on "crawlers".



AWS Glue

Search results for 'AWS Glue'

Services (66)

Features (106)

Documentation (692,319)

Marketplace (111)

Services

See all 66 results ▸

AWS Glue
AWS Glue is a fully managed ETL (extract, transform, and load) service

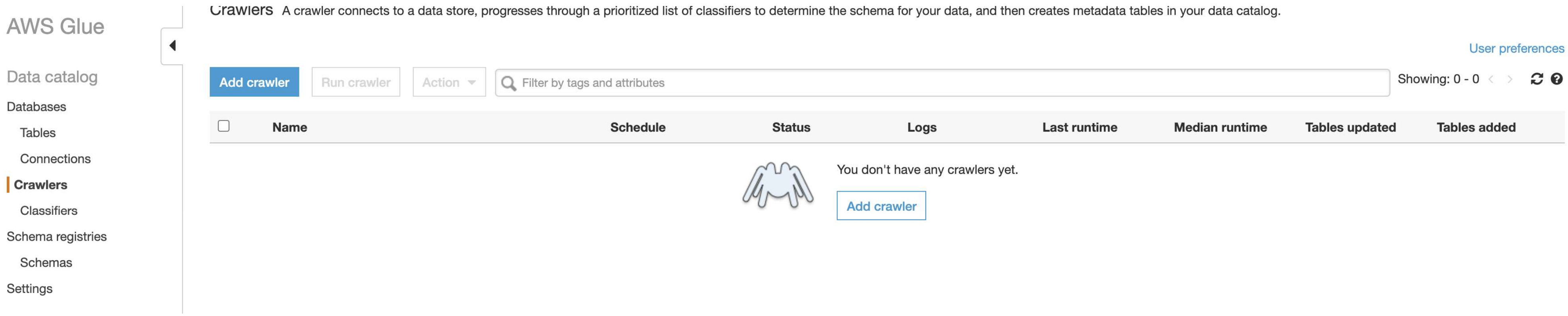
AWS Glue DataBrew
Visual data preparation tool to clean and normalize data for analytics and machine learning

AWS Signer
Ensuring trust and integrity of your code

AWS AppConfig

Configure AWS Glue Operation

- Click on Add Crawler and enter the crawler name (eg, dataLakeCrawler) and click on the “Next button”.



Crawlers A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

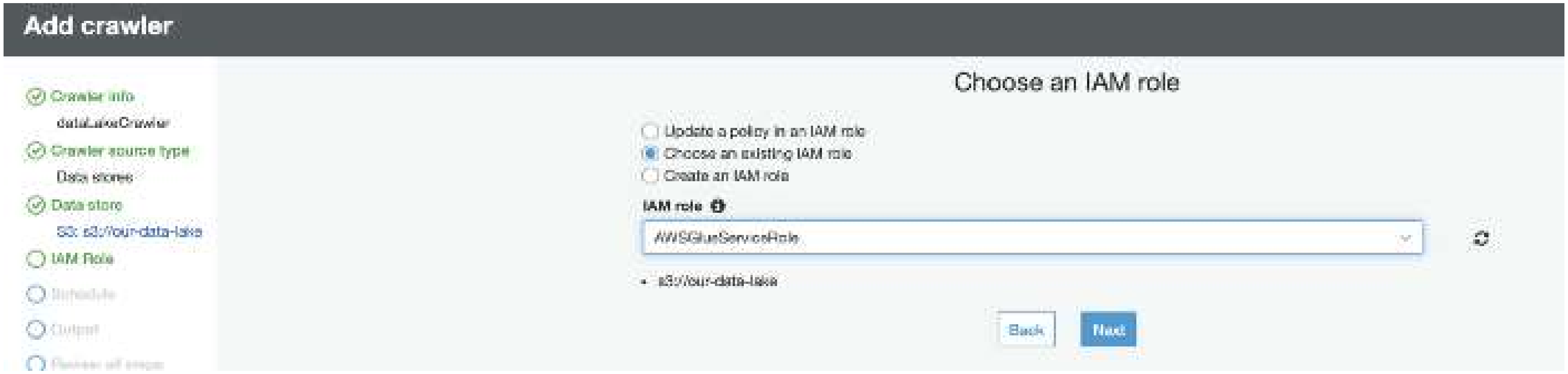
☐	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added

You don't have any crawlers yet.

- Select the data store as “S3” and give the path of the data that we have stored in our S3 bucket and click on “Next”
- Now skip adding another data store for now and click on “Next”.

Configure AWS Glue Operation

- Select the IAM role which we have created before and click “Next”



Add crawler

Choose an IAM role

○ Update a policy in an IAM role
● Choose an existing IAM role
○ Create an IAM role

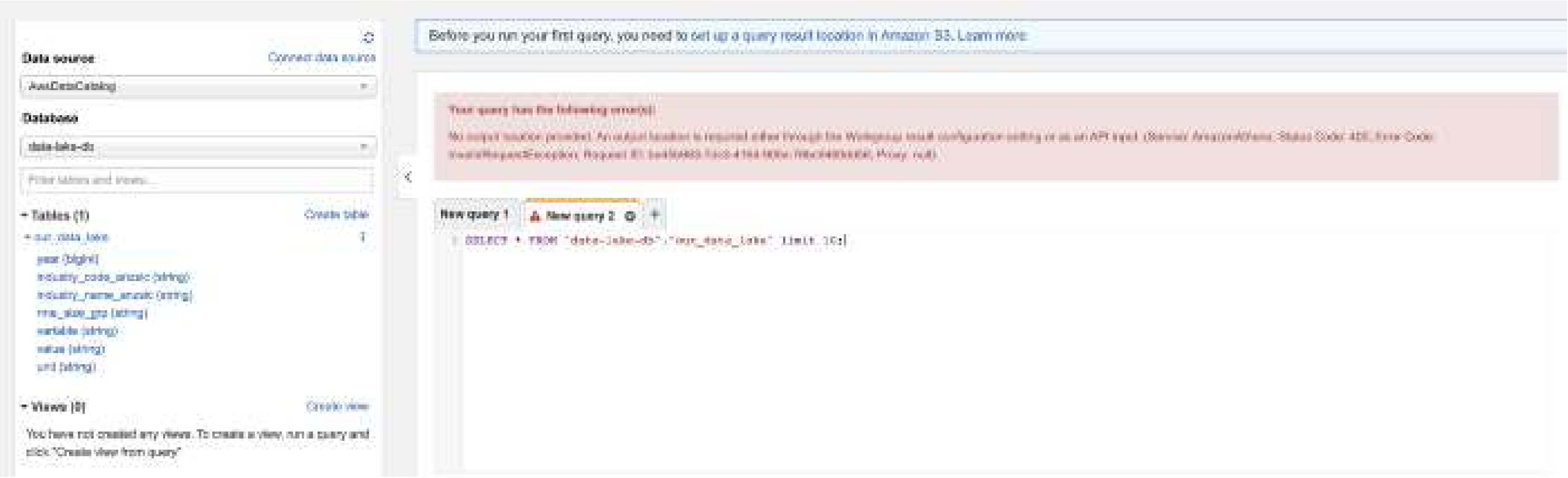
IAM role: AWSGlueServiceRole

- s3://our-data-lake

- Select the “Run on Demand” option and click “Next”.
- Click on “Add Database” and give the name “data-lake-db” then, click on “Next”.
- Review the AWS Glue crawler configuration and click on “Finish”

Running Query in Athena

- We can now
  - Go to AWS Athena
  - Select the database that we have created above (data-lake-db)
  - Execute our query using standard MySQL
    - SELECT * FROM “data-lake-db”.”<your-name>_data_lake” limit 10;



You have not created any views. To create a view, run a query and click "Create view from query"

Before you run your first query, you need to set up a query result location in Amazon S3. Learn more

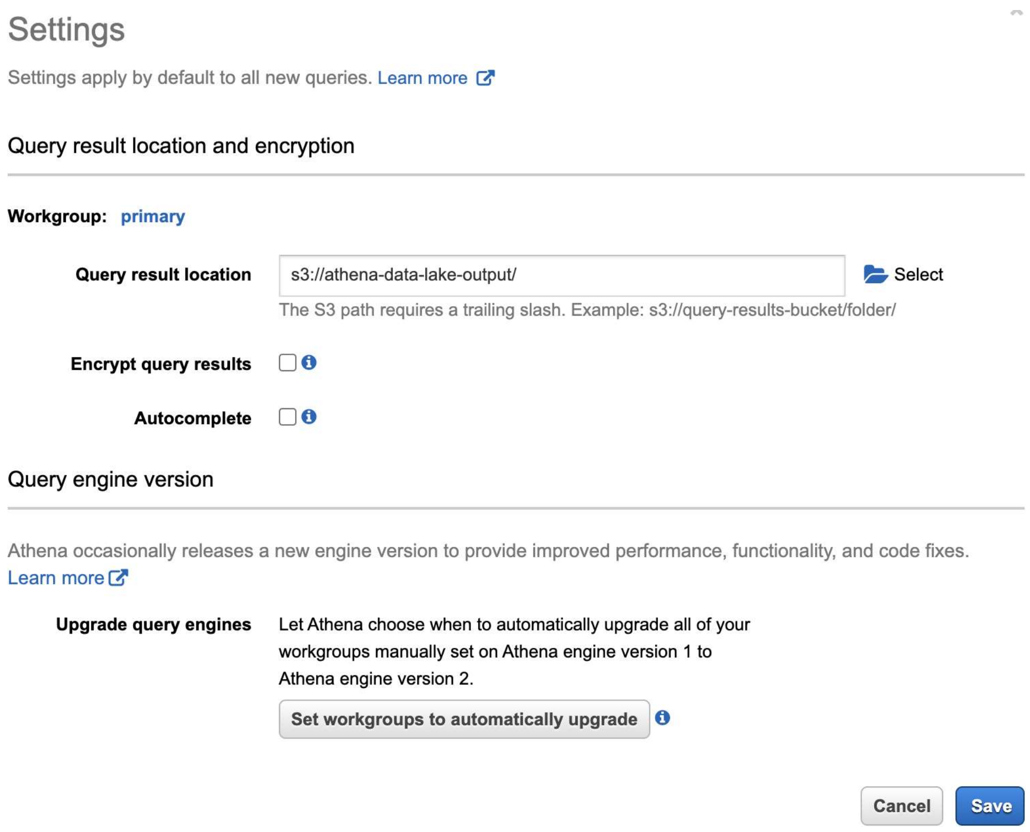
Your query has the following error(s):

No output location provided. An output location is required either through the Workgroup result configuration setting or as an API input. (Service: AmazonAthena; Status Code: 400; Error Code: InvalidRequestException; Request ID: [illegible]; Proxy: null)

```
SELECT * FROM "data-lake-db"."our_data_lake" limit 10;
```

Running Query in Athena

- We got an error stating that we need to provide “Output Location” before executing the query
- Now create an S3 bucket name as “athena-data-lake-output” and store the output of the query in this bucket by clicking on the “set up a query result location in Amazon S3” tab on the Athena management console



Settings

Settings apply by default to all new queries. Learn more

Query result location and encryption

Workgroup: primary

Query result location: s3://athena-data-lake-output/

The S3 path requires a trailing slash. Example: s3://query-results-bucket/folder/

Encrypt query results □

Autocomplete □

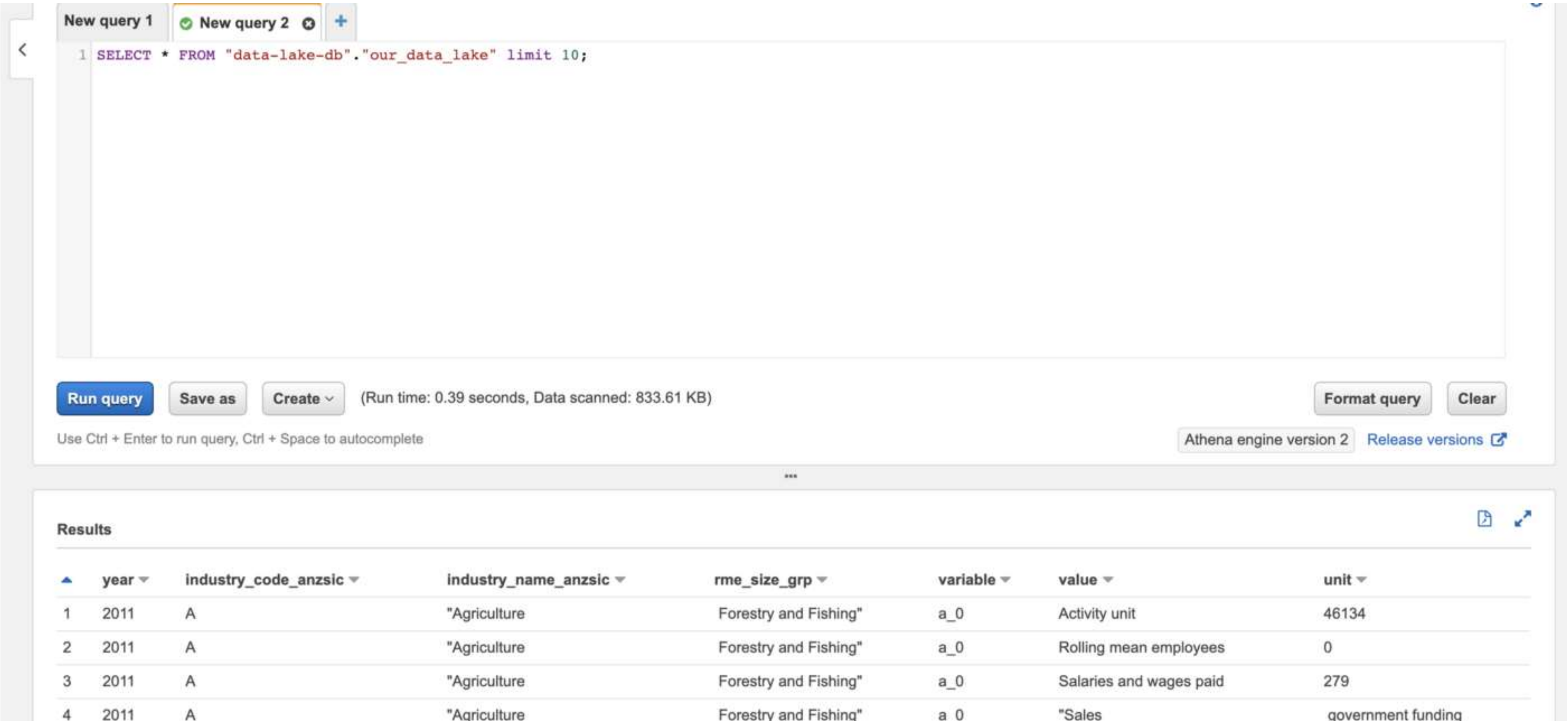
Query engine version

Athena occasionally releases a new engine version to provide improved performance, functionality, and code fixes. Learn more

Upgrade query engines: Let Athena choose when to automatically upgrade all of your workgroups manually set on Athena engine version 1 to Athena engine version 2.

Running Query in Athena

- Finally, we can run the same query and analyze the output.



```
SELECT * FROM "data-lake-db"."our_data_lake" limit 10;
```

(Run time: 0.39 seconds, Data scanned: 833.61 KB)

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

	year	industry_code_anzsic	industry_name_anzsic	rme_size_grp	variable	value	unit
1	2011	A	"Agriculture	Forestry and Fishing"	a_0	Activity unit	46134
2	2011	A	"Agriculture	Forestry and Fishing"	a_0	Rolling mean employees	0
3	2011	A	"Agriculture	Forestry and Fishing"	a_0	Salaries and wages paid	279
4	2011	A	"Agriculture	Forestry and Fishing"	a_0	"Sales	government funding

Thanks