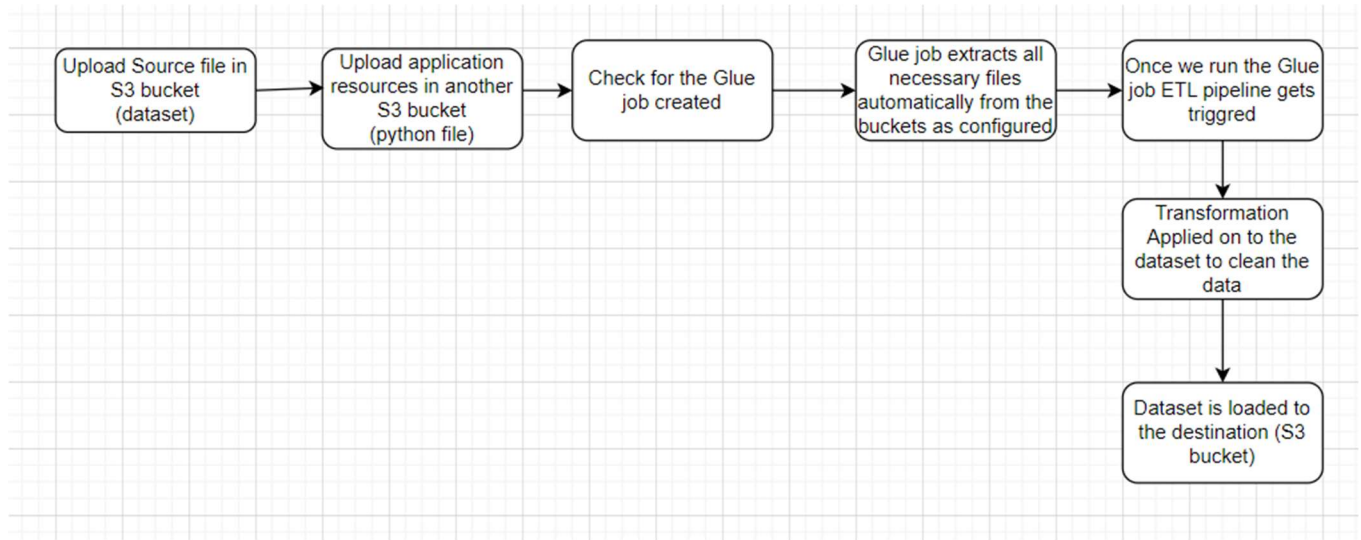


Documentation

Workflow



Initial Setup and configuration

Step1: Create AWS IAM user account and provide console access and programmatic access to the user account.

myuseraccount

Delete

Summary

ARN [redacted]	Console access Enabled without MFA	Access key 1 Used today. 3 days old.
Created June 22, 2023, 23:30 (UTC-04:00)	Last console sign-in Today	Access key 2 Not enabled

[Permissions](#) | [Groups](#) | [Tags](#) | [Security credentials](#) | [Access Advisor](#)

Permissions policies (3)
Permissions are defined by policies attached to the user directly or through groups.

Filter by Type: All types

Policy name	Type	Attached via
AmazonS3FullAccess	AWS managed	Directly
AWSGlueConsoleFullAccess	AWS managed	Directly
inlinepolicy	Customer inline	Inline

Permissions boundary (not set)

Step2: Configure the access key and secret key with the local environment and terraform installed.

Step3: Develop Terraform scripts to provision S3 buckets for the source, destination, and to store application resources (python files).

Buckets (3) [Info](#)

Copy content

Empty

Delete

Create bucket

< 1 >

	Name ▲	AWS Region ▼	Access ▼	Creation date ▼
<input type="radio"/>	my-destination-bucket-data	US East (Ohio) us-east-2	Objects can be public	June 25, 2023, 09:38:59 (UTC-04:00)
<input type="radio"/>	my-python-script-bucket	US East (Ohio) us-east-2	Objects can be public	June 25, 2023, 09:38:59 (UTC-04:00)
<input type="radio"/>	my-source-bucket-data	US East (Ohio) us-east-2	Objects can be public	June 25, 2023, 09:38:59 (UTC-04:00)

Step4: Develop Terraform script to create a Glue job and provide path to the S3 bucket fetch the python file.

Source

Amazon S3
JSON, CSV, or Parquet files stored in S3.

→

Target

Amazon S3
S3 bucket by specifying a bucket path as the data target.

Your jobs (1) [Info](#)

Actions ▼

Run job

< 1 >

<input type="checkbox"/>	Job name ▼	Type	Last modified ▼	AWS Glue version ▼
<input type="checkbox"/>	gluejob	Glue ETL	6/26/2023, 12:18:33 PM	3.0

Step5: Create an IAM role with permissions to access AWS Glue and S3 service.

gluerole Delete

Summary Edit

Creation date June 25, 2023, 11:59 (UTC-04:00)	ARN [REDACTED]	Link to switch roles in console [REDACTED]
Last activity 2 hours ago	Maximum session duration 1 hour	

Permissions | Trust relationships | Tags | Access Advisor | Revoke sessions

Permissions policies (3) Info Refresh Simulate Remove Add permissions

You can attach up to 10 managed policies.

Filter policies by property or policy name and press enter.

<input type="checkbox"/>	Policy name	Type	Description
<input type="checkbox"/>	AmazonS3FullAccess	AWS managed	Provides full access to all buckets via the AWS Management Console.
<input type="checkbox"/>	AWSGlueServiceRole	AWS managed	Policy for AWS Glue service role which allows access to related services including EC2, S3...
<input type="checkbox"/>	AWSGlueConsoleFullAccess	AWS managed	Provides full access to AWS Glue via the AWS Management Console

Step6: Provide AWS Glue service and S3 service permissions to the IAM user created and pass the role to the IAM user with the help of an inline policy.

```
1 {
2   "Version": "2012-10-17",
3   "Statement": [
4     {
5       "Effect": "Allow",
6       "Action": [
7         "iam:GetRole",
8         "iam:PassRole"
9       ],
10      "Resource": "arn:aws:iam::[REDACTED]:role/gluerole"
11    }
12  ]
13 }
```

Deploying and Testing ETL pipeline

Step1: once we run “terraform apply” command all the infrastructure gets provisioned.

Step2: Ensure we have the .csv file (dataset) in source S3 bucket, if not upload the file.

my-source-bucket-data [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (1)
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](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	airquality.csv	csv	June 26, 2023, 12:46:53 (UTC-04:00)	3.1 KB	Standard

Step3: Ensure we have the python file uploaded in my-python-script-bucket, if not upload the file.

my-python-script-bucket [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (1)
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](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	pythonscript.py	py	June 26, 2023, 12:53:48 (UTC-04:00)	1.8 KB	Standard

Step4: Check if the glue job has been created. If it is created, then we should be able to get the python script into the glue job (remember we configured the path of S3 bucket using terraform script).

Step5: Once we run the glue job, the python file will be able to extract the dataset, transform it and load it to the destination.

AWS Glue > Monitoring

Monitoring [Info](#)

Date range
7 Day ▼

Job runs summary

Total runs	Running	Canceled	Success	Failed	Success rate	DPU hours
120	0	0	25	95	21%	16

Job runs (121) [Info](#) [Refresh](#) [Actions](#) [View CloudWatch logs](#) [View run details](#)

Filter job runs by property

	Job name ▼	Type ▼	Start time ▼	End time ▼	Run status ▼	Run time ▼	Capacity ▼	Worker type ▼	DPU hours
<input type="radio"/>	gluejob	Glue ETL	06/26/2023 15:45:36	06/26/2023 15:47:08	✓ Succeeded	1 minute	10	G.1X	0.24
<input type="radio"/>	gluejob	Glue ETL	06/26/2023 12:54:00	06/26/2023 12:55:19	✓ Succeeded	1 minute	10	G.1X	0.20

Step6: python code in the glue will check for the null values in the dataset and remove them thus handles the inconsistencies in the dataset. Then the dataset gets loaded to the destination S3 bucket.

my-destination-bucket-data [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (25)

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](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

Find objects by prefix

<input type="checkbox"/>	Name ▲	Type ▼	Last modified ▼	Size ▼	Storage class ▼
<input type="checkbox"/>	run-1687704295530-part-r-00000	-	June 25, 2023, 10:45:13 (UTC-04:00)	3.2 KB	Standard
<input type="checkbox"/>	run-1687710075385-part-r-00000	-	June 25, 2023, 12:21:21 (UTC-04:00)	3.2 KB	Standard
<input type="checkbox"/>	run-1687710270460-part-r-00000	-	June 25, 2023, 12:24:49 (UTC-04:00)	3.2 KB	Standard
<input type="checkbox"/>	run-1687711284199-	-	June 25, 2023, 12:41:41	3.2 KB	Standard

Step7: Python script runs the ETL pipeline once the glue job is triggered.

