Social Media Data Integration & Analysis

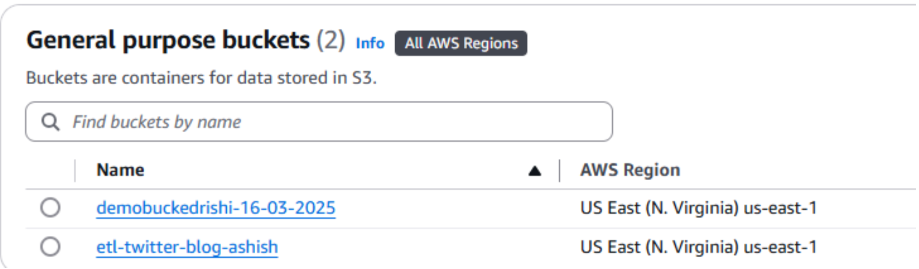
MADE BY- Ashish Chamel

COURSE- Extract, Transform, and Load (ETL)

DATE OF SUBMISSON-05/04/2025

Step-1

1. Creation of a input bucket in s3 ie etl-twitter-blog-ashish



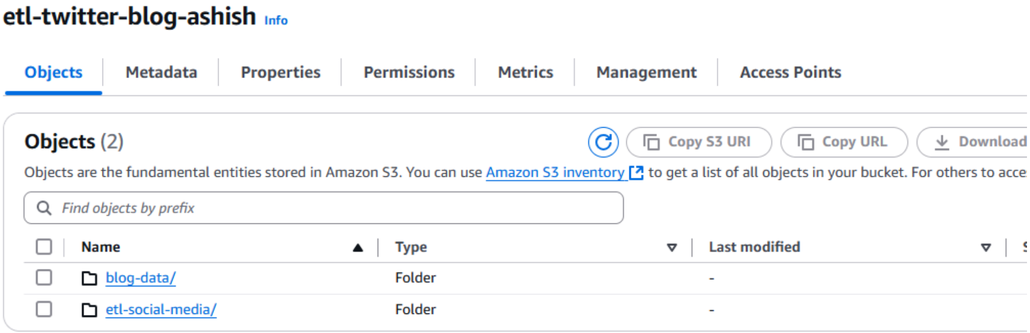
Buckets are containers for data stored in S3.

	Name ▲	AWS Region
○	demobuckedrishi-16-03-2025	US East (N. Virginia) us-east-1
○	etl-twitter-blog-ashish	US East (N. Virginia) us-east-1

Step-2

1. Creation of two folders inside etl-twitter-blog-ashish bucket ie

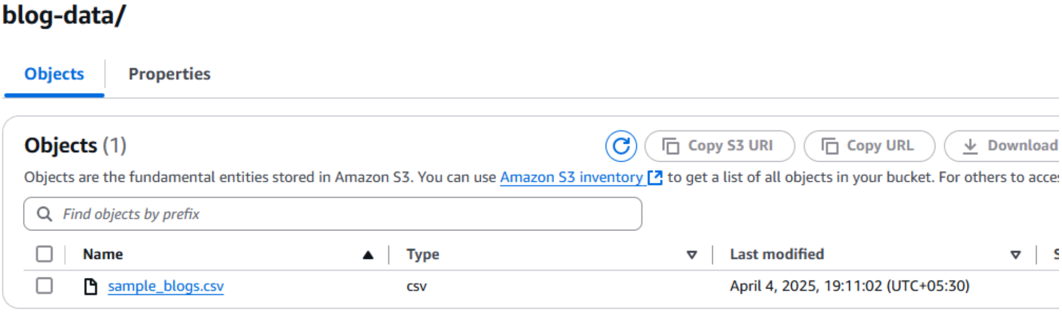
   - blog-data and etl-social-media



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 acce

□	Name ▲	Type ▽	Last modified ▽
□	blog-data/	Folder	-
□	etl-social-media/	Folder	-

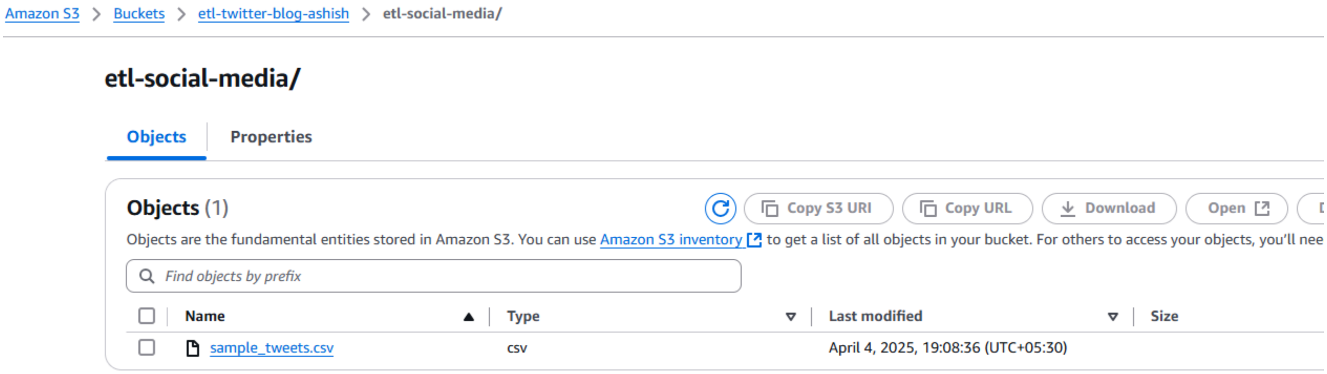
2. In blog-data folder add sample_blogs.csv



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 acce

□	Name ▲	Type ▽	Last modified ▽
□	sample_blogs.csv	csv	April 4, 2025, 19:11:02 (UTC+05:30)

3. In etl-social-media folder add sample_tweets.csv


Amazon S3 > Buckets > etl-twitter-blog-ashish > etl-social-media/

etl-social-media/

Objects (1)

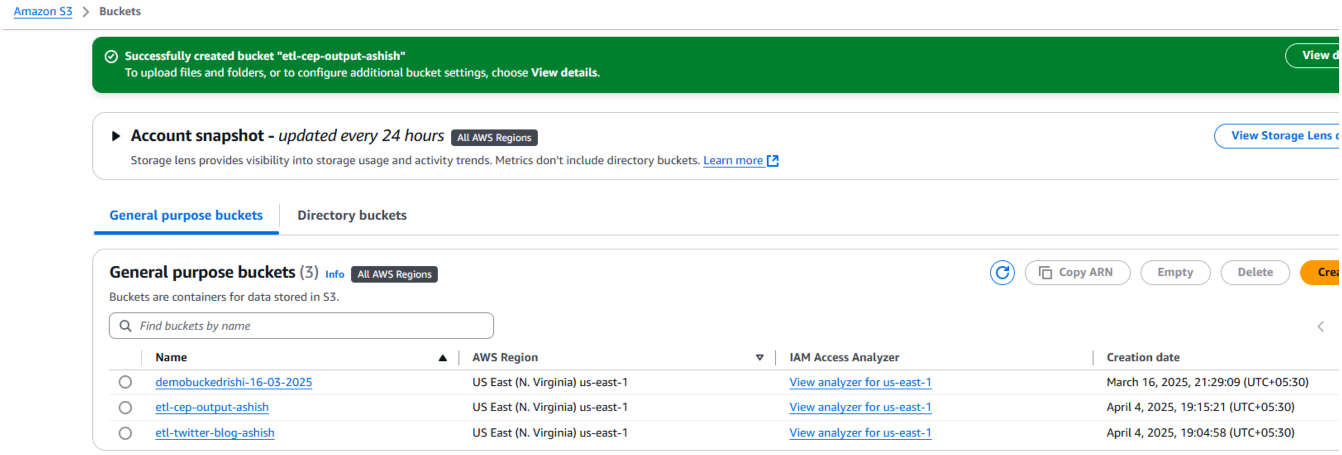
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll nee

Find objects by prefix

□	Name	Type	Last modified	Size
□	sample_tweets.csv	csv	April 4, 2025, 19:08:36 (UTC+05:30)	

Step-3

1. Creation of a new output bucket named etl-cep-output-ashish


Amazon S3 > Buckets

Successfully created bucket "etl-cep-output-ashish"
To upload files and folders, or to configure additional bucket settings, choose **View details**.

Account snapshot - *updated every 24 hours* All AWS Regions

Storage lens provides visibility into storage usage and activity trends. Metrics don't include directory buckets. Learn more

General purpose buckets (3) Info All AWS Regions

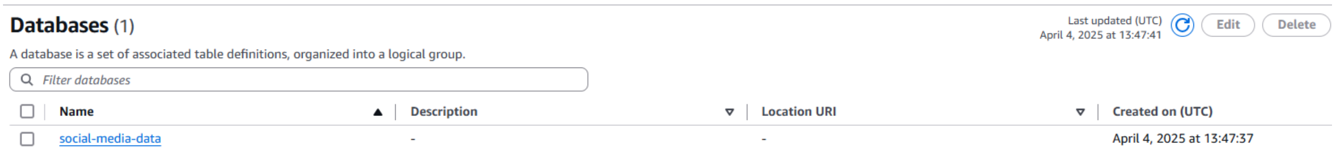
Buckets are containers for data stored in S3.

Find buckets by name

	Name	AWS Region	IAM Access Analyzer	Creation date
○	demobuckedrishi-16-03-2025	US East (N. Virginia) us-east-1	View analyzer for us-east-1	March 16, 2025, 21:29:09 (UTC+05:30)
○	etl-cep-output-ashish	US East (N. Virginia) us-east-1	View analyzer for us-east-1	April 4, 2025, 19:15:21 (UTC+05:30)
○	etl-twitter-blog-ashish	US East (N. Virginia) us-east-1	View analyzer for us-east-1	April 4, 2025, 19:04:58 (UTC+05:30)

Step-4

1. Navigating to AWS GLUE and creating a new database ie. Social-media-data


Databases (1)

Last updated (UTC)
April 4, 2025 at 13:47:41

A database is a set of associated table definitions, organized into a logical group.

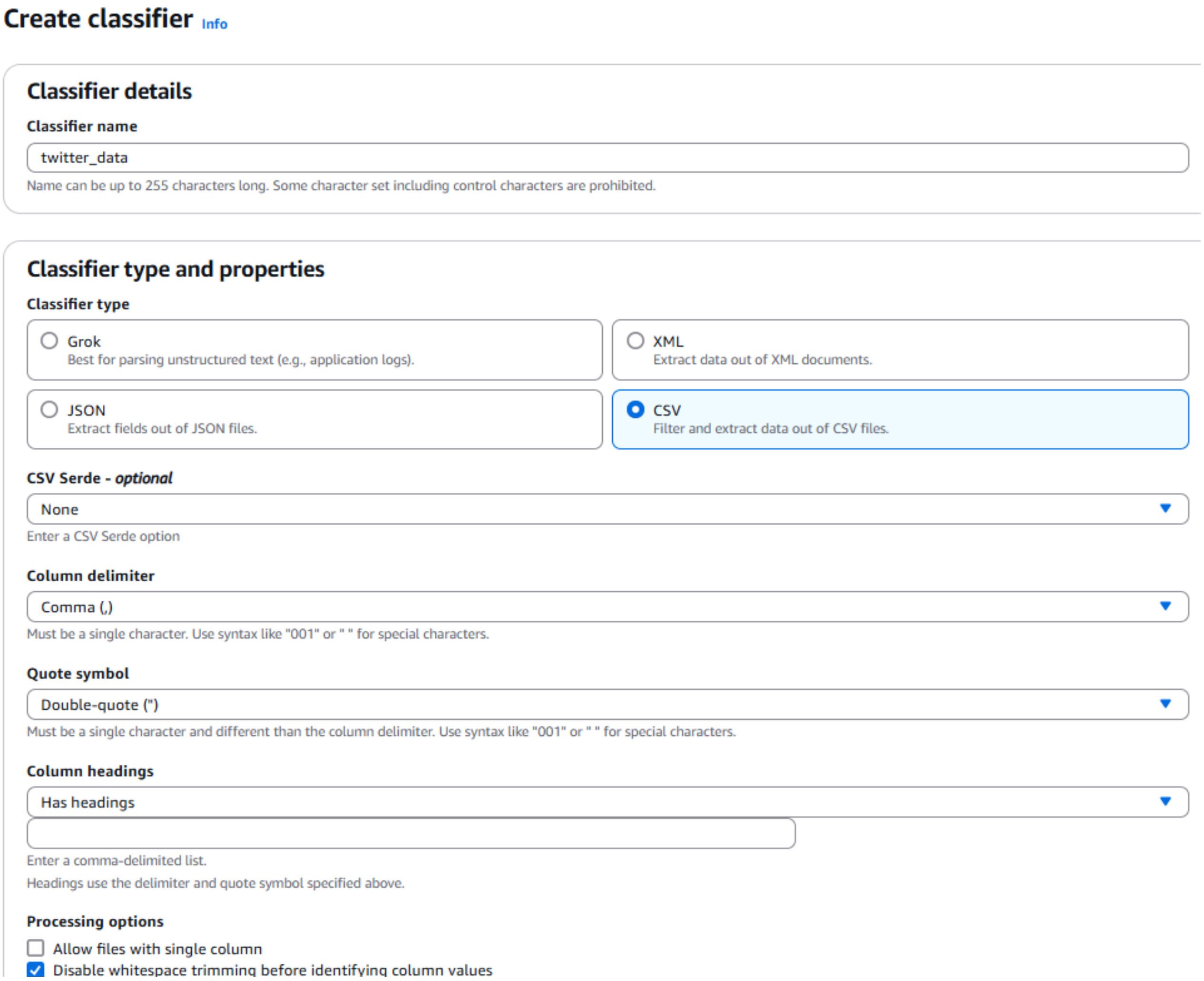
Filter databases

□	Name	Description	Location URI	Created on (UTC)
□	social-media-data	-	-	April 4, 2025 at 13:47:37

Step-5

1. Classifier creation ie
   - Twitter_data
   - Blog_data

2. Creation of twitter_data classifier its settings are as follows
    - Classifier name as twitter_data
    - Classifier type and properties as CSV
    - CSV Serde – optional as None o Column delimiter as comma(,)
    - Quote symbol as Double-quote(“)
    - Column headings as Has headings



Create classifier

Classifier details

Classifier name

twitter_data

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Classifier type and properties

Classifier type

○ Grok
Best for parsing unstructured text (e.g., application logs).

○ XML
Extract data out of XML documents.

○ JSON
Extract fields out of JSON files.

● CSV
Filter and extract data out of CSV files.

CSV Serde - *optional*

None

Enter a CSV Serde option

Column delimiter

Comma (,)

Must be a single character. Use syntax like "001" or " " for special characters.

Quote symbol

Double-quote (")

Must be a single character and different than the column delimiter. Use syntax like "001" or " " for special characters.

Column headings

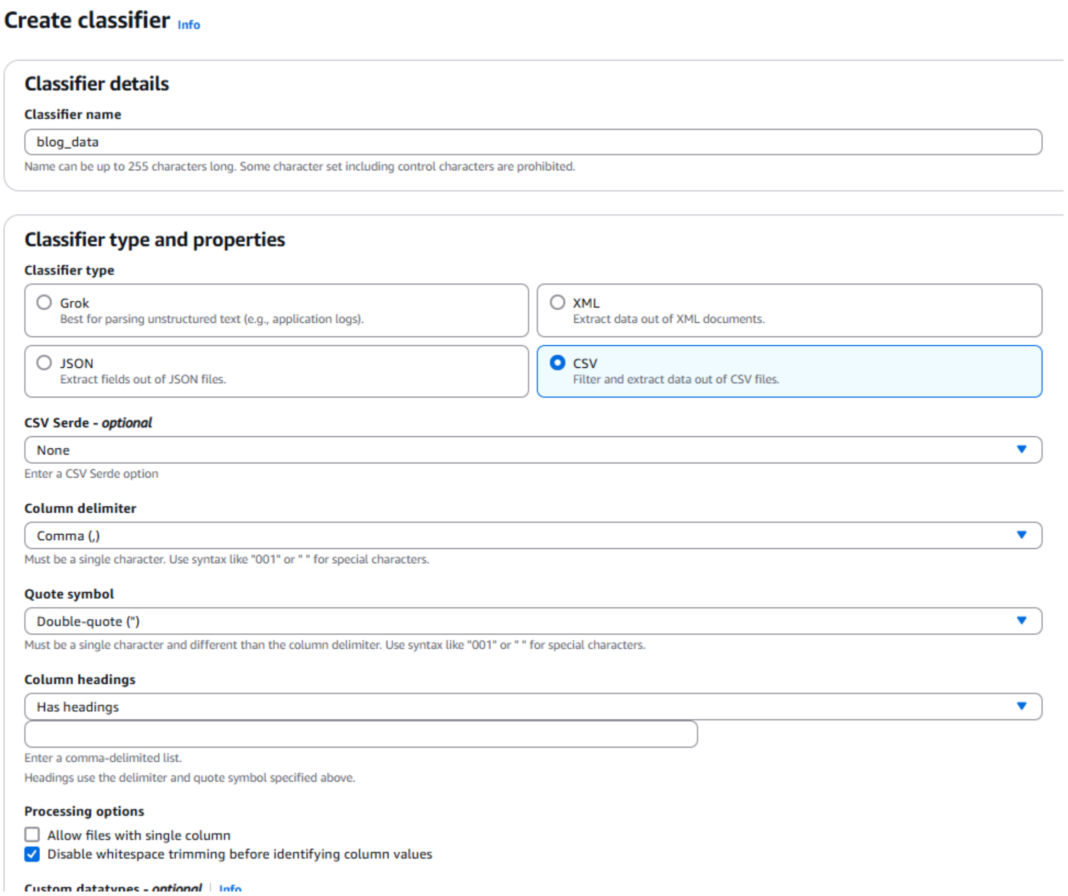
Has headings

Enter a comma-delimited list.
Headings use the delimiter and quote symbol specified above.

Processing options

□ Allow files with single column
☑ Disable whitespace trimming before identifying column values

3. Creation of a 2nd classifier ie blog_data and its settings are as follows-
    - Classifier name as blog_data
    - Classifier type and properties as CSV
    - CSV Serde – optional as None
    - Column delimiter as comma(,)
    - Quote symbol as Double-quote(“)
    - Column headings as Has headings



Create classifier Info

Classifier details

Classifier name

blog_data

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Classifier type and properties

Classifier type

○ Grok
Best for parsing unstructured text (e.g., application logs).

○ XML
Extract data out of XML documents.

○ JSON
Extract fields out of JSON files.

● CSV
Filter and extract data out of CSV files.

CSV Serde - *optional*

None

Enter a CSV Serde option

Column delimiter

Comma (,)

Must be a single character. Use syntax like "001" or " " for special characters.

Quote symbol

Double-quote (")

Must be a single character and different than the column delimiter. Use syntax like "001" or " " for special characters.

Column headings

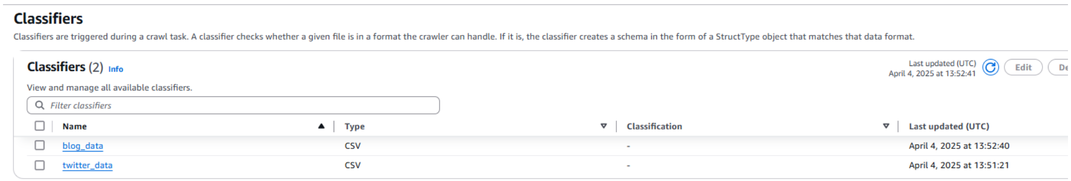
Has headings

Enter a comma-delimited list.
Headings use the delimiter and quote symbol specified above.

Processing options

□ Allow files with single column
☑ Disable whitespace trimming before identifying column values

Custom datatypes - *optional* | Info



Classifiers

Classifiers are triggered during a crawl task. A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Classifiers (2) Info

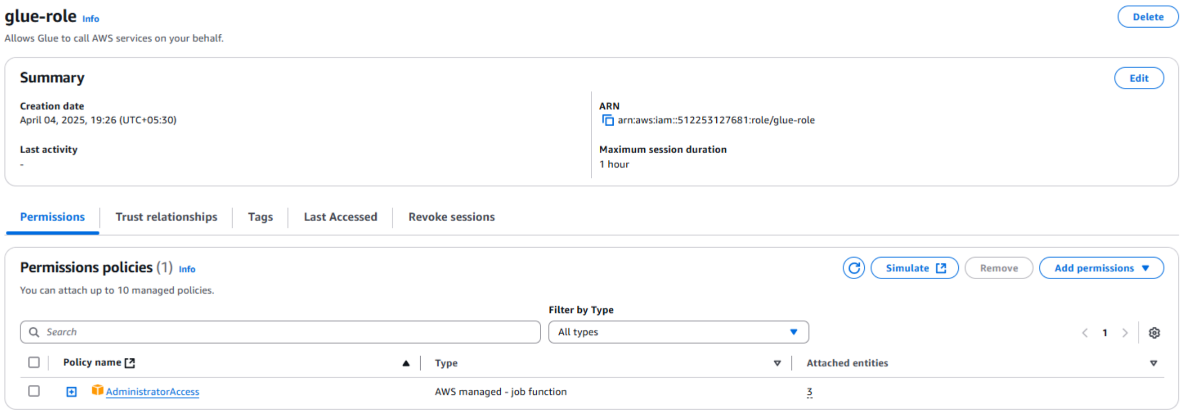
View and manage all available classifiers.

Last updated (UTC)
April 4, 2025 at 13:52:41

□	Name	Type	Classification	Last updated (UTC)
□	blog_data	CSV	-	April 4, 2025 at 13:52:40
□	twitter_data	CSV	-	April 4, 2025 at 13:51:21

Step-6

Navigated to the AWS console and configured an IAM role ie glue-role given administrator access.



glue-role Info

Allows Glue to call AWS services on your behalf.

Summary

Creation date
April 04, 2025, 19:26 (UTC+05:30)

Last activity
-

ARN
arn:aws:iam::512253127681:role/glue-role

Maximum session duration
1 hour

Permissions policies (1) Info

You can attach up to 10 managed policies.

Filter by Type

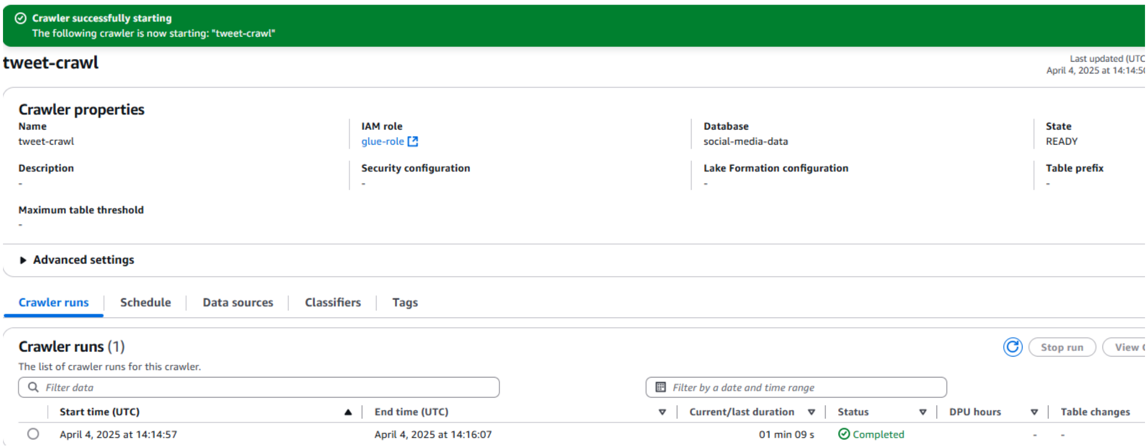
All types

□	Policy name	Type	Attached entities
□	AdministratorAccess	AWS managed - job function	3

Step-7

1. In AWS GLUE creation of the 1st crawler as tweet-crawl where
    - IAM role is glue-role
    - Database is social-media data

2. Running the tweet-crawl successfully to get metadata



Crawler successfully starting
The following crawler is now starting: "tweet-crawl"

tweet-crawl

Last updated (UTC
April 4, 2025 at 14:14:5

Crawler properties

Name tweet-crawl	**IAM role** glue-role	**Database** social-media-data	**State** READY
Description -	**Security configuration** -	**Lake Formation configuration** -	**Table prefix** -
Maximum table threshold -			

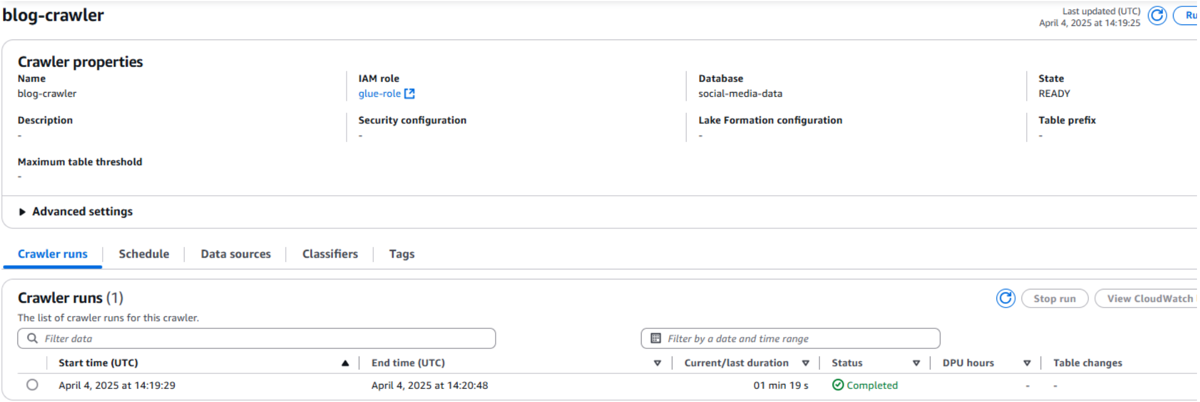
Crawler runs (1)

The list of crawler runs for this crawler.

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	April 4, 2025 at 14:14:57	April 4, 2025 at 14:16:07	01 min 09 s	✓ Completed	-	-

3. In AWS GLUE creation of a 2nd crawler as blog-crawler
    - IAM role is glue-role
    - Database is social-media data
    -

4. Running the blog-crawler successfully to get metadata



blog-crawler

Last updated (UTC)
April 4, 2025 at 14:19:25

Crawler properties

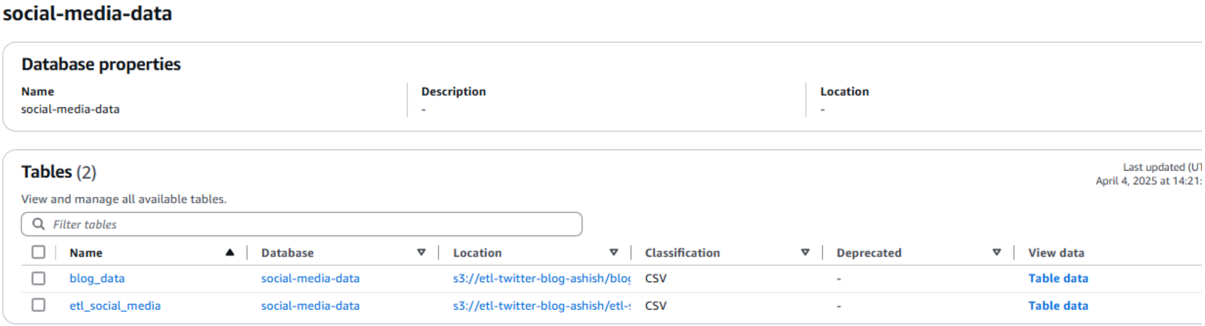
Name blog-crawler	**IAM role** glue-role	**Database** social-media-data	**State** READY
Description -	**Security configuration** -	**Lake Formation configuration** -	**Table prefix** -
Maximum table threshold -			

Crawler runs (1)

The list of crawler runs for this crawler.

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	April 4, 2025 at 14:19:29	April 4, 2025 at 14:20:48	01 min 19 s	✓ Completed	-	-

5. Metadata inside the social-media-data database as follows:



social-media-data

Database properties

Name social-media-data	**Description** -	**Location** -

Tables (2)

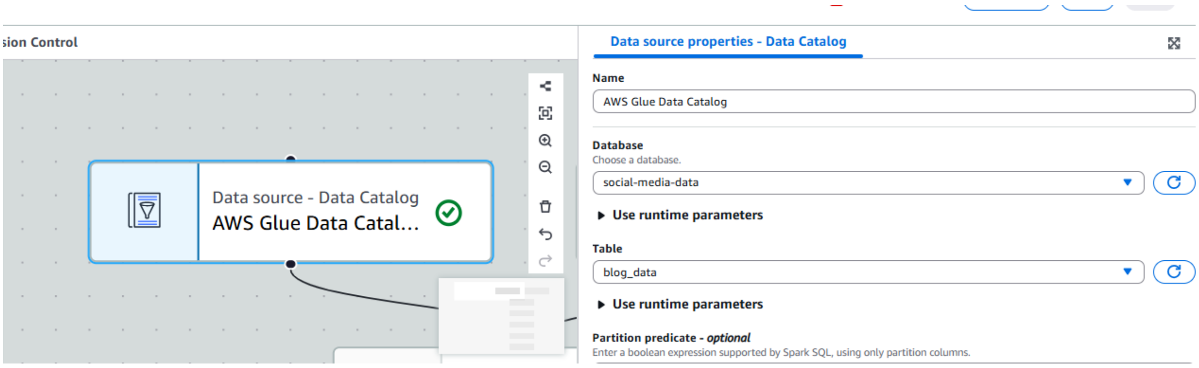
Last updated (U
April 4, 2025 at 14:21:

View and manage all available tables.

□	Name	Database	Location	Classification	Deprecated	View data
□	blog_data	social-media-data	s3://etl-twitter-blog-ashish/blo	CSV	-	Table data
□	etl_social_media	social-media-data	s3://etl-twitter-blog-ashish/etl-	CSV	-	Table data

Step-8

1. Navigating to AWS GLUE in ETL jobs using Visual ETL for a job creation.
2. Importing AWS GLUE DATA CATALOG 1, settings are as follows
    - Database- social-media-data
    - Table- blog_data



Data source properties - Data Catalog

Name
AWS Glue Data Catalog

Database
Choose a database.
social-media-data

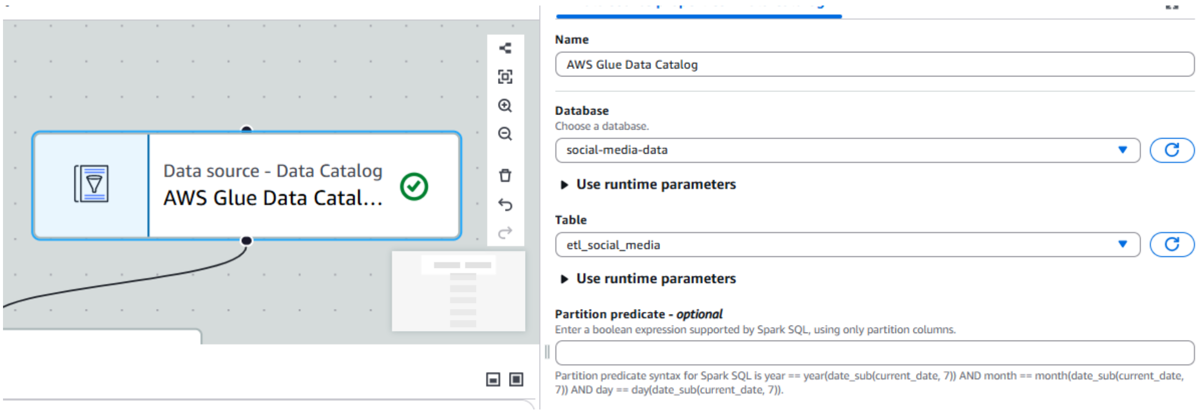
▶ Use runtime parameters

Table
blog_data

▶ Use runtime parameters

Partition predicate - *optional*
Enter a boolean expression supported by Spark SQL, using only partition columns.

3. Importing AWS GLUE DATA CATALOG 2, settings are as follows
    - Database- social-media-data
    - Table- etl_social_media



Name
AWS Glue Data Catalog

Database
Choose a database.
social-media-data

▶ Use runtime parameters

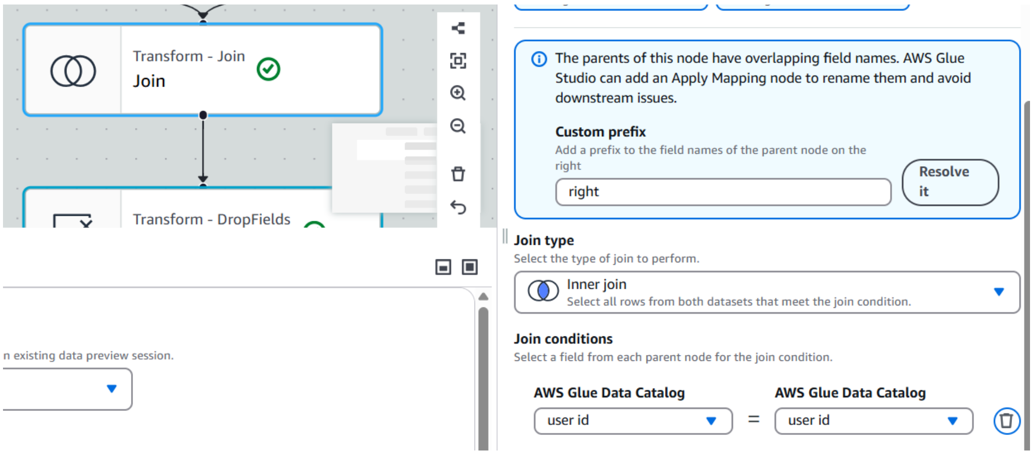
Table
etl_social_media

▶ Use runtime parameters

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)).

4. Importing Join from transforms and allotting parent nodes as AWS GLUE DATA CATALOG 1 and AWS GLUE DATA CATALOG 2 settings are as follows-
    - Custom prefix right
    - Join type – inner join
    - Join conditions- User id= User id



The parents of this node have overlapping field names. AWS Glue Studio can add an Apply Mapping node to rename them and avoid downstream issues.

Custom prefix
Add a prefix to the field names of the parent node on the right
right

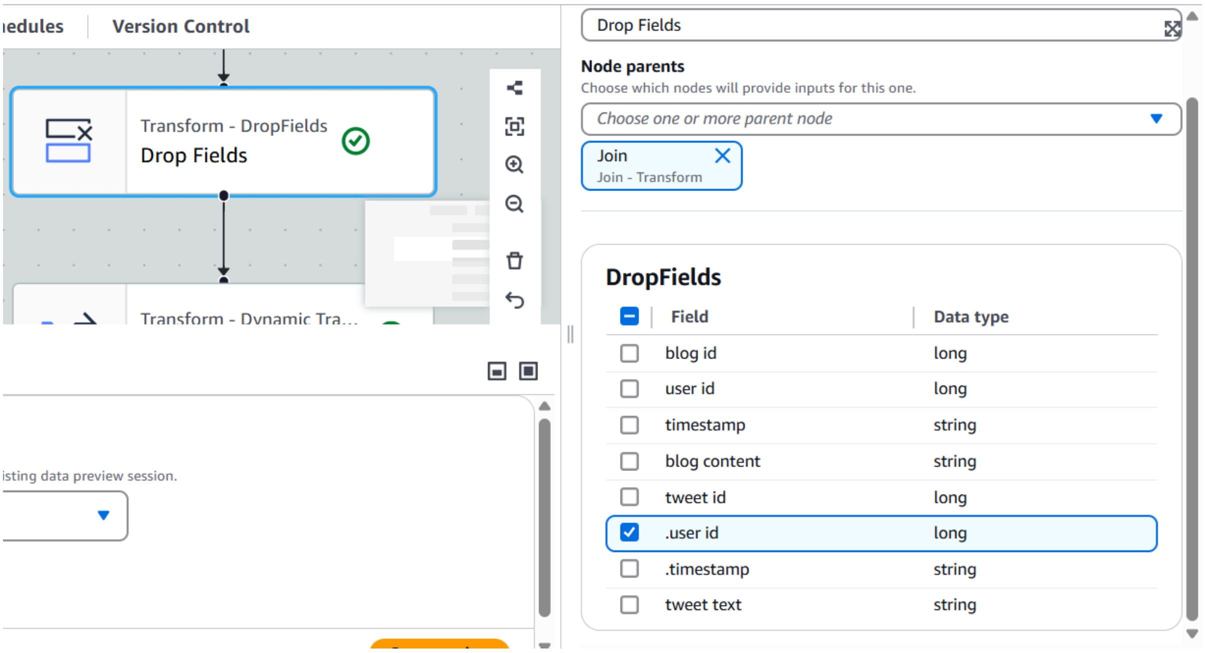
Resolve it

Join type
Select the type of join to perform.
Inner join
Select all rows from both datasets that meet the join condition.

Join conditions
Select a field from each parent node for the join condition.

AWS Glue Data Catalog		AWS Glue Data Catalog
user id	=	user id

5. Importing Drop Fields from transforms where parent nodes are Join and its settings are as follows-
    - Drop fields as .userid



Drop Fields

Node parents
Choose which nodes will provide inputs for this one.

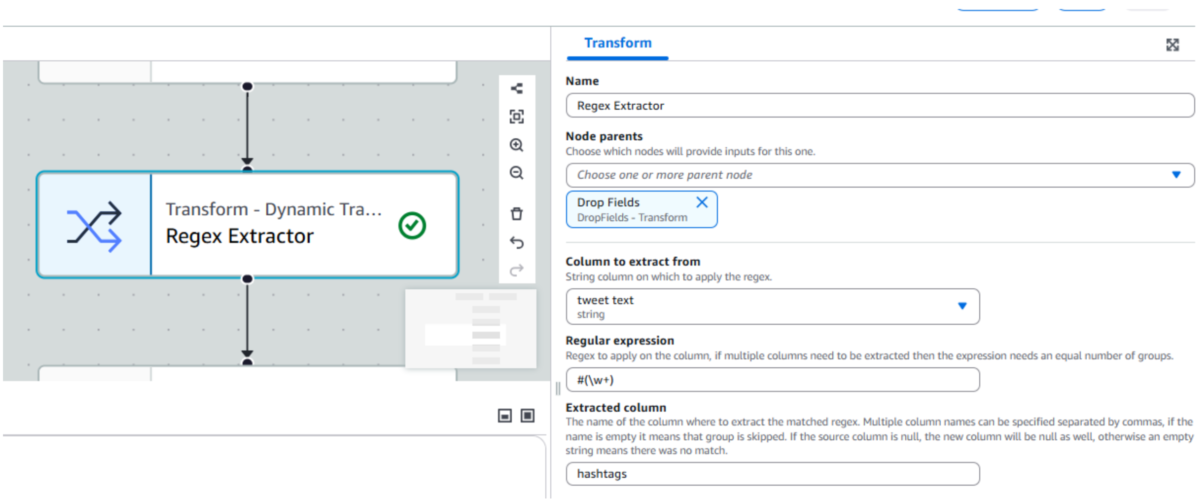
Choose one or more parent node

Join
Join - Transform

DropFields

	Field	Data type
☐	blog id	long
☐	user id	long
☐	timestamp	string
☐	blog content	string
☐	tweet id	long
☑	.user id	long
☐	.timestamp	string
☐	tweet text	string

6. Importing Regex Extractor from Transforms choosing node parents as Drop Fields. Settings are as follows-
    - Column to extract from – tweet text
    - Regular expression #\(w+)
    - Extracted column as hashtags



Name
Regex Extractor

Node parents
Choose which nodes will provide inputs for this one.

Choose one or more parent node

Drop Fields
DropFields - Transform

Column to extract from
String column on which to apply the regex.

tweet text
string

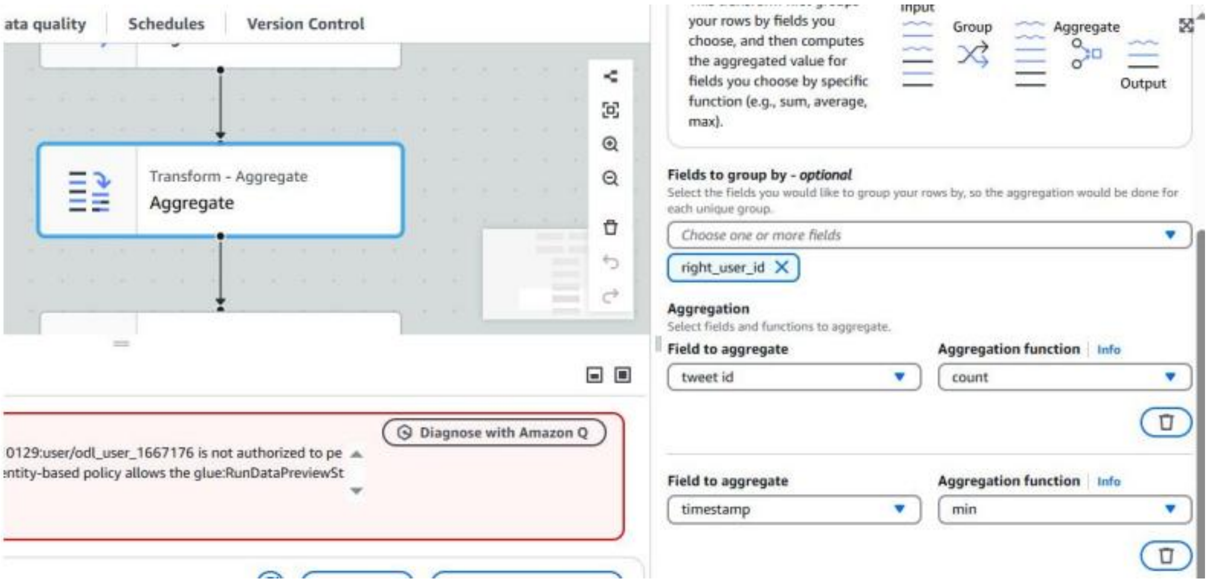
Regular expression
Regex to apply on the column, if multiple columns need to be extracted then the expression needs an equal number of groups.

#(\w+)

Extracted column
The name of the column where to extract the matched regex. Multiple column names can be specified separated by commas, if the name is empty it means that group is skipped. If the source column is null, the new column will be null as well, otherwise an empty string means there was no match.

hashtags

7. Importing Aggregate from Transforms the settings are as follows-
    - Fields to group by as right_user id
    - Field to aggregate as tweet id and Aggregation function as count
    - In Aggregate a column,
      Field to aggregate as timestamp and Aggregation function as min



0129:user/odl_user_1667176 is not authorized to pe
entity-based policy allows the glue:RunDataPreviewSt

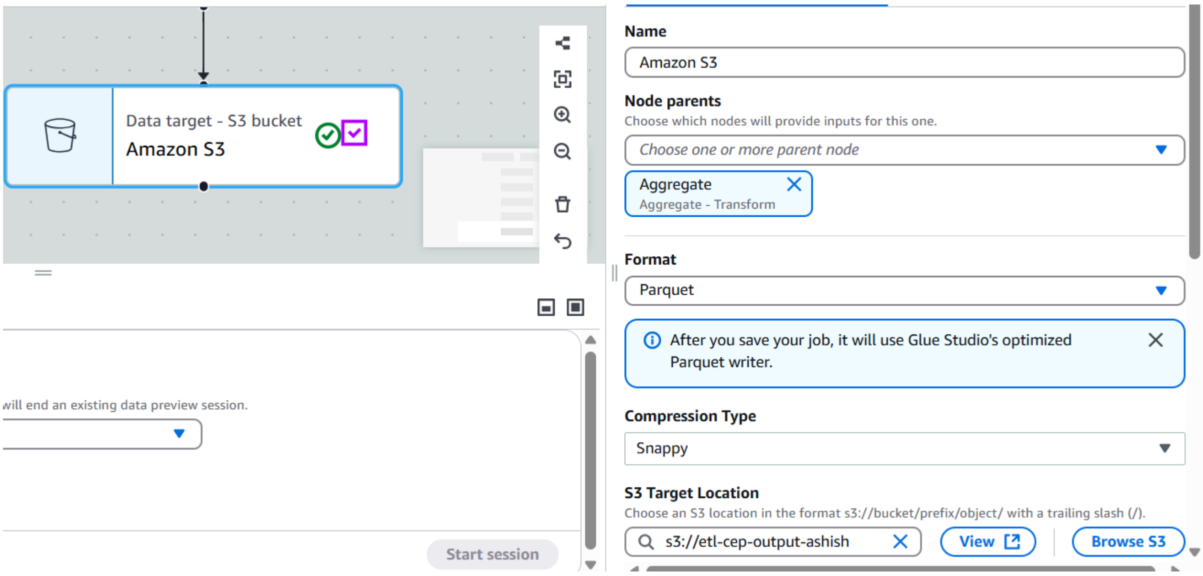
your rows by fields you choose, and then computes the aggregated value for fields you choose by specific function (e.g., sum, average, max).

Fields to group by - *optional*
Select the fields you would like to group your rows by, so the aggregation would be done for each unique group.
Choose one or more fields
right_user_id

Aggregation
Select fields and functions to aggregate.

Field to aggregate	Aggregation function Info
tweet id	count
timestamp	min

8. Importing Amazon S3 from Data Target node parent as Aggregate.
    - S3 target location as – s3://etl-cep-output-ashish



will end an existing data preview session.

Name
Amazon S3

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Aggregate
Aggregate - Transform

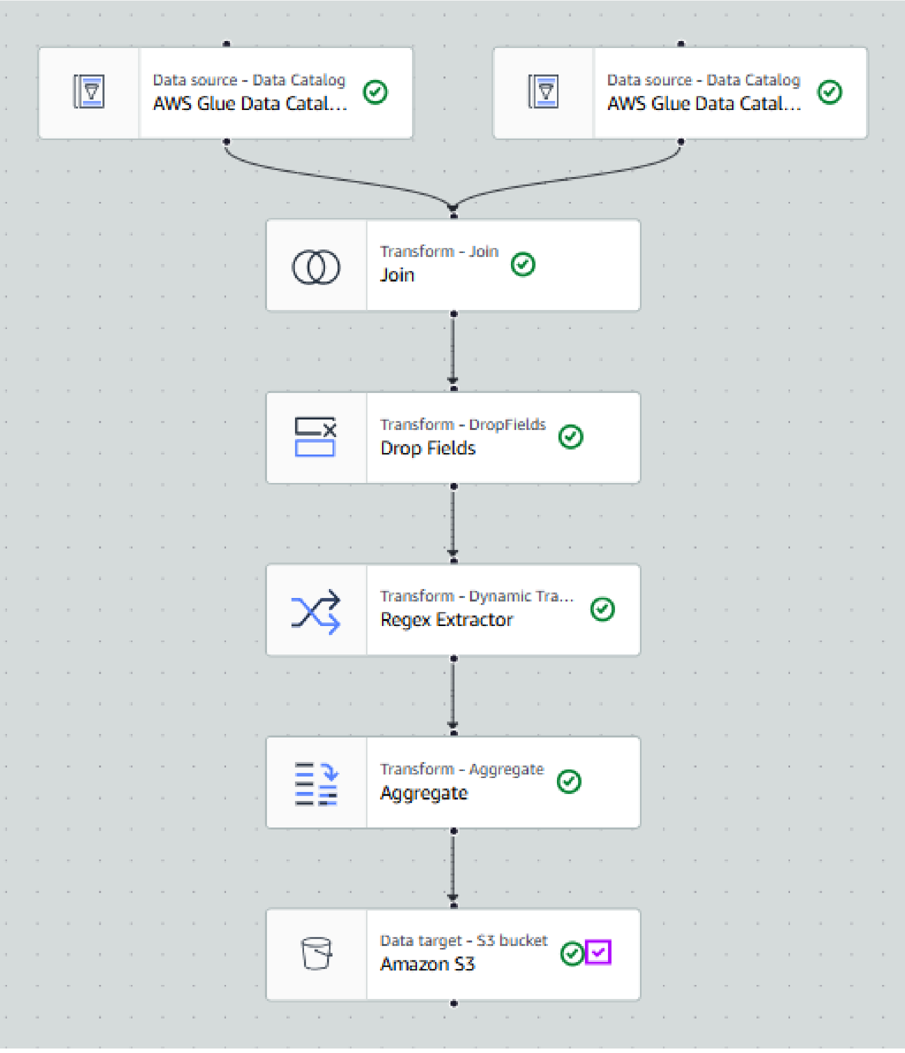
Format
Parquet

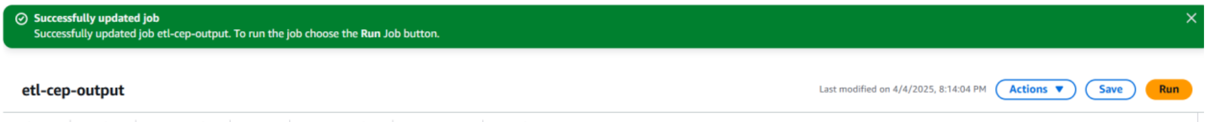
After you save your job, it will use Glue Studio's optimized Parquet writer.

Compression Type
Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://etl-cep-output-ashish
View
Browse S3

9. Final outlook after the job creation
    - Job name- etl-cep-output

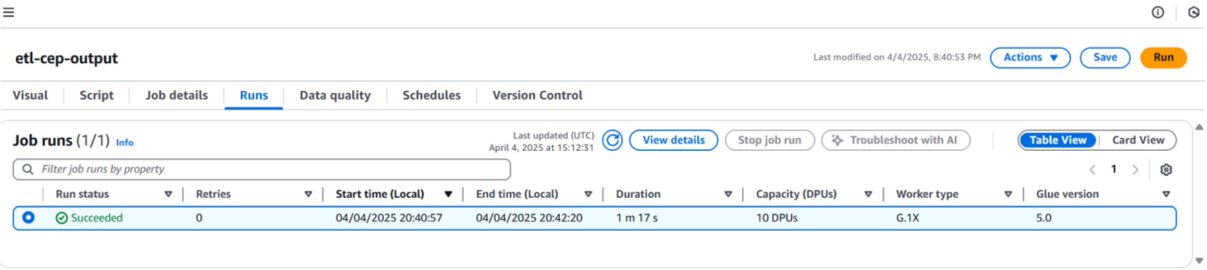



Successfully updated job
Successfully updated job etl-cep-output. To run the job choose the **Run** Job button.

etl-cep-output

Last modified on 4/4/2025, 8:14:04 PM

10. Job ran successfully


etl-cep-output

Last modified on 4/4/2025, 8:40:53 PM

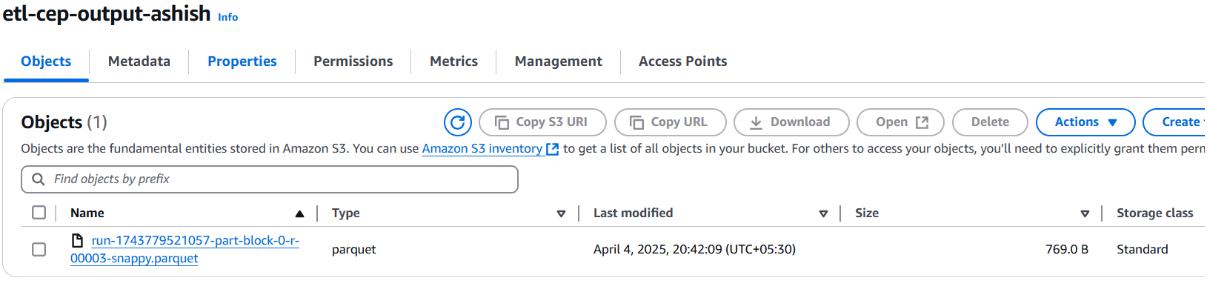
Job runs (1/1) Info

Last updated (UTC)
April 4, 2025 at 15:12:31

	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
●	✓ Succeeded	0	04/04/2025 20:40:57	04/04/2025 20:42:20	1 m 17 s	10 DPUs	G.1X	5.0

Step-9

11. The output folder now has a file that is generated from the ETL job I performed



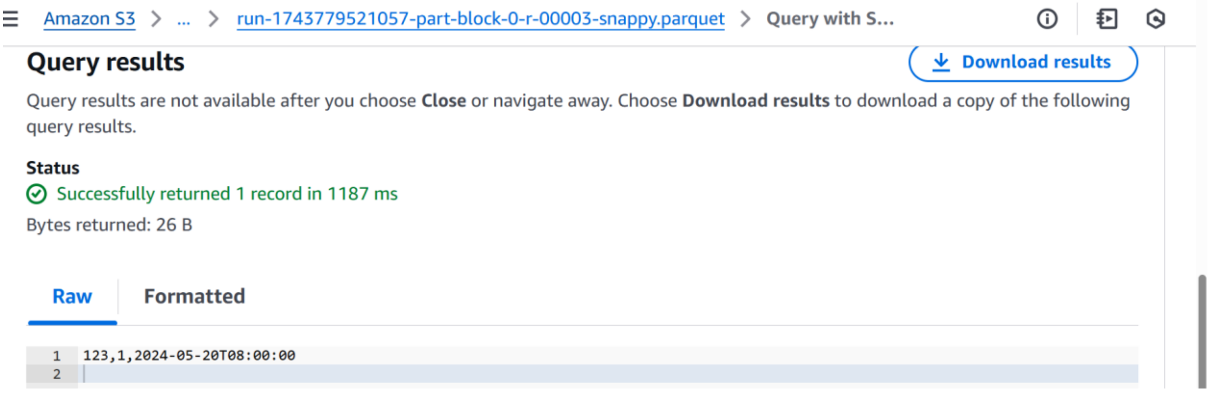
etl-cep-output-ashish Info

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 pern

□	Name	Type	Last modified	Size	Storage class
□	run-1743779521057-part-block-0-r-00003-snappy.parquet	parquet	April 4, 2025, 20:42:09 (UTC+05:30)	769.0 B	Standard

12. Running the SQL query to obtain the final result



Query results

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

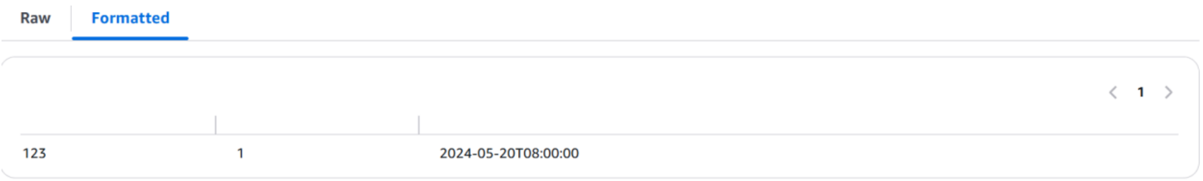
Status

✓ Successfully returned 1 record in 1187 ms

Bytes returned: 26 B

```
123,1,2024-05-20T08:00:00

```



123	1	2024-05-20T08:00:00

User Id	Blog ID	Timestamp
123	1	2024-05-20T08:00:00