Analyzing Retail Sales of Electricity

Milestone 6

Group 15

Manaswini Kamtam

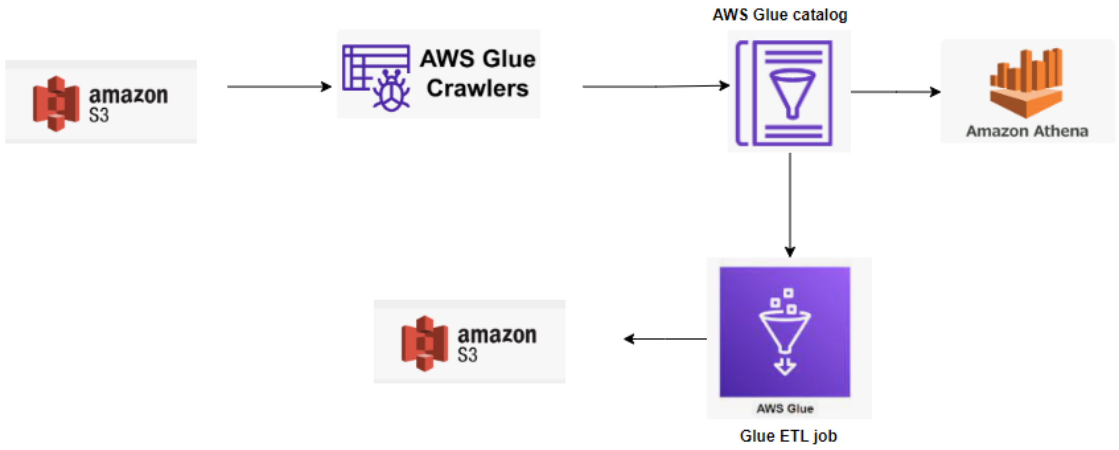
Aditya Bharadwaj Shivapura Guruprasad

kamtam.m@northeastern.edu

shivapuraguruprasa.a@northeastern.edu

Submission Date:________12/01/2023__________________________________

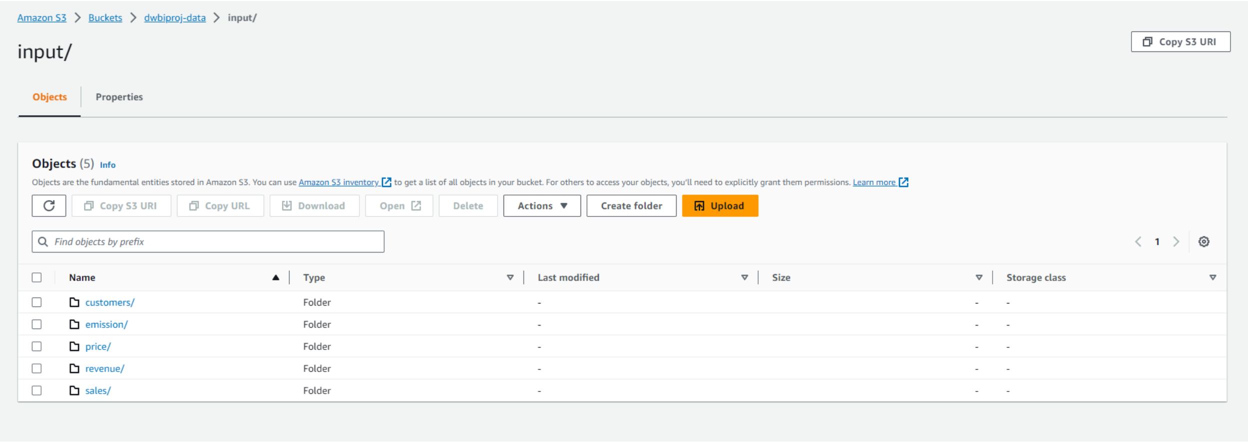
Architecture Diagram of the Data Pipeline



Steps involved in building the pipeline:

1. Loaded data in S3 bucket

At first, we loaded the csv data to the S3 bucket by manual upload. Separate folders are created for each file



input/

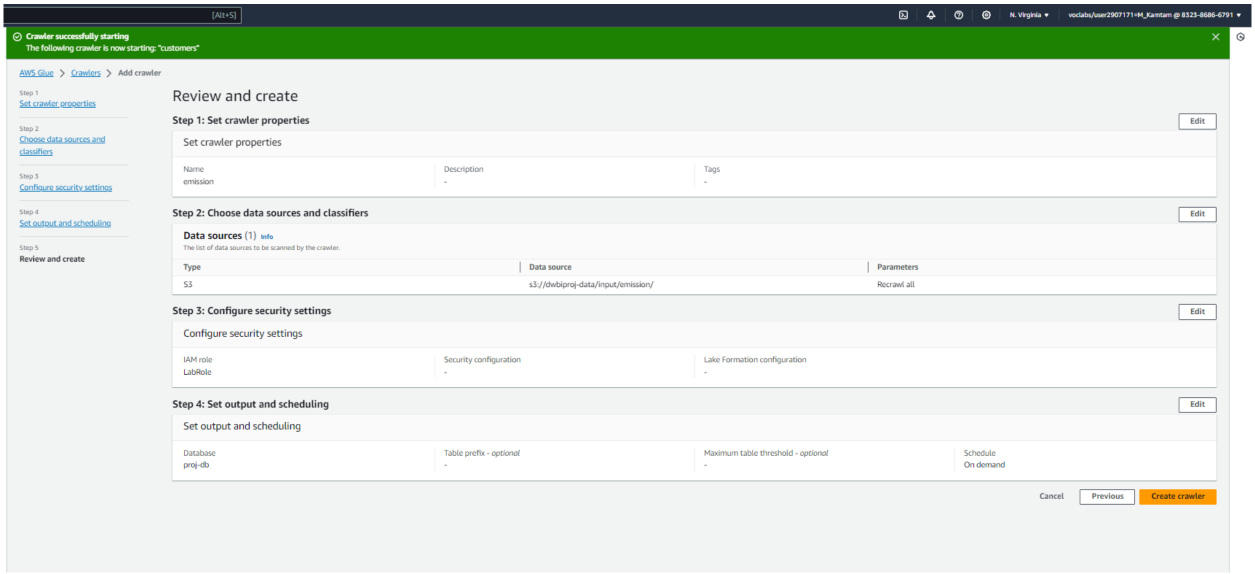
Objects (5) Info

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
□	customers/	Folder	-	-	-
□	emission/	Folder	-	-	-
□	price/	Folder	-	-	-
□	revenue/	Folder	-	-	-
□	sales/	Folder	-	-	-

2. Created Glue crawlers for each data table

Then AWS Glue crawler is created for each of the csv file and a new database is created in the catalog to store the data



[Alt+S]
N. Virginia ▾
voclabs/user2907171=M_Kamtam @ 8323-8686-6791 ▾

Crawler successfully starting
The following crawler is now starting: "customers"

AWS Glue > Crawlers > Add crawler

Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create

Review and create

Step 1: Set crawler properties

Set crawler properties

Name	Description	Tags
emission	-	-

Step 2: Choose data sources and classifiers

Data sources (1) Info
The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://dwbiproj-data/input/emission/	Recrawl all

Step 3: Configure security settings

Configure security settings

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

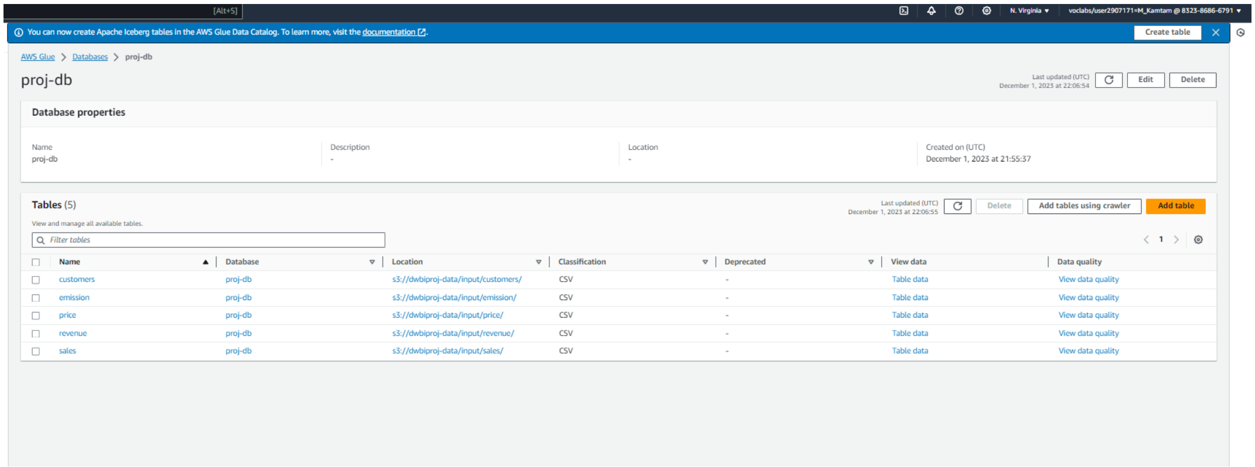
Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - *optional*	Maximum table threshold - *optional*	Schedule
proj-db	-	-	On demand

3. Ran the crawler and loaded the data in Glue catalog

Crawler is run to read the schema of the data present in the s3 bucket and load the data in the catalog DB



You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the documentation.

AWS Glue > Databases > proj-db

proj-db

Last updated (UTC)
December 1, 2023 at 22:06:54

Database properties

Name	Description	Location	Created on (UTC)
proj-db	-	-	December 1, 2023 at 21:55:37

Tables (5)

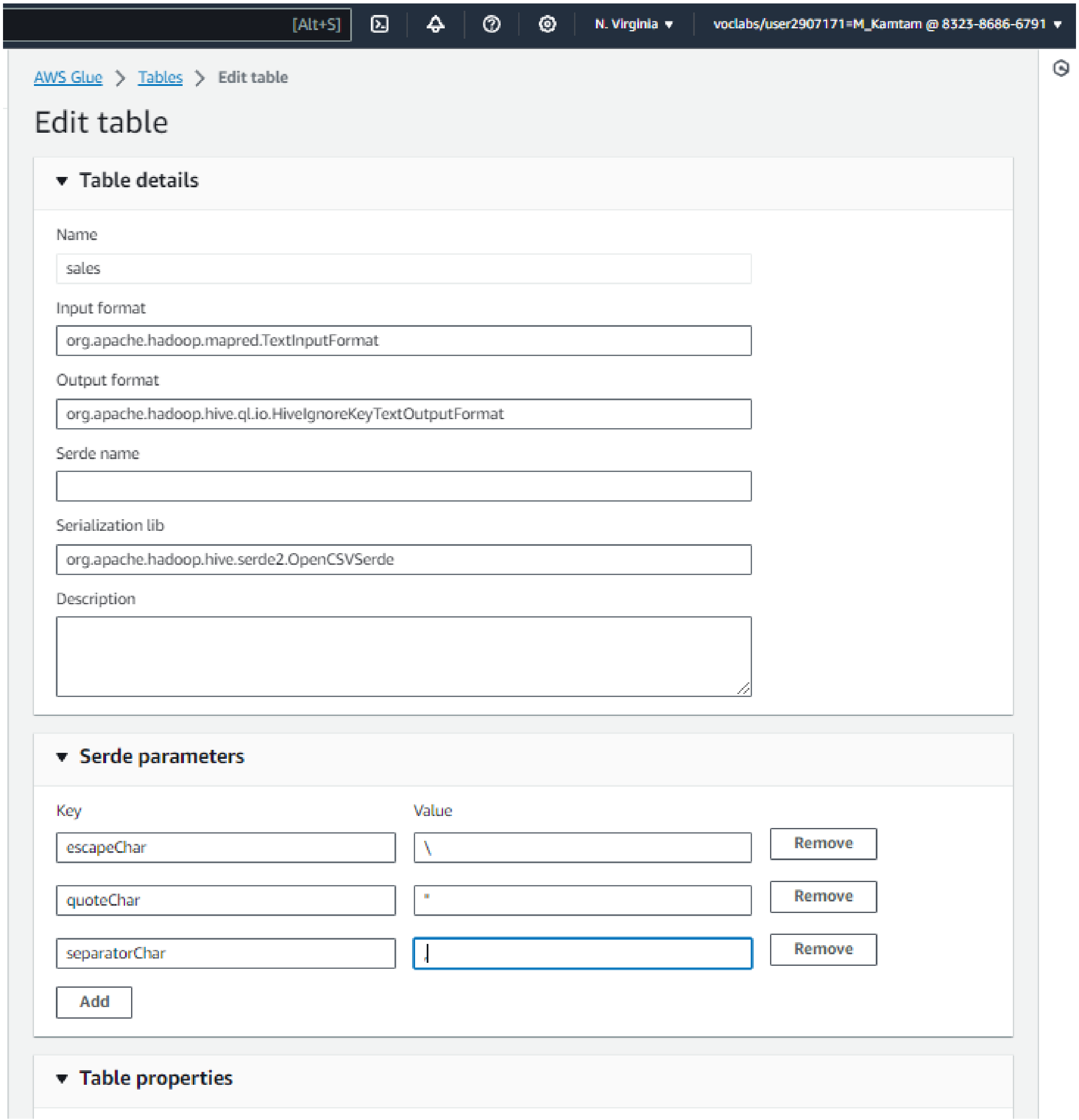
View and manage all available tables.

Last updated (UTC)
December 1, 2023 at 22:06:55

Name	Database	Location	Classification	Deprecated	View data	Data quality
customers	proj-db	s3://dwbiproj-data/input/customers/	CSV	-	Table data	View data quality
emission	proj-db	s3://dwbiproj-data/input/emission/	CSV	-	Table data	View data quality
price	proj-db	s3://dwbiproj-data/input/price/	CSV	-	Table data	View data quality
revenue	proj-db	s3://dwbiproj-data/input/revenue/	CSV	-	Table data	View data quality
sales	proj-db	s3://dwbiproj-data/input/sales/	CSV	-	Table data	View data quality

Made the following change to table schema since input data was hasquotes in the csv file

Data shifting was happening without the below configuration. Delimeters were introduced in the below steps


[Alt+S]
N. Virginia ▾
voclabs/user2907171=M_Kamtam @ 8323-8686-6791 ▾
AWS Glue > Tables > Edit table

Edit table

▾ Table details

Name: sales

Input format: org.apache.hadoop.mapred.TextInputFormat

Output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

Serde name:

Serialization lib: org.apache.hadoop.hive.serde2.OpenCSVSerde

Description:

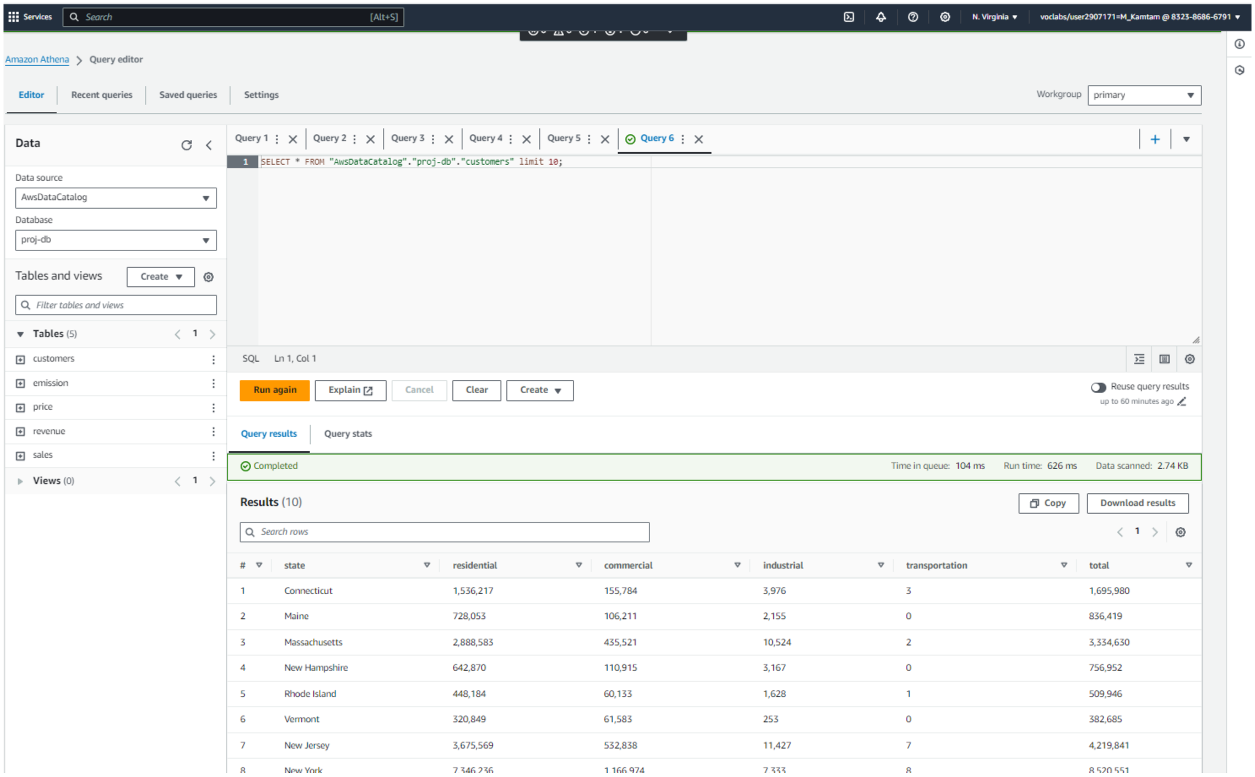
▾ Serde parameters

Key	Value	
escapeChar	\	Remove
quoteChar	"	Remove
separatorChar	,	Remove

Add

▾ Table properties

4. The data is queried using Aws Athena to read the data from the catalog DB



Workgroup: primary

Data source: AwsDataCatalog

Database: proj-db

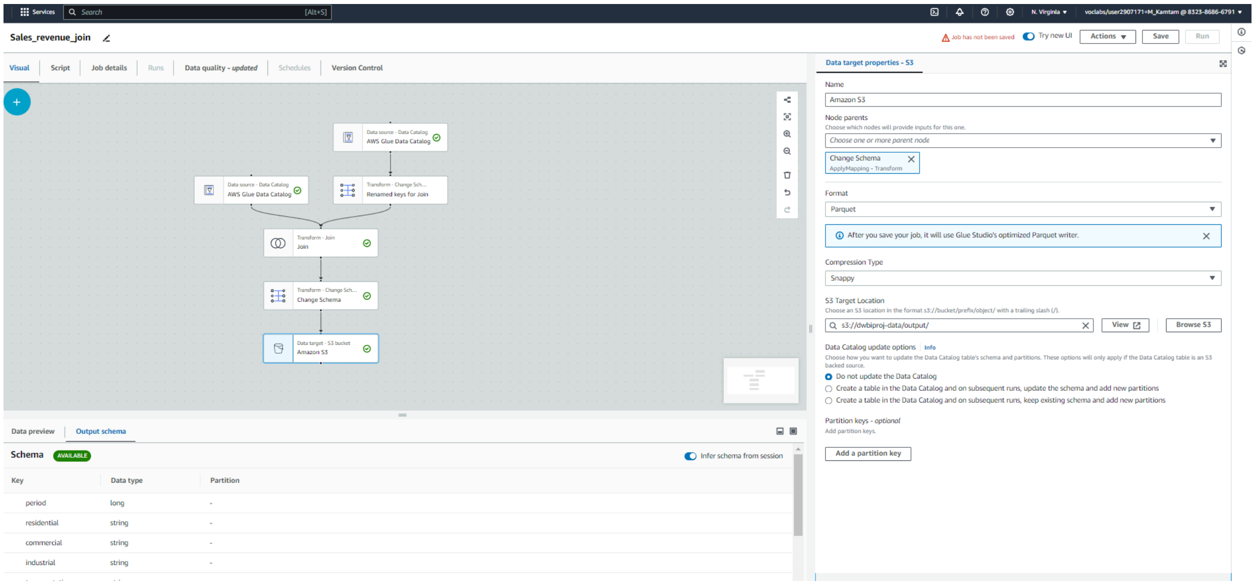
```
SELECT * FROM "AwsDataCatalog"."proj-db"."customers" limit 10;
```

Completed — Time in queue: 104 ms — Run time: 626 ms — Data scanned: 2.74 KB

Results (10)

#	state	residential	commercial	industrial	transportation	total
1	Connecticut	1,536,217	155,784	3,976	3	1,695,980
2	Maine	728,053	106,211	2,155	0	836,419
3	Massachusetts	2,888,583	435,521	10,524	2	3,334,630
4	New Hampshire	642,870	110,915	3,167	0	756,952
5	Rhode Island	448,184	60,133	1,628	1	509,946
6	Vermont	320,849	61,583	253	0	382,685
7	New Jersey	3,675,569	532,838	11,427	7	4,219,841
8	New York	7,346,236	1,166,974	7,333	8	8,520,551

5. Created a new Glue ETL visual based job with the below configuration



Sales_revenue_join

Data target properties - S3

Name: Amazon S3

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

Choose one or more parent node

Change Schema — ApplyMapping - 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://dwbiproj-data/output/

Data Catalog update options

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

- ● Do not update the Data Catalog
- ○ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
- ○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

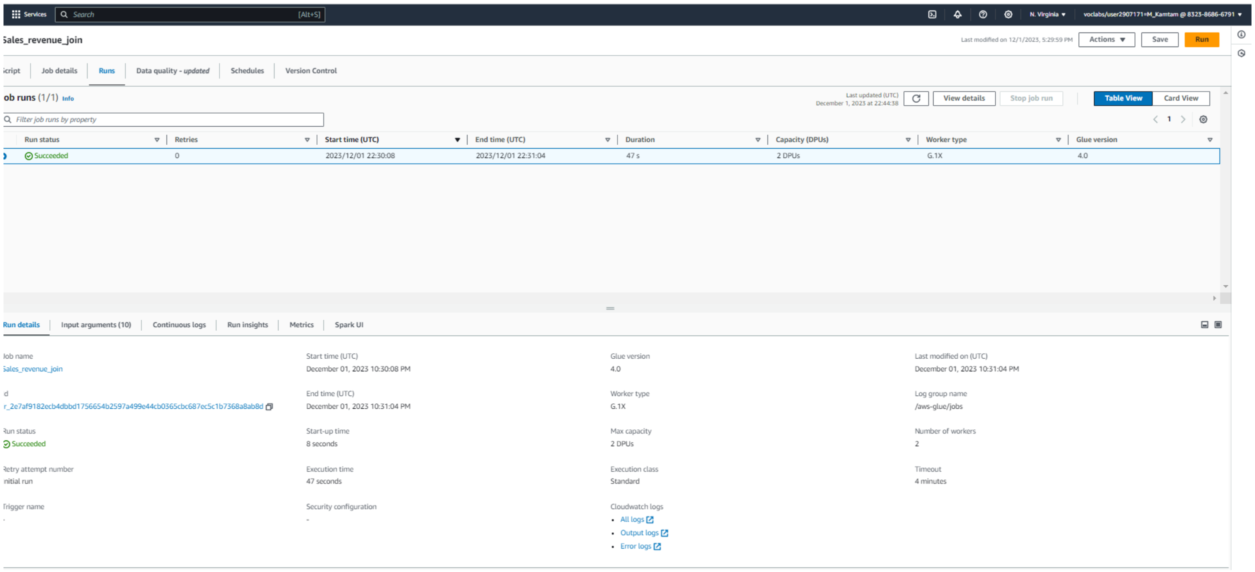
Partition keys - *optional*

Add partition keys.

Schema AVAILABLE

Key	Data type	Partition
period	long	-
residential	string	-
commercial	string	-
industrial	string	-

6. Joining two input tables (sales and revenue) and loading data to S3 bucket - as a csv



Sales_revenue_join

Last modified on 12/1/2023, 5:29:59 PM

Job runs (1/1)

Last updated (UTC)
December 1, 2023 at 22:44:38

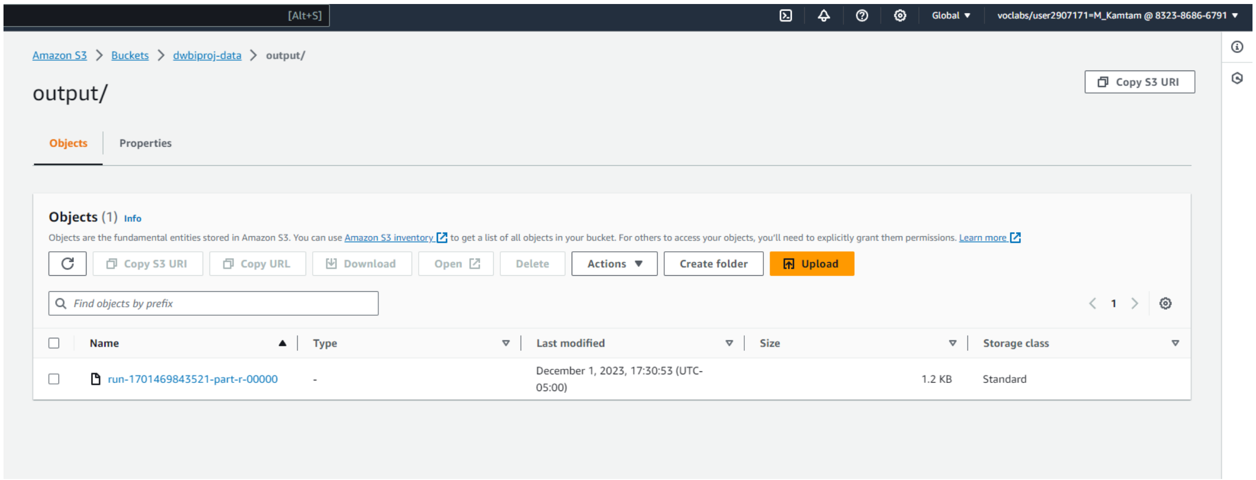
Run status	Retries	Start time (UTC)	End time (UTC)	Duration	Capacity (DPUs)	Worker type	Glue version
✓ Succeeded	0	2023/12/01 22:30:08	2023/12/01 22:31:04	47 s	2 DPUs	G.1X	4.0

Run details

Job name Sales_revenue_join	Start time (UTC) December 01, 2023 10:30:08 PM	Glue version 4.0	Last modified on (UTC) December 01, 2023 10:31:04 PM
Id jr_2e7af9182ecb4dbbd1756654b2597a499e44cb0365cbc687ec5c1b7368a8ab8d	End time (UTC) December 01, 2023 10:31:04 PM	Worker type G.1X	Log group name /aws-glue/jobs
Run status ✓ Succeeded	Start-up time 8 seconds	Max capacity 2 DPUs	Number of workers 2
Retry attempt number Initial run	Execution time 47 seconds	Execution class Standard	Timeout 4 minutes
Trigger name -	Security configuration -	Cloudwatch logs • All logs • Output logs • Error logs	

Job by default writes to multiple csv parallely. This feature is disabled by updating the job script and only one csv is obtained

7. ETL job output as seen in s3 bucket



Amazon S3 > Buckets > dwbiproj-data > output/

output/

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

☐	Name	Type	Last modified	Size	Storage class
☐	run-1701469843521-part-r-00000	-	December 1, 2023, 17:30:53 (UTC-05:00)	1.2 KB	Standard