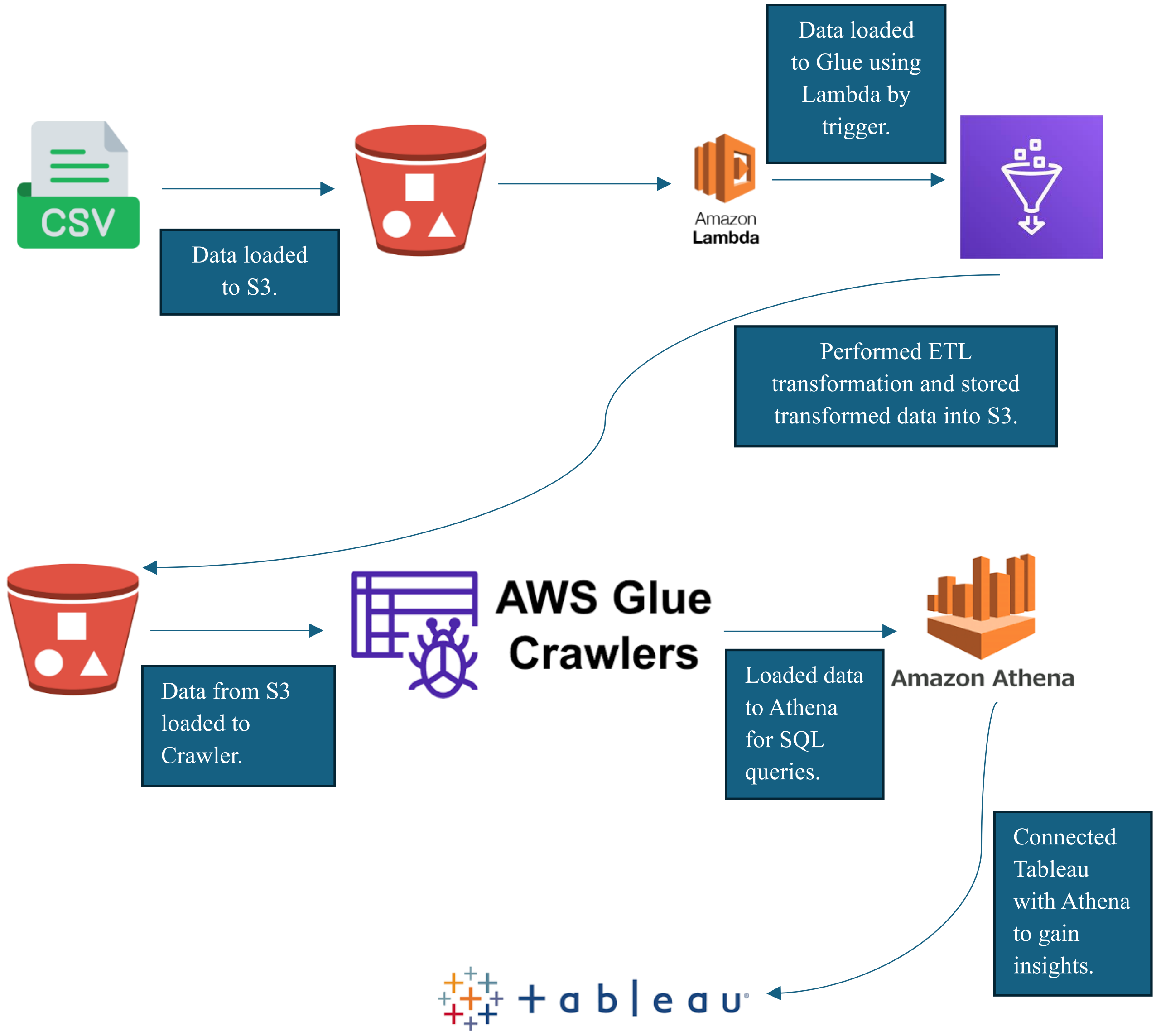
Milestone 6

Data Architecture



Data loaded to S3.

Data loaded to Glue using Lambda by trigger.

Performed ETL transformation and stored transformed data into S3.

Data from S3 loaded to Crawler.

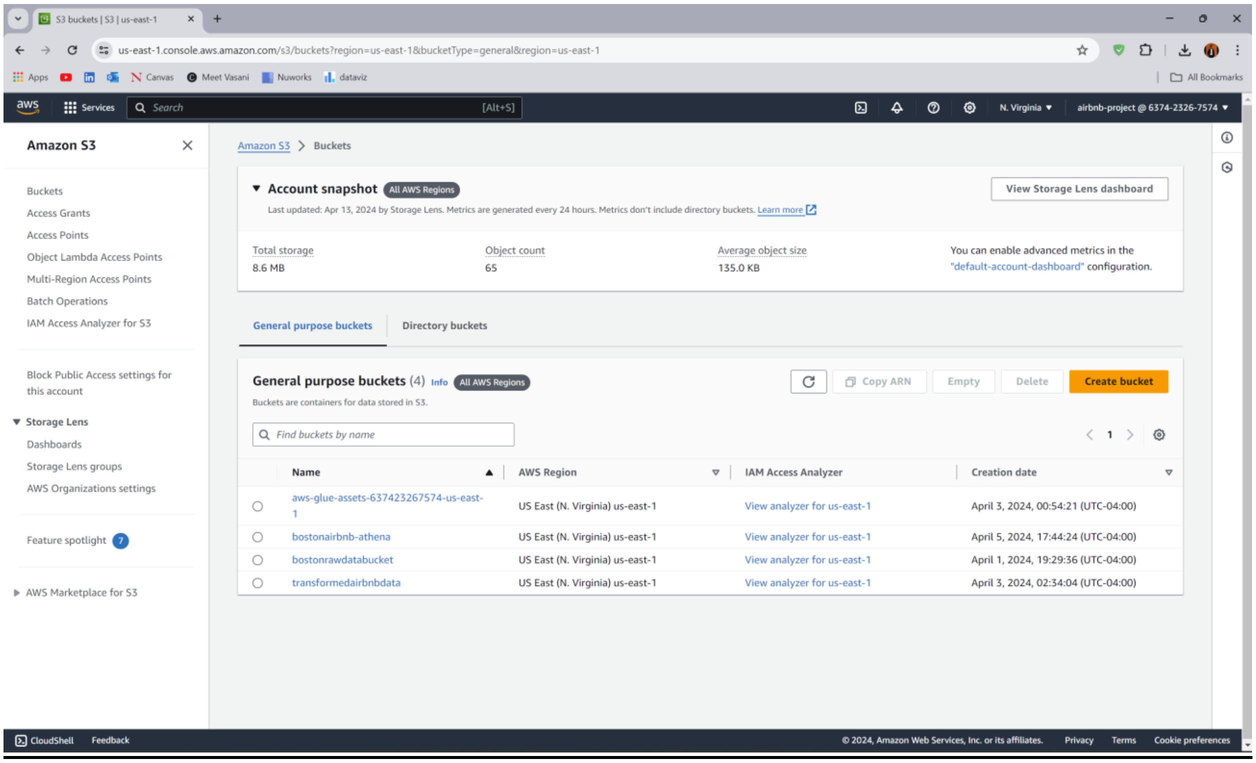
Loaded data to Athena for SQL queries.

Connected Tableau with Athena to gain insights.

Data Ingestion

- Prepared structured data in CSV format.
- Created an S3 bucket, load csv file to S3.
- Used AWS lambda to fetch data from S3 and pass it to AWS glue.
- Used AWS glue for ETL to transform data.
- After transformation, sent data to the new S3 bucket.
- Loaded the transformed data to AWS Athena by AWS Glue Crawler.
- Used Athena to run queries and get granular insights from the data.

- Connected Athena with Tableau to visually get some insights.


S3 buckets | S3 | us-east-1
us-east-1.console.aws.amazon.com/s3/buckets?region=us-east-1&bucketType=general&region=us-east-1
Services
Search
[Alt+S]
N. Virginia
airbnb-project @ 6374-2326-7574
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets
View Storage Lens dashboard
General purpose buckets
Directory buckets
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Account snapshot All AWS Regions

Last updated: Apr 13, 2024 by Storage Lens. Metrics are generated every 24 hours. Metrics don't include directory buckets. Learn more

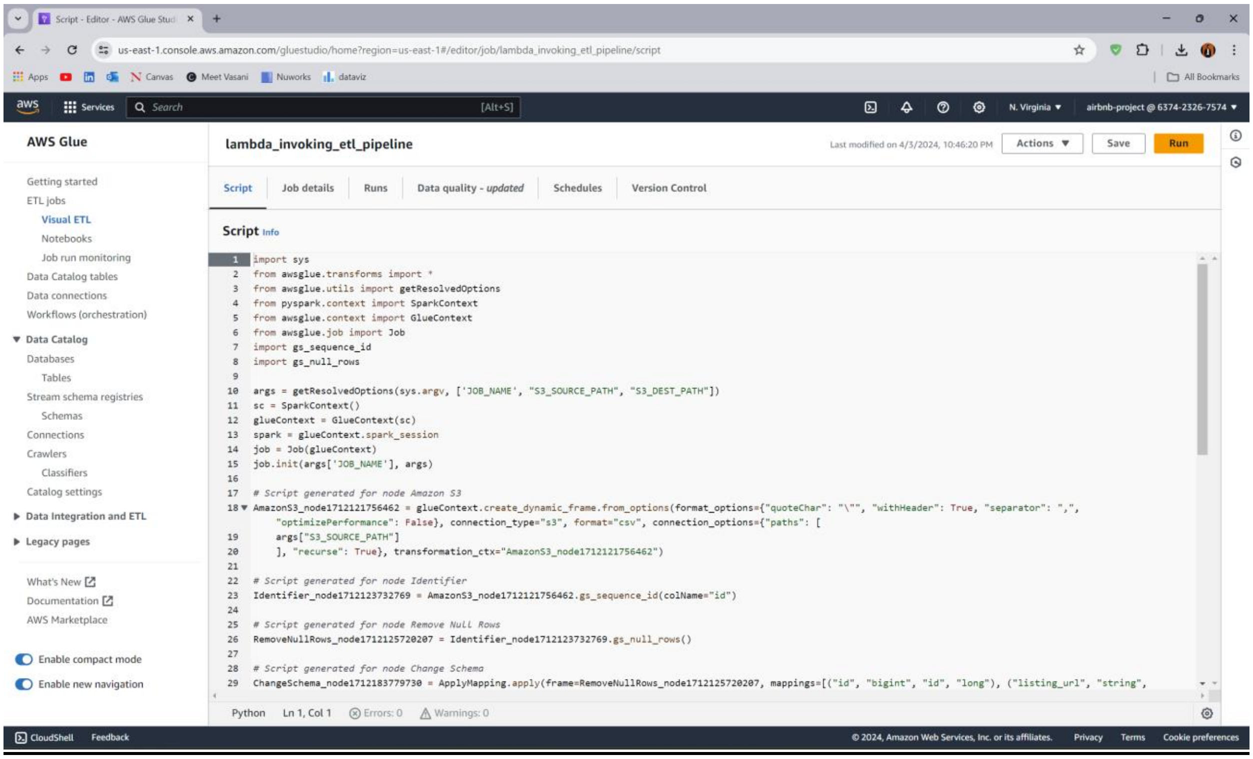
Total storage	Object count	Average object size	
8.6 MB	65	135.0 KB	You can enable advanced metrics in the "default-account-dashboard" configuration.

General purpose buckets (4) Info All AWS Regions

Buckets are containers for data stored in S3.

Find buckets by name

	Name	AWS Region	IAM Access Analyzer	Creation date
○	aws-glue-assets-637423267574-us-east-1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	April 3, 2024, 00:54:21 (UTC-04:00)
○	bostonairbnb-athena	US East (N. Virginia) us-east-1	View analyzer for us-east-1	April 5, 2024, 17:44:24 (UTC-04:00)
○	bostonrawdatabucket	US East (N. Virginia) us-east-1	View analyzer for us-east-1	April 1, 2024, 19:29:36 (UTC-04:00)
○	transformedairbnbdata	US East (N. Virginia) us-east-1	View analyzer for us-east-1	April 3, 2024, 02:34:04 (UTC-04:00)

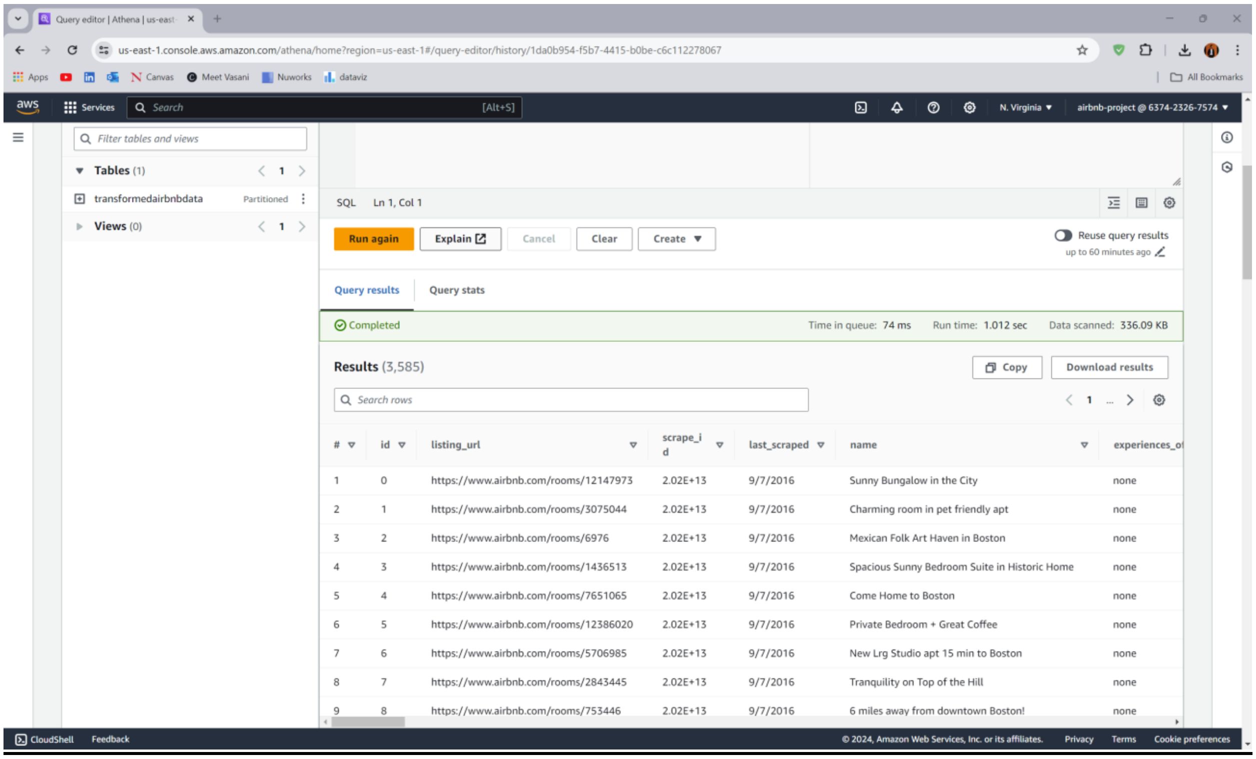


lambda_invoking_etl_pipeline

Last modified on 4/3/2024, 10:46:20 PM

Script Info

```python
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
import gs_sequence_id
import gs_null_rows

args = getResolvedOptions(sys.argv, ['JOB_NAME', "S3_SOURCE_PATH", "S3_DEST_PATH"])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)

# Script generated for node Amazon S3
AmazonS3_node1712121756462 = glueContext.create_dynamic_frame.from_options(format_options={"quoteChar": "\"", "withHeader": True, "separator": ",",
    "optimizePerformance": False}, connection_type="s3", format="csv", connection_options={"paths": [
    args["S3_SOURCE_PATH"]
    ], "recurse": True}, transformation_ctx="AmazonS3_node1712121756462")

# Script generated for node Identifier
Identifier_node1712123732769 = AmazonS3_node1712121756462.gs_sequence_id(colName="id")

# Script generated for node Remove Null Rows
RemoveNullRows_node1712125720207 = Identifier_node1712123732769.gs_null_rows()

# Script generated for node Change Schema
ChangeSchema_node1712183779730 = ApplyMapping.apply(frame=RemoveNullRows_node1712125720207, mappings=[("id", "bigint", "id", "long"), ("listing_url", "string",
```



Completed   Time in queue: 74 ms   Run time: 1.012 sec   Data scanned: 336.09 KB

Results (3,585)

#	id	listing_url	scrape_id	last_scraped	name	experiences_o
1	0	https://www.airbnb.com/rooms/12147973	2.02E+13	9/7/2016	Sunny Bungalow in the City	none
2	1	https://www.airbnb.com/rooms/3075044	2.02E+13	9/7/2016	Charming room in pet friendly apt	none
3	2	https://www.airbnb.com/rooms/6976	2.02E+13	9/7/2016	Mexican Folk Art Haven in Boston	none
4	3	https://www.airbnb.com/rooms/1436513	2.02E+13	9/7/2016	Spacious Sunny Bedroom Suite in Historic Home	none
5	4	https://www.airbnb.com/rooms/7651065	2.02E+13	9/7/2016	Come Home to Boston	none
6	5	https://www.airbnb.com/rooms/12386020	2.02E+13	9/7/2016	Private Bedroom + Great Coffee	none
7	6	https://www.airbnb.com/rooms/5706985	2.02E+13	9/7/2016	New Lrg Studio apt 15 min to Boston	none
8	7	https://www.airbnb.com/rooms/2843445	2.02E+13	9/7/2016	Tranquility on Top of the Hill	none
9	8	https://www.airbnb.com/rooms/753446	2.02E+13	9/7/2016	6 miles away from downtown Boston!	none