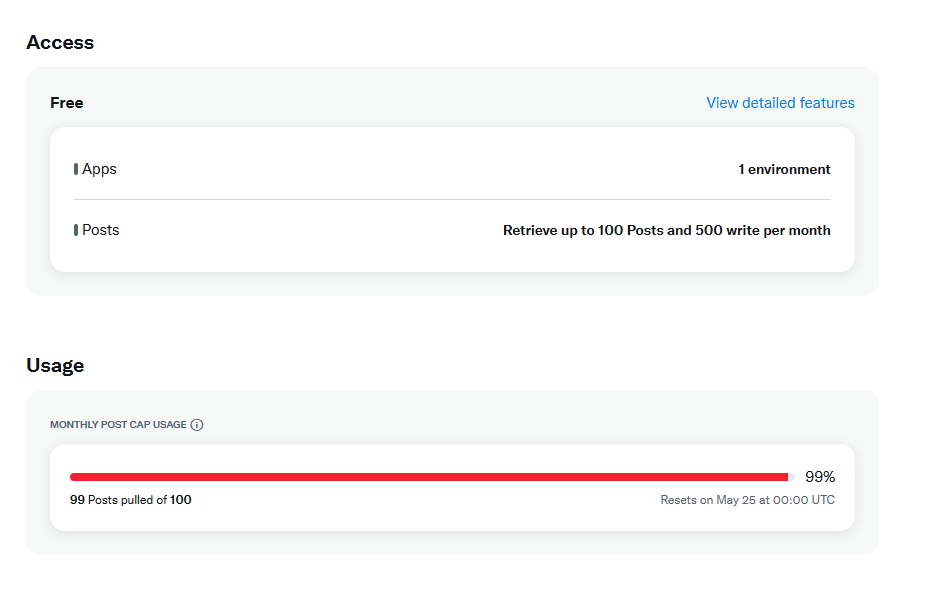
**Real-Time News Ingestion Pipeline using AWS Kinesis Firehose and Lambda**

**Overview**

This document outlines the design and implementation of a real-time news ingestion pipeline using Amazon Kinesis Data Firehose, AWS Lambda, and Amazon S3. The pipeline fetches live news articles related to specific topics using [NewsAPI](https://newsapi.org/account) and delivers the structured data directly into an S3 bucket for storage and downstream analytics.

**Objective**

Due to the exhaustion of Twitter API quota (As I have used twitter API for project work before the beginning of these tasks), the objective was redefined to build a scalable serverless pipeline that ingests real-time news articles based on specified keywords, stores them in an S3 bucket, and can later be integrated with analytics services such as AWS Glue, Amazon Athena, or Amazon QuickSight.



**Architecture Overview**

The final architecture consisted of the following components:

* **NewsAPI:** A free third-party API that provides real-time news data from global publishers.
* **AWS Lambda:** A serverless function that fetches news from NewsAPI and pushes the data to Kinesis Firehose.
* **Amazon Kinesis Data Firehose:** A managed service that delivers streaming data directly to destinations, in this case, Amazon S3.
* **Amazon S3:** Serves as the durable and scalable storage for raw ingested news data.

**Process Summary**

**API Key Acquisition**

* Obtained a free NewsAPI key to access the /everything endpoint, enabling keyword-based queries for recent news articles.

**Lambda Function Development**

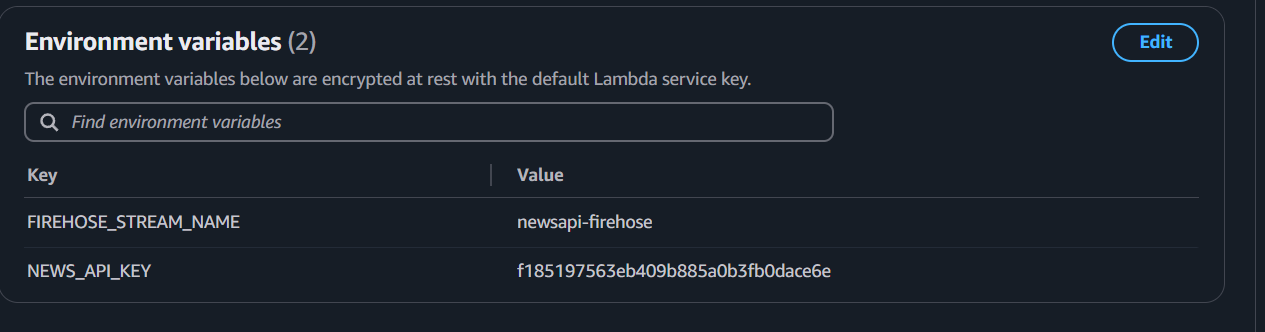
Wrote a Python-based Lambda function that:

* Constructs a NewsAPI request using urllib (avoiding external dependencies like requests).
* Authenticates using an API key from environment variables.
* Parses and formats the response payload (title, description, URL, timestamp, and source).
* Streams the formatted records to a specified Firehose delivery stream.

**Environment Variables Configuration**

Defined the following Lambda environment variables:

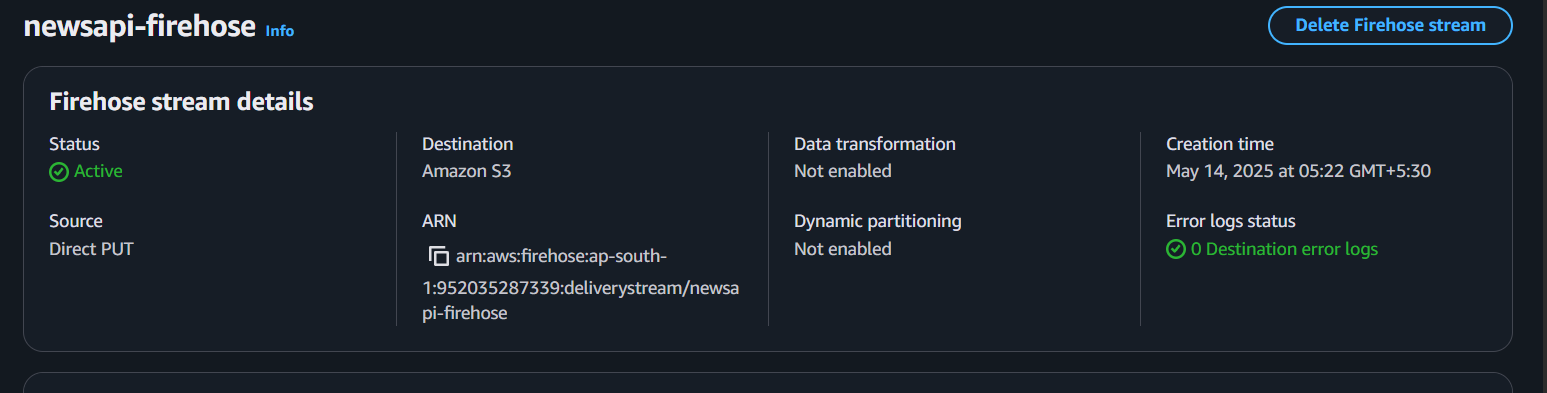
* **NEWS\_API\_KEY:** Stores the API key securely.
* **FIREHOSE\_STREAM\_NAME:** Identifies the target Firehose stream.



**Kinesis Firehose Setup**

Configured a Firehose delivery stream to:

* Receive records from the Lambda function.
* Deliver them in batch format to a designated Amazon S3 bucket.



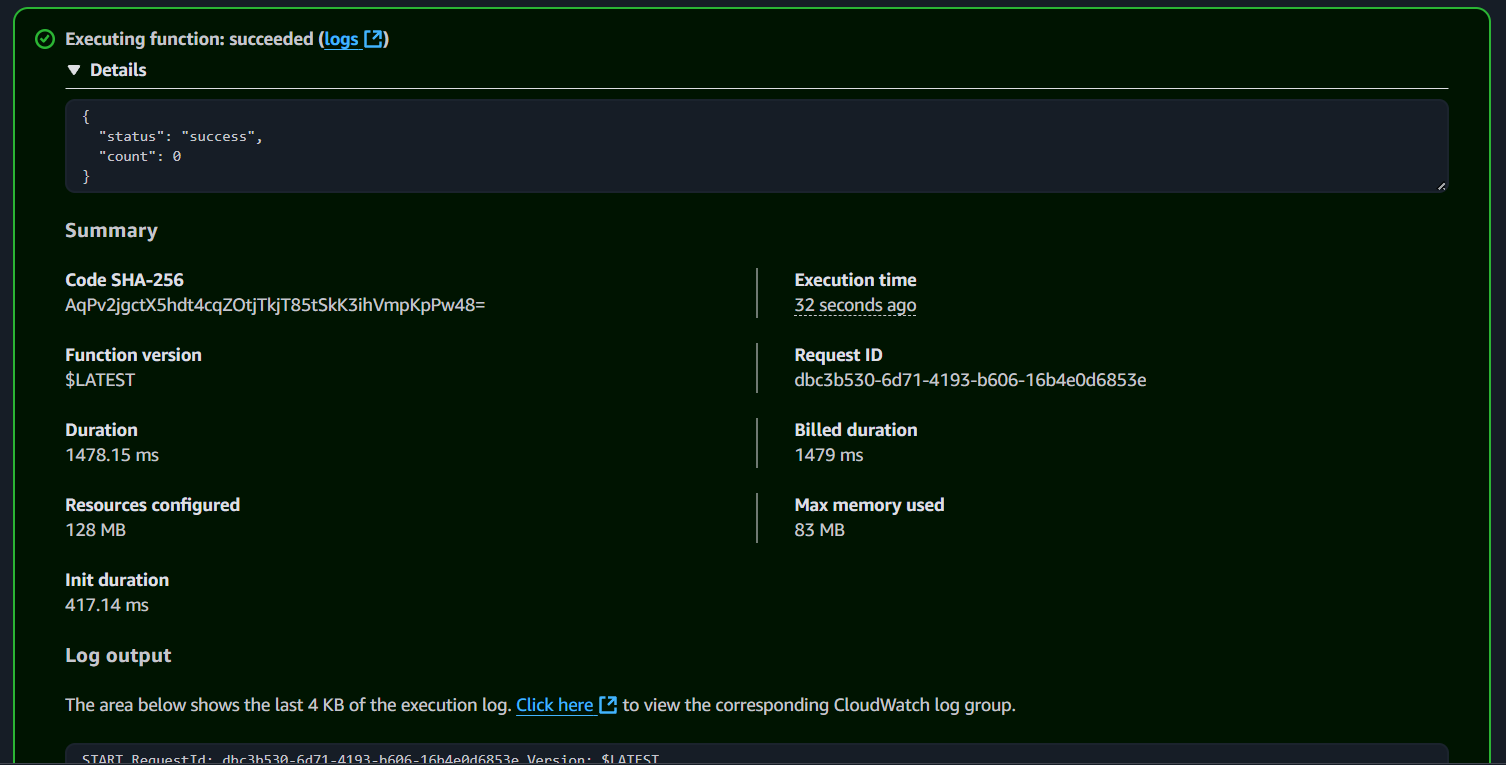
**Permission and IAM Setup**

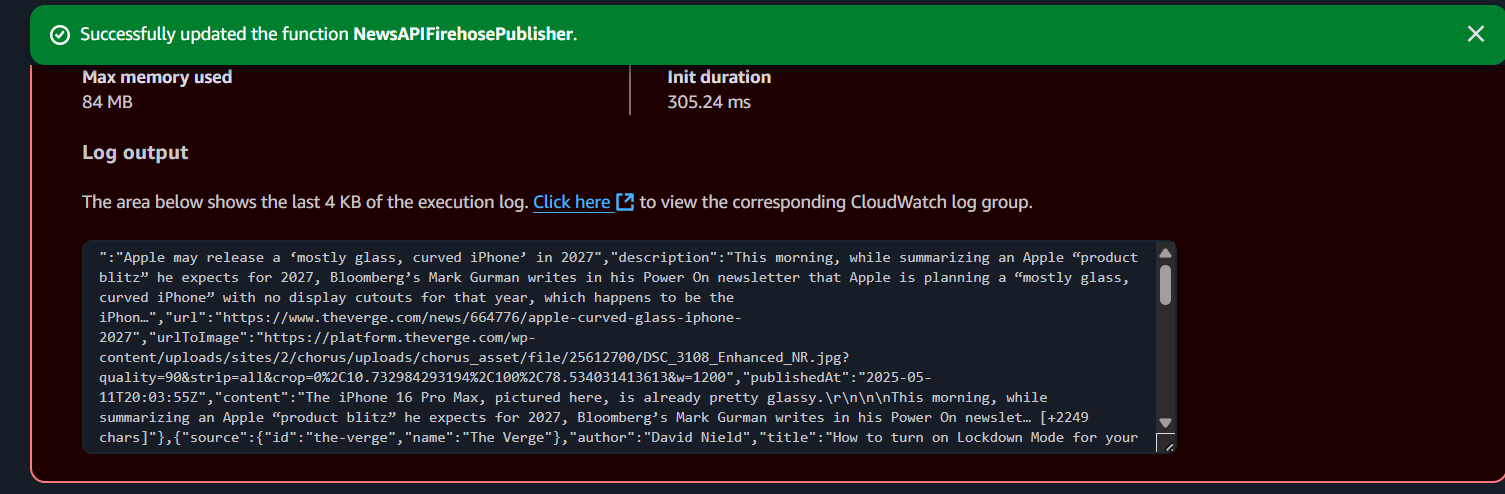
Ensured the Lambda function’s execution role had the necessary permissions:

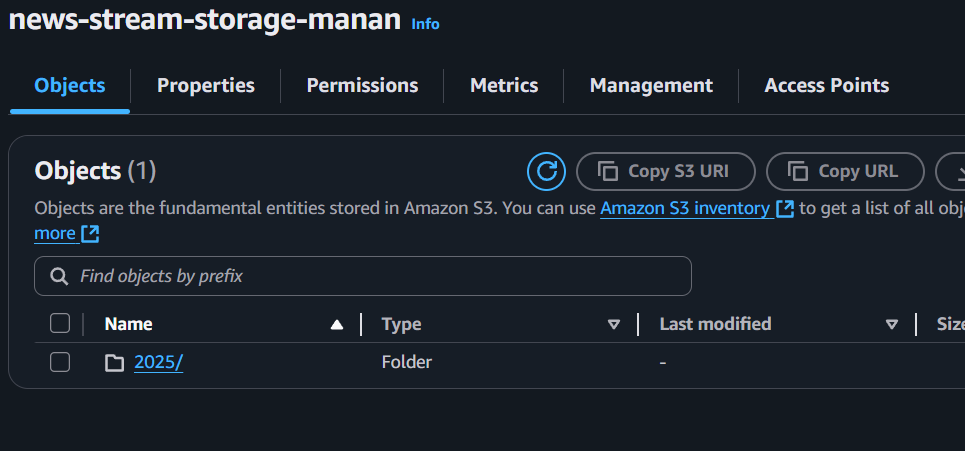
* firehose:PutRecord on the delivery stream
* Logging permissions (logs:CreateLogGroup, logs:CreateLogStream, logs:PutLogEvents)

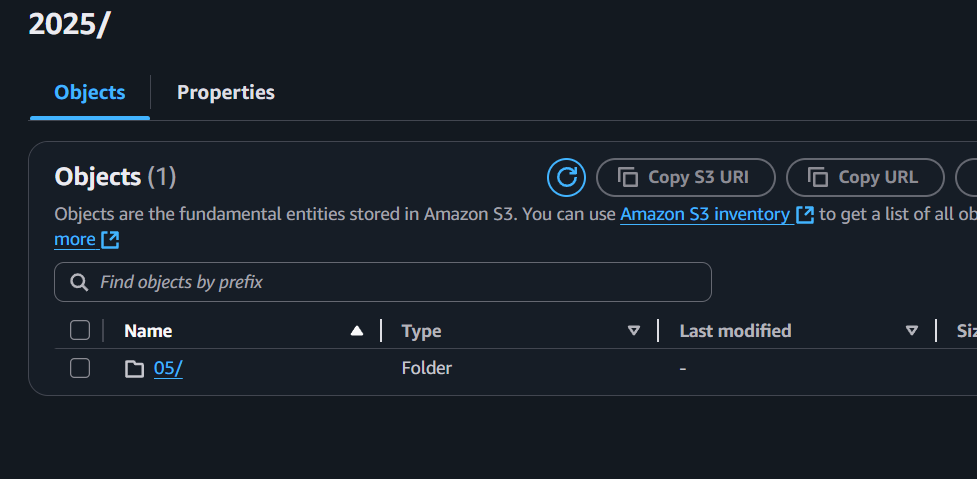
**Testing and Validation**

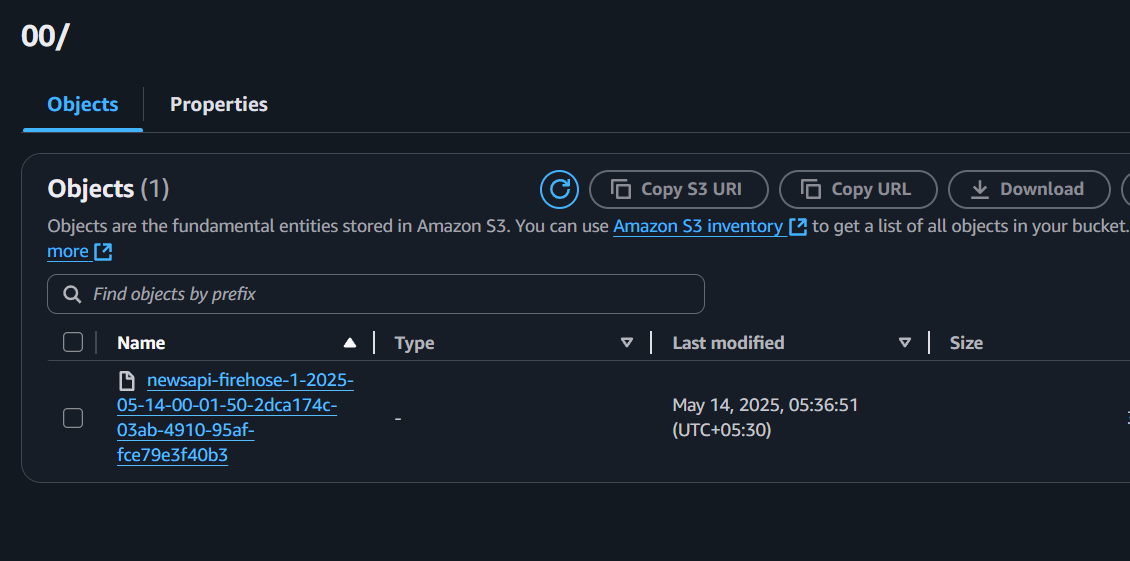
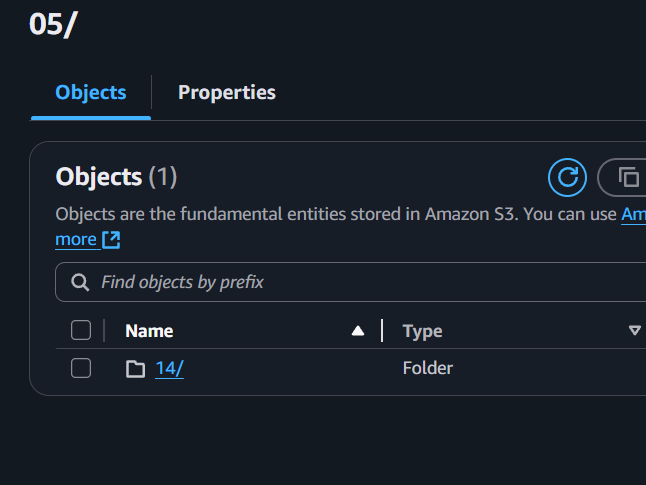
* Triggered the Lambda function manually to initiate the API request.
* Verified S3 bucket content — confirmed the receipt of newline-delimited JSON articles.

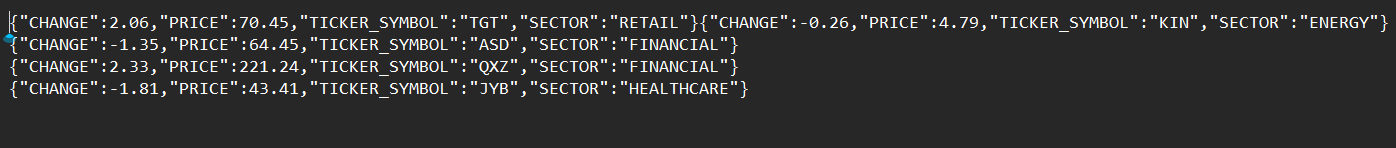












**Outcomes**

* Successfully implemented a scalable, serverless news ingestion pipeline on AWS.
* Ingested and stored real-time structured news data in a centralized S3 bucket.
* Built a foundation for real-time dashboards or analytics using AWS-native tools.