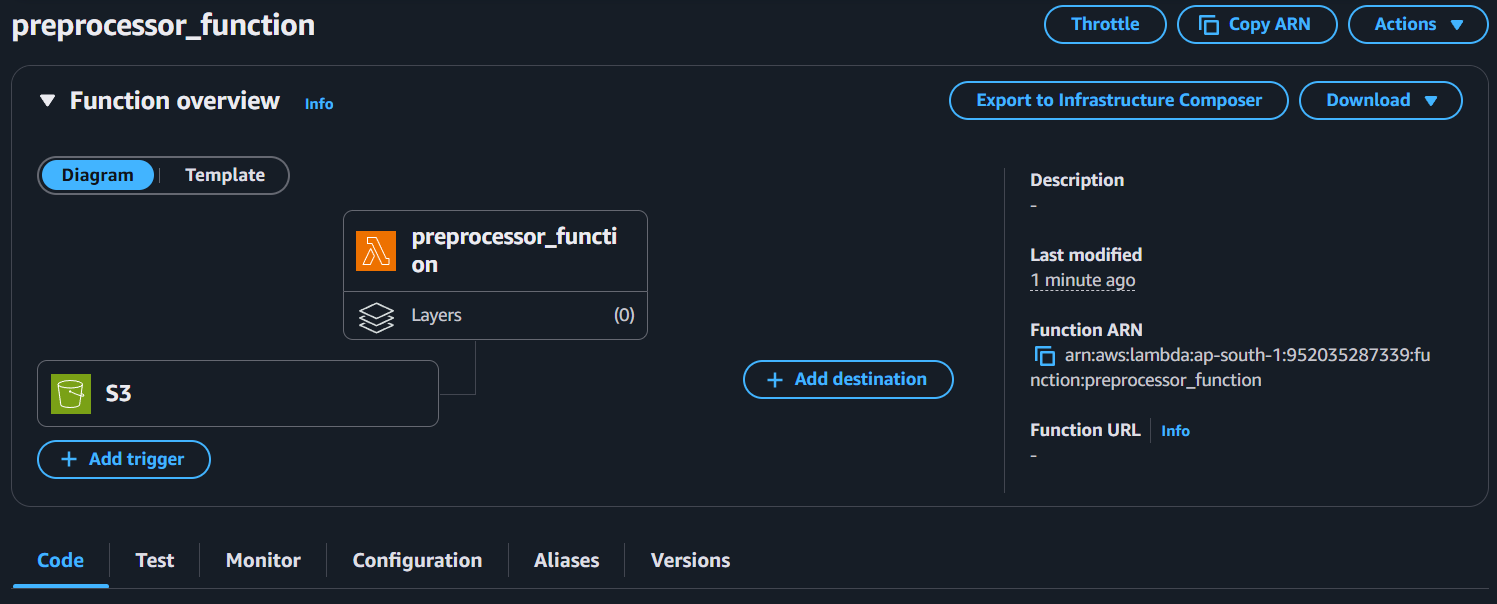
**AWS Lambda-Based Automated Data Cleaning upon S3 Upload**

**Overview**

This solution implements an automated data cleaning and preprocessing pipeline using AWS Lambda, which is triggered when a CSV file is uploaded to a designated folder in an S3 bucket. The processed output is stored in a separate folder within the same bucket.

**Components**

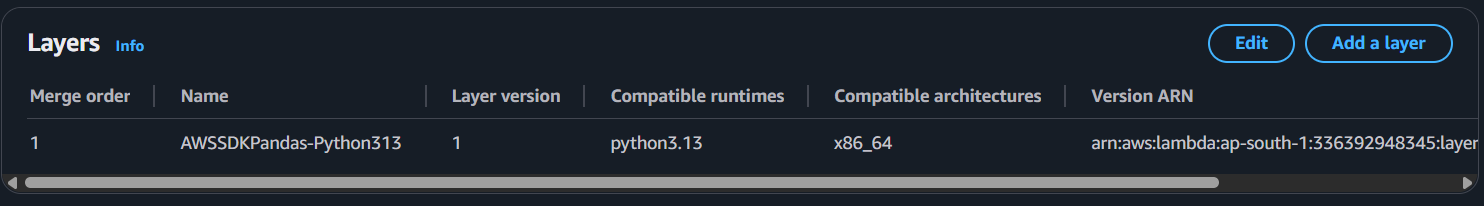
AWS Lambda Function (Preprocessor\_function).



* Automatically triggered when a .csv file is uploaded to incoming/ path in the S3 bucket.
* **S3 Bucket:** smartwatch-health-data-bucket
* **Trigger Suffix:** .csv
* Output Folder: cleaned/

Lambda Layer

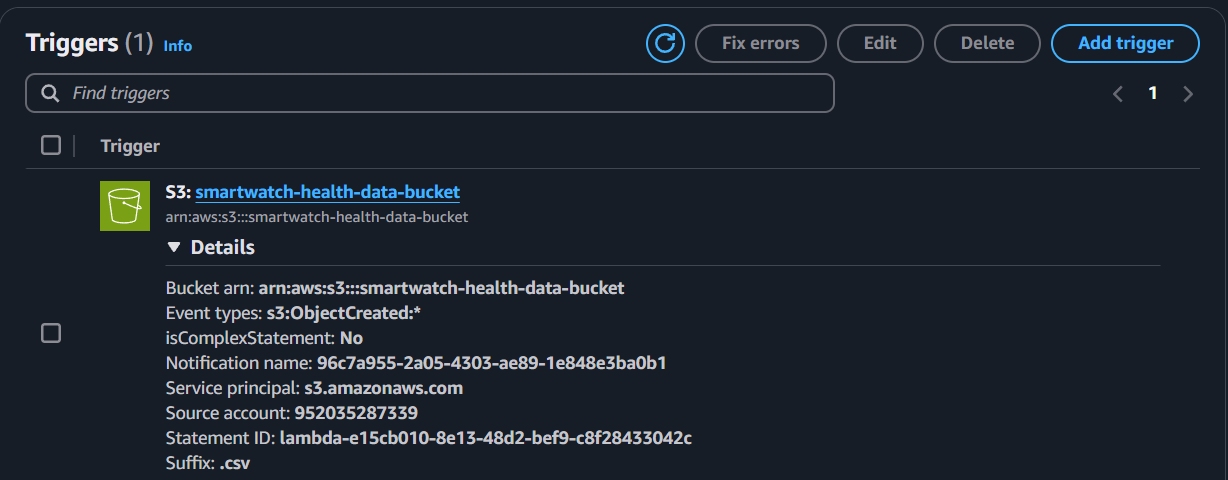
A Lambda layer containing the pandas library was added to support data processing.

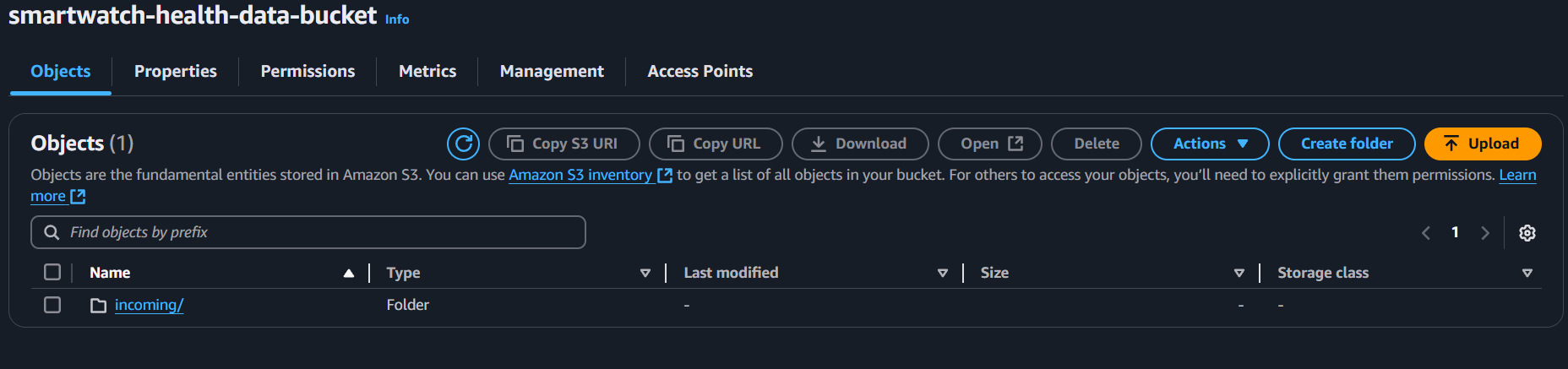


**Functionality**

**Trigger:**

When a new .csv file is uploaded to incoming/ path in the S3 bucket.



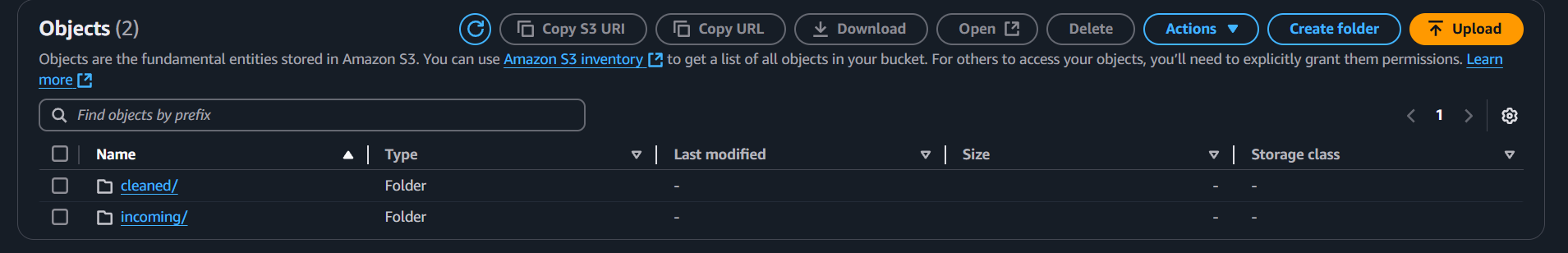


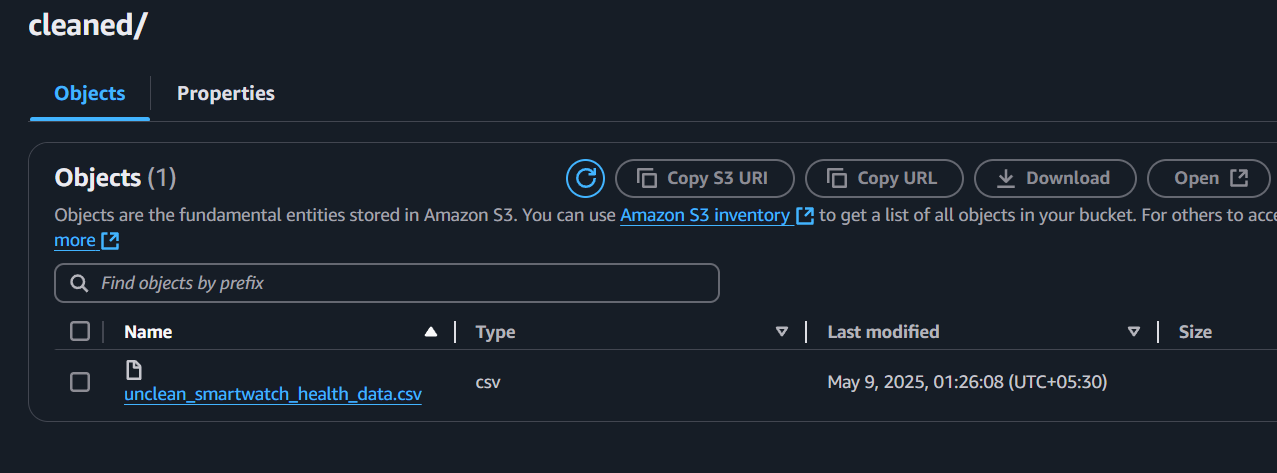
Processing Steps:

* Read File from S3
* Preprocess Using Pandas:
* Remove duplicate rows
* Drop fully empty rows
* Forward-fill missing values
* Normalize column names (lowercase, underscores)

**Save and Upload:**

* Save cleaned data in memory
* Upload to cleaned/ folder in the same bucket





**Lambda Function Code (Python)**

***import boto3***

***import pandas as pd***

***import io***

***def lambda\_handler(event, context):***

***s3 = boto3.client('s3')***

***source\_bucket = event['Records'][0]['s3']['bucket']['name']***

***source\_key = event['Records'][0]['s3']['object']['key']***

***destination\_bucket = source\_bucket***

***destination\_key = f"cleaned/{source\_key.split('/')[-1]}"***

***try:***

***response = s3.get\_object(Bucket=source\_bucket, Key=source\_key)***

***df = pd.read\_csv(io.BytesIO(response['Body'].read()))***

***df.drop\_duplicates(inplace=True)***

***df.dropna(how='all', inplace=True)***

***df.fillna(method='ffill', inplace=True)***

***df.columns = [col.strip().lower().replace(" ", "\_") for col in df.columns]***

***buffer = io.StringIO()***

***df.to\_csv(buffer, index=False)***

***buffer.seek(0)***

***s3.put\_object(Bucket=destination\_bucket, Key=destination\_key, Body=buffer.getvalue())***

***return {***

***'statusCode': 200,***

***'body': f'Successfully cleaned and uploaded to {destination\_key}'***

***}***

***except Exception as e:***

***return {***

***'statusCode': 500,***

***'body': f'Error processing file: {str(e)}'***

***}***

**Output Example**

* Input: incoming/raw\_data.csv
* Output: cleaned/raw\_data.csv (Cleaned and preprocessed)

**Security & IAM**

Ensure the Lambda function has the following IAM permissions:

***{***

***"Effect": "Allow",***

***"Action": [***

***"s3:GetObject",***

***"s3:PutObject"***

***],***

***"Resource": "arn:aws:s3:::smartwatch-health-data-bucket/\*"***

***}***

**Summary**

This project implements an automated, serverless data cleaning pipeline using AWS Lambda triggered by uploads to an Amazon S3 bucket. When a CSV file is placed in the incoming/ directory of the smartwatch-health-data-bucket, the Lambda function is triggered to preprocess the data using the pandas library. The process includes removing duplicates, handling missing values, and standardizing column names. The cleaned output is then stored in the cleaned/ directory within the same bucket. This solution enables real-time preprocessing of health-related smartwatch data with no manual intervention, ensuring data quality and consistency for downstream analytics or machine learning applications.