

### **3\*\*. Serverless Data Processing Pipeline\*\***

**Objective: Build a serverless pipeline for processing data (e.g., log processing or ETL jobs).**

Approach:

- Data Ingestion: Use AWS services like S3 or Kinesis to ingest data.
- Processing: Create Lambda functions to process the ingested data.
- Storage: Store the processed data in an appropriate AWS service, like S3 or DynamoDB.

● Monitoring: Set up CloudWatch to monitor the pipeline's performance and to log any issues.

**Goal: Learn to build a serverless data processing pipeline, understanding the flow of data through various AWS services.**

- Create lambda function and update the code to access the txt file from input directory of S3 bucket and generate uppercase output in output directory

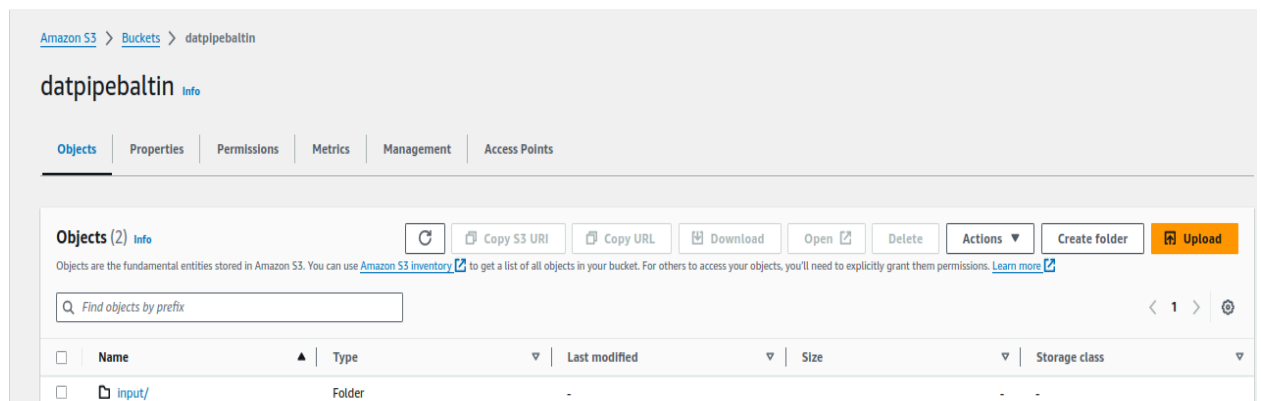


```

1 import boto3
2 from io import StringIO
3
4 def lambda_handler(event, context):
5     # Define the S3 bucket and file path
6     bucket_name = 'datpipebaltin'
7     input_key = event['Records'][0]['s3']['object']['key']
8
9     # Check if the file is in the 'input' folder
10    if not input_key.startswith('input/'):
11        print(f'The file '{input_key}' is not in the 'input' folder. Skipping.')
12        return
13
14    # Create an S3 client
15    s3 = boto3.client('s3')
16
17    # Read the content of the file from S3
18    response = s3.get_object(Bucket=bucket_name, Key=input_key)
19    content = response['Body'].read().decode('utf-8')
20
21    # Transform the content to uppercase
22    transformed_content = content.upper()
23
24    # Upload the transformed content to the 'output' folder in the same bucket
25    output_key = input_key.replace('input/', 'output/')
26    s3.put_object(Body=transformed_content, Bucket=bucket_name, Key=output_key)
27
28    print(f'File '{input_key}' successfully transformed and uploaded to '{output_key}' in S3.')
29


```

- Create S3 bucket and create folder(input) to upload file in lowercase text(.txt file)



- Add trigger

**Trigger configuration** [Info](#)

 **S3**

aws asynchronous storage

▼

**Bucket**  
Choose or enter the ARN of an S3 bucket that serves as the event source. The bucket must be in the same region as the function.  

✕ ↻

Bucket region: us-east-1

**Event types**  
Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.  

▼

POST ✕

**Prefix - optional**  
Enter a single optional prefix to limit the notifications to objects with keys that start with matching characters.

**Suffix - optional**  
Enter a single optional suffix to limit the notifications to objects with keys that end with matching characters.

**Recursive invocation**  
If your function writes objects to an S3 bucket, ensure that you are using different S3 buckets for input and output. Writing to the same bucket increases the risk of creating a recursive invocation, which can result in increased Lambda usage and increased costs. [Learn](#)

- Upload file in input folder of the S3 bucket

[Amazon S3](#) > [Buckets](#) > [datpipebaltin](#) > [input/](#) > Upload

**Upload** [Info](#)

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

**Files and folders (1 Total, 41.0 B)** Remove Add files Add folder

All files and folders in this table will be uploaded.

< 1 >

<input type="checkbox"/>	Name	Folder
<input type="checkbox"/>	random.txt	-

**Destination** [Info](#)

**Destination**  
`s3://datpipebaltin/input/`

**Destination details**  
Bucket settings that impact new objects stored in the specified destination.

**Permissions**  
Grant public access and access to other AWS accounts.

**Properties**  
Specify storage class, encryption settings, tags, and more.

Cancel Upload

- **Output is generated**

[Amazon S3](#) > [Buckets](#) > [datapipebaltin](#) > [output/](#)

output/

Copy S3 URI

Objects

Properties

Objects (1) Info

Refresh

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

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](#)

Find objects by prefix

< 1 > ⚙

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	random.txt	txt	March 5, 2024, 10:55:39 (UTC+05:45)	41.0 B	Standard

- **Query the output object**

SQL query

Add SQL from templates Run SQL query

Amazon S3 Select supports only the SELECT SQL command. Using the S3 console, you can extract up to 40 MB of records from an object that is up to 128 MB in size. To work with larger files or more records, use the AWS CLI, AWS SDK, or Amazon S3 REST API. For more complex SQL queries, use [Amazon Athena](#)

```
1 /* To create reference point for writing SQL queries, you can display the first 5 records of input data by running the following SQL query: SELECT * FROM s3object s LIMIT 5 */
2 SELECT * FROM s3object s LIMIT 5
```

SQL Ln 1, Col 1 Errors: 0 Warnings: 0

Query results

Download 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 1303 ms

Bytes returned: 41 B

```
1 A CRAZY BROWN FOX JUMPED OVER A LAZY DOG
2
```

- Go to Cloudwatch log group and select datapipe lambda function

CloudWatch

Favorites and recents

Dashboards

Alarms 2 6 0

In alarm

All alarms

Billing

Logs

Log groups

Log Anomalies

Live Tail

Logs Insights

Metrics

X-Ray traces

Events

Application Signals

Network monitoring

Insights

CloudWatch > Log groups

Log groups (1/10)

By default, we only load up to 10000 log groups.

☐ Exact match

< 1 >

<input type="checkbox"/>	Log group	Log class	Anomaly d...	Data prote...	Sensitive ...	Retention	Metric filters
<input type="checkbox"/>	/aws/lambda/RedshiftEventSubscription	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/RedshiftOverwatch	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/RoleCreationFunction	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/ToDoFunction	Standard	Configure	-	-	Never expire	-
<input checked="" type="checkbox"/>	/aws/lambda/datapipe	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/datapipefunc	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/datapipefunction	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/serverless-api	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/serverlessLambdaFunction	Standard	Configure	-	-	Never expire	-
<input type="checkbox"/>	/aws/lambda/toDoApp	Standard	Configure	-	-	Never expire	-

- View logs

CloudWatch > Log groups > /aws/lambda/datapipe > 2024/03/05/[\$LATEST]f287302d1eee412c8752414492f34447

Log events

Clear

1m

30m

1h

12h

Custom

Local timezone

Display

Timestamp

Message

No older events at this moment. [Retry](#)

▶

2024-03-05T10:55:35.671+05:45

INIT\_START Runtime Version: python:3.12.v19 Runtime Version ARN: arn:aws:lambda:us-east-1::runtime:a79ae1de439e89e7ald:89465a221e8fe9bb3c495...

▶

2024-03-05T10:55:35.961+05:45

START RequestId: 78424bb0-88f0-497e-a962-bbald8bdf007 Version: \$LATEST

▶

2024-03-05T10:55:38.704+05:45

File 'input/random.txt' successfully transformed and uploaded to 'output/random.txt' in S3.

▶

2024-03-05T10:55:38.739+05:45

END RequestId: 78424bb0-88f0-497e-a962-bbald8bdf007

▶

2024-03-05T10:55:38.740+05:45

REPORT RequestId: 78424bb0-88f0-497e-a962-bbald8bdf007 Duration: 2778.92 ms Billed Duration: 2779 ms Memory Size: 128 MB Max Memory Used: 81...

▶

2024-03-05T10:55:39.576+05:45

START RequestId: f27f36af-fef8-4788-a67e-53d29b7cba46 Version: \$LATEST

▶

2024-03-05T10:55:39.576+05:45

The file 'output/random.txt' is not in the 'input' folder. Skipping.

▶

2024-03-05T10:55:39.577+05:45

END RequestId: f27f36af-fef8-4788-a67e-53d29b7cba46

▶

2024-03-05T10:55:39.577+05:45

REPORT RequestId: f27f36af-fef8-4788-a67e-53d29b7cba46 Duration: 1.48 ms Billed Duration: 2 ms Memory Size: 128 MB Max Memory Used: 81 MB

No newer events at this moment. [Auto retry paused.](#) [Resume](#)