Data Pipeline on AWS (ETL)

Description:

1. User: Starts the process by uploading a CSV file to the S3 bucket.

2. AWS S3 Bucket:

- Triggers an event when a new file is uploaded.
- Monitoring: CloudWatch logs S3 events for successful or failed uploads.

3. Lambda Function 1 (TransformData):

- Processes the uploaded CSV file.
- Sends a message to the Amazon SQS queue.
- Monitoring**:** CloudWatch tracks Lambda execution metrics such as invocation count, duration, errors, and throttles.

4. Amazon SQS Queue:

- Acts as a message buffer, ensuring asynchronous communication.
- Monitoring: CloudWatch tracks queue depth, message age, and any dead-letter queue (DLQ) activity.

5. Lambda Function 2 (SaveToMongo):

- Consumes messages from SQS.
- Writes the processed data to MongoDB.
- Monitoring: CloudWatch monitors invocation metrics and logs errors or exceptions.

MongoDB Atlas:

- The final destination where the data is stored.
- Monitoring: MongoDB Atlas provides its own monitoring tools for database performance and query metrics.

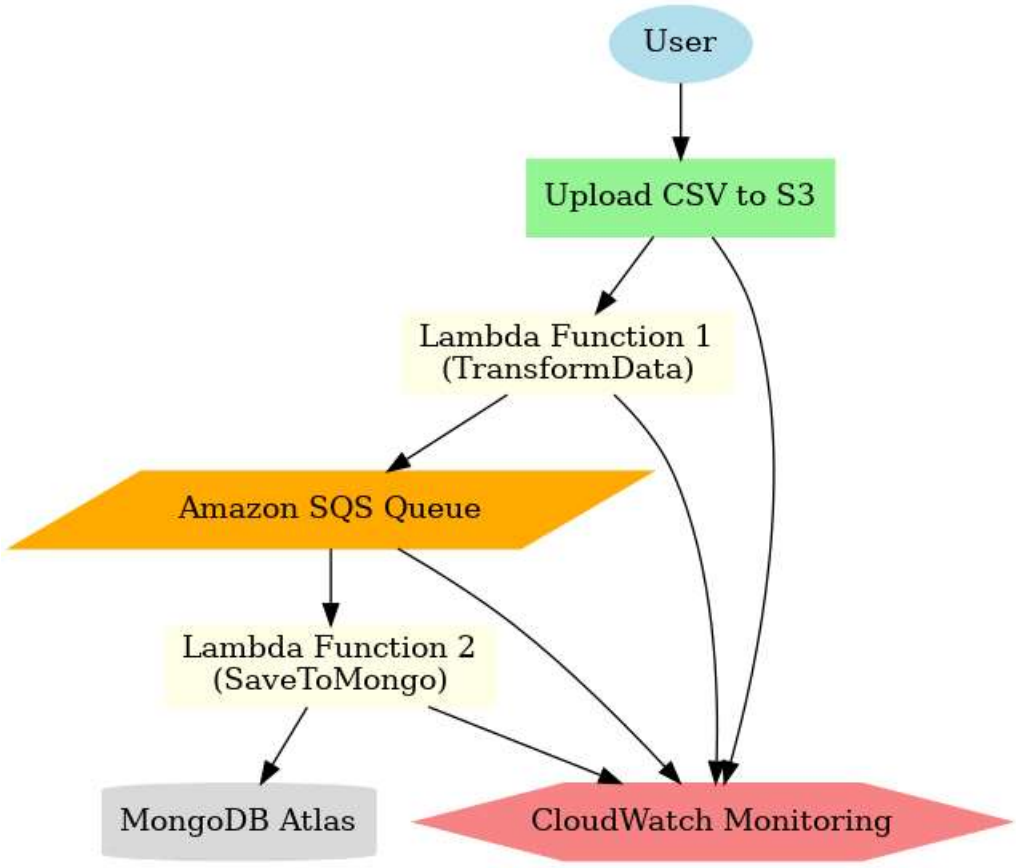
Dataset:

Financial Transactions

Dataset: Cards transactions (anonymized).

Source: Kaggle

Flow Chart:



Appendix:

Appendix A: Code

Lambda Function 1 (TransformData): Function name is TriggerSparkJob.

Code is attached as a seperate file.

File Name : LambdaFunction1-TransformData.py

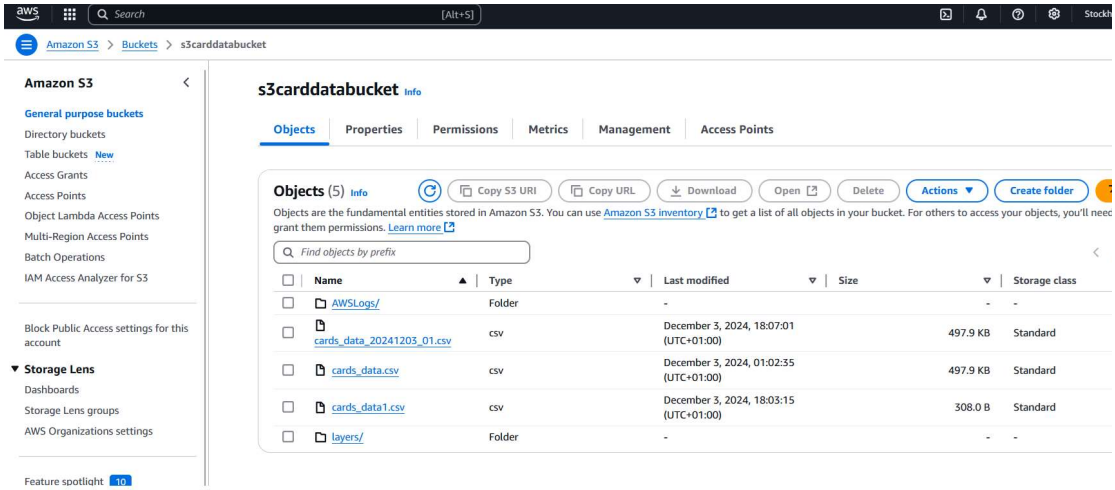
Lambda Function 2(MongoDBInsertion):

Code is attached as a seperate file.

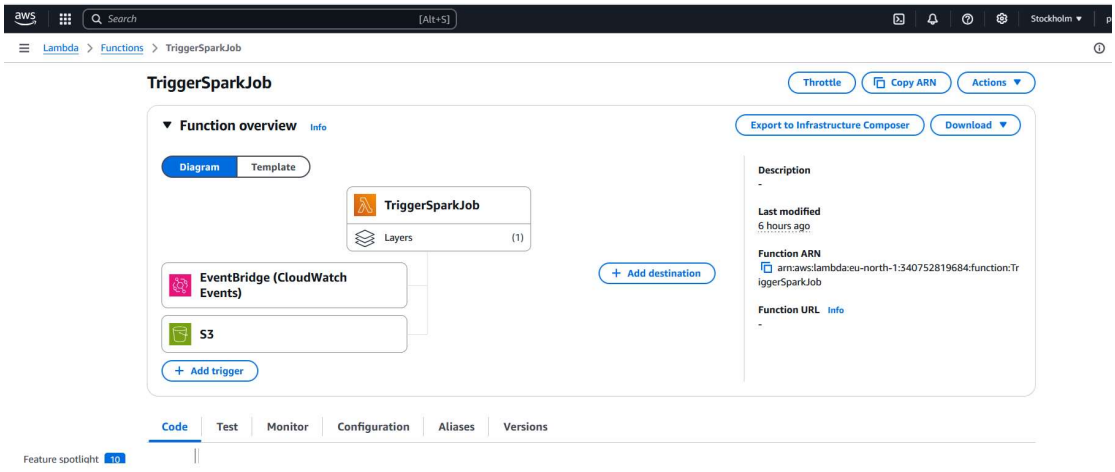
File Name : LambdaFunction2-MongoDBInsertion.py

Appendix B: Screenshots

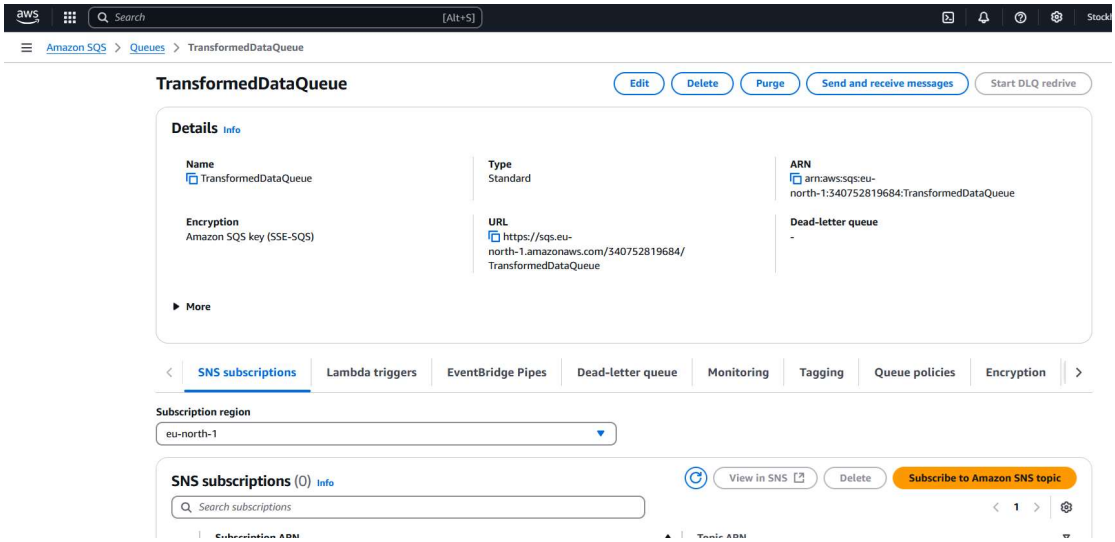
AWS S3 Bucket:



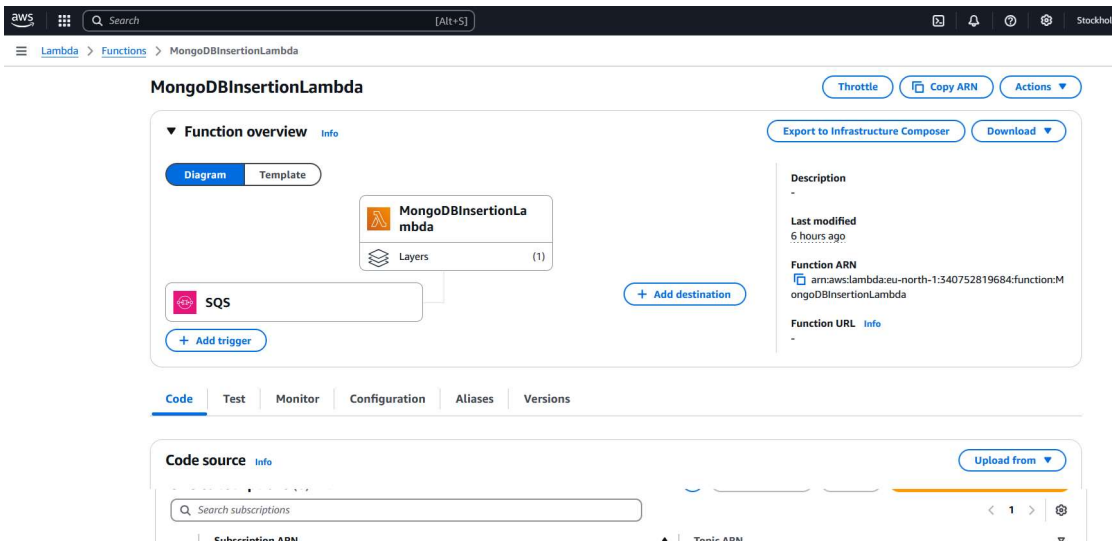
Lambda Function 1 (TransformData): Function name is TriggerSparkJob.



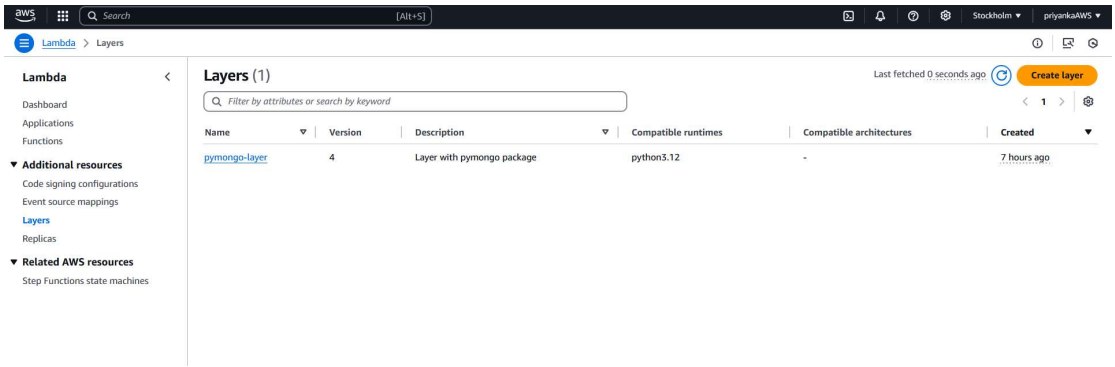
Amazon SQS Queue:



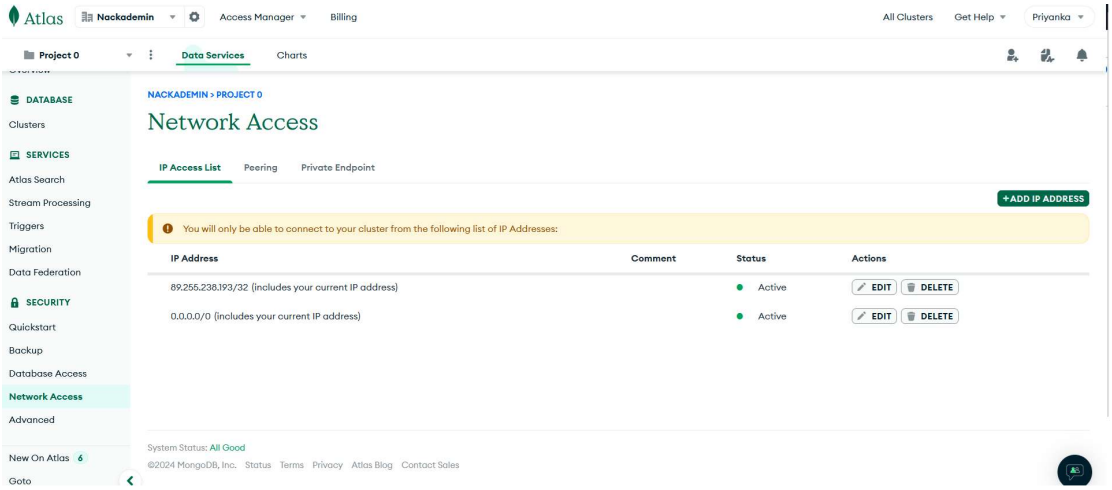
Lambda Function 2 (MongoDBInsertionLambda):

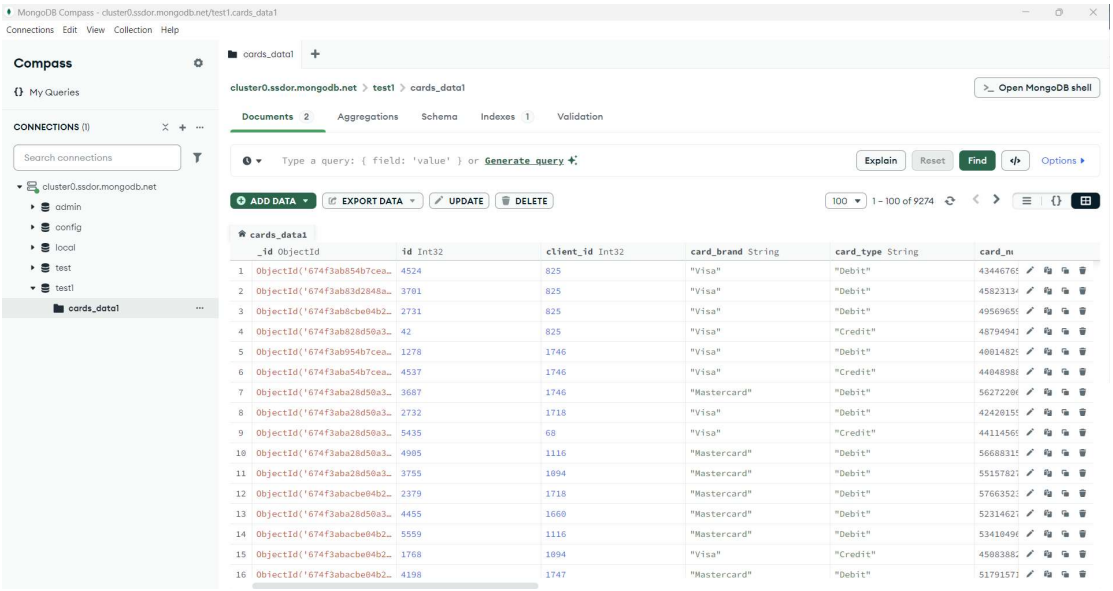


Lambda Layers:

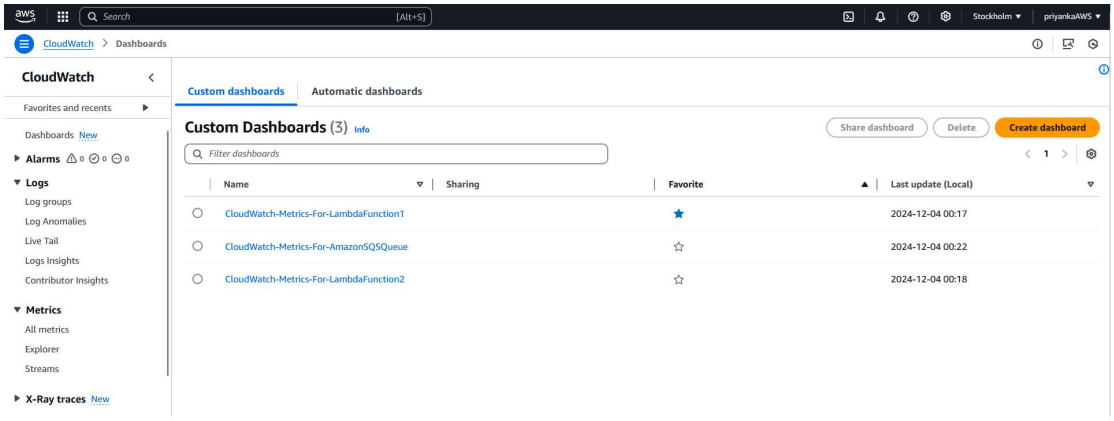


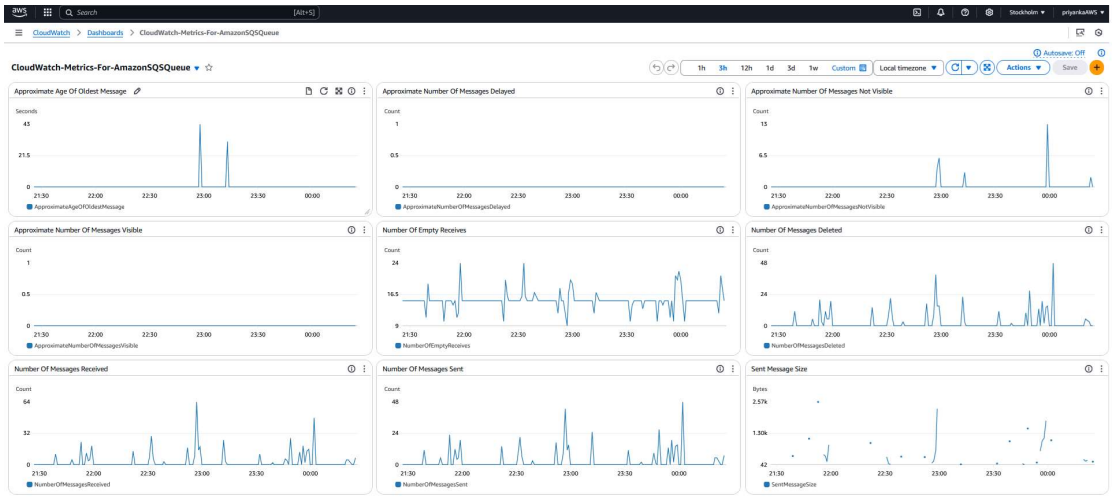
MongoDB :

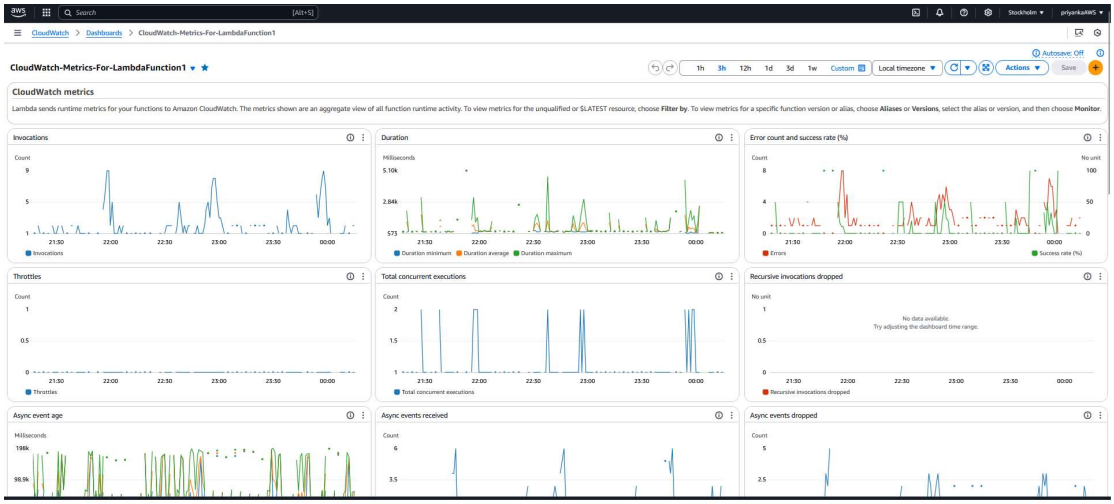


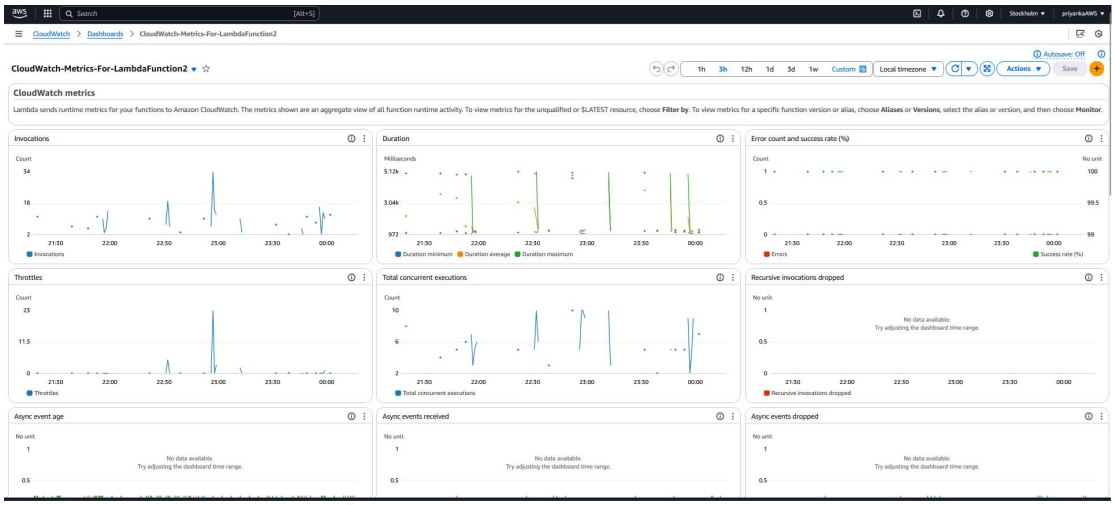


Appendix C: Dashboards


aws
Search
[Alt+S]
Stockholm
priyankaAWS
CloudWatch
Dashboards
CloudWatch
Favorites and recents
Dashboards New
Alarms 0 0 0
Logs
Log groups
Log Anomalies
Live Tail
Logs Insights
Contributor Insights
Metrics
All metrics
Explorer
Streams
X-Ray traces New
Custom dashboards
Automatic dashboards
Custom Dashboards (3) Info
Share dashboard
Delete
Create dashboard
Filter dashboards
Name
Sharing
Favorite
Last update (Local)
CloudWatch-Metrics-For-LambdaFunction1
2024-12-04 00:17
CloudWatch-Metrics-For-AmazonSQSQueue
2024-12-04 00:22
CloudWatch-Metrics-For-LambdaFunction2
2024-12-04 00:18


CloudWatch-Metrics-For-AmazonSQSQueue
Approximate Age Of Oldest Message
Approximate Number Of Messages Delayed
Approximate Number Of Messages Not Visible
Approximate Number Of Messages Visible
Number Of Empty Receives
Number Of Messages Deleted
Number Of Messages Received
Number Of Messages Sent
Sent Message Size





NOTE: Public Links for CloudWatch Dashboards are given in CloudWatch-Dashboard-Public-Links.rtf document.