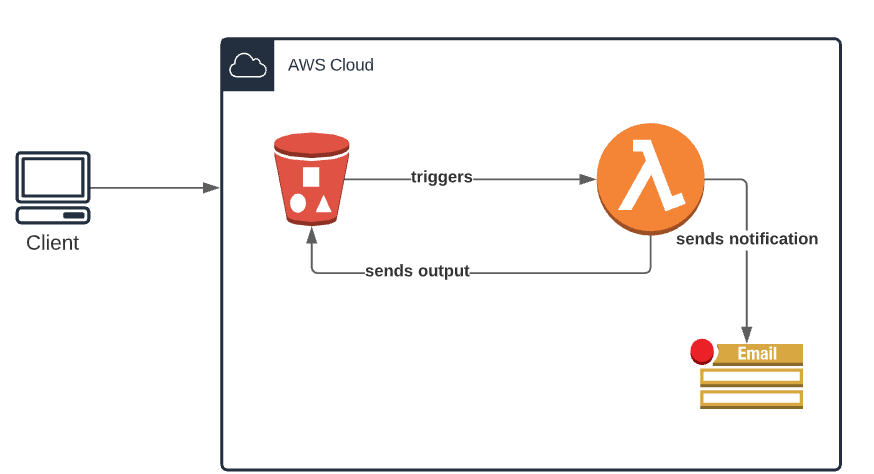
**Technical Details**

**Project:** **Web Traffic Analysis**

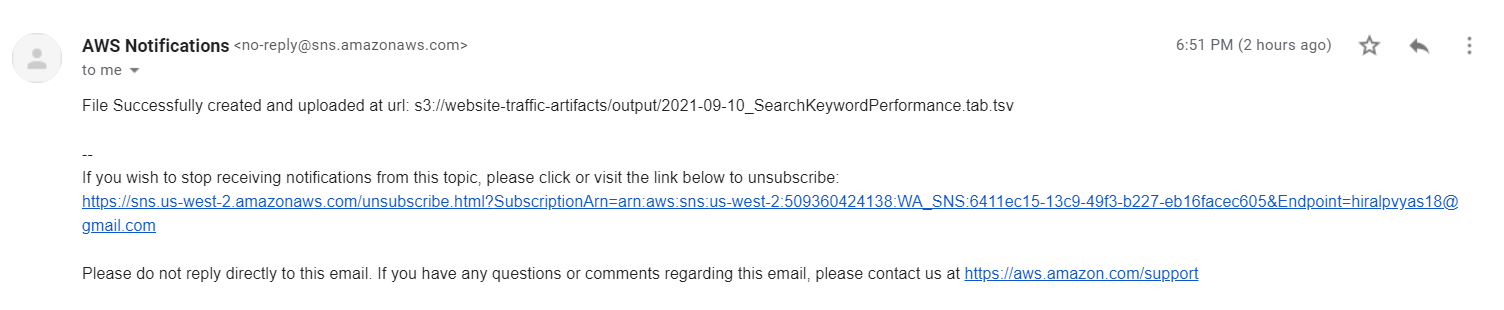
**High level Architecture:**



**Workflow**:

1. Client sends hit level data file to Adobe's Team for further analysis of the data.
2. File lands in S3 bucket's input folder, as soon as file lands in S3 it invokes a Lambda Function
3. All Business logic and analysis of business problem is done in python script hosted on lambda
4. Once Lambda function sends output file desired by client to S3's output folder an SNS notification is send to the respective parties describing the s3 URL path for the output

Sample AWS Notification:



**Deployment Details:**

AWS Infrastructure components like S3, Lambda Function, Lambda Layer, Role, Policies, SNS Topic are created using Terraform Template.

**main.tf** has all the resources

**variables**.**tf** hold all variables data like s3 bucket name, lambda function name etc.

**Commands to run Terraform Template:**

terraform init

terraform plan

terraform apply

**Potential Improvements in Pipeline**

1. SQS queue can be implemented if multiple files are uploaded at once
2. Multiple Alternatives like Dask, pandas chunksize or pyspark can be used if larger files are uploaded for processing
3. Architecture can be made more modular by using step functions
4. Docker container can be used and deployed on lambda to create lambda layer.Also if we have huge size of the of the code and dependencies.

**Potential KPI’s to track**

* Geographic data of visitors to know which product is more prevalent in which areas
* Count of users based on keyword searches, category of search, browser, number of visitors(track by time), page hits

Tracking Activity and Inactivity:

Activity peaks

* Look at Aggregated area of interest (Which categories user looks at in consecutive visits or visits over a period)
* Check Returning users based on number of visits
* Average length of a session- Can help determine the response time
* Diversity of referrers for a user

Inactivity and gaps

* What do they not respond to or pages they don't visit ever- Can help gauge their interests so client can provide relevant ads, recommendations for the respective categories?
* Average time between purchases