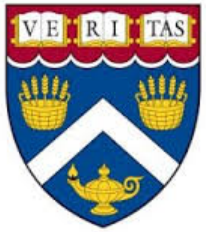
DEPLOYING A REAL TIME AND BATCH PREDICTION SERVICE

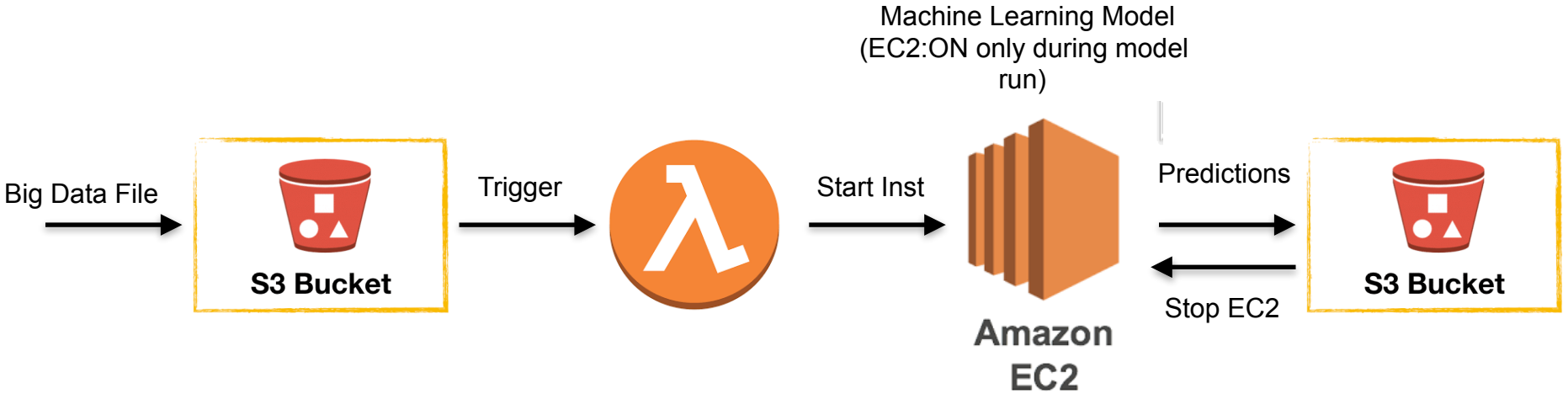
Amit R. Gupta



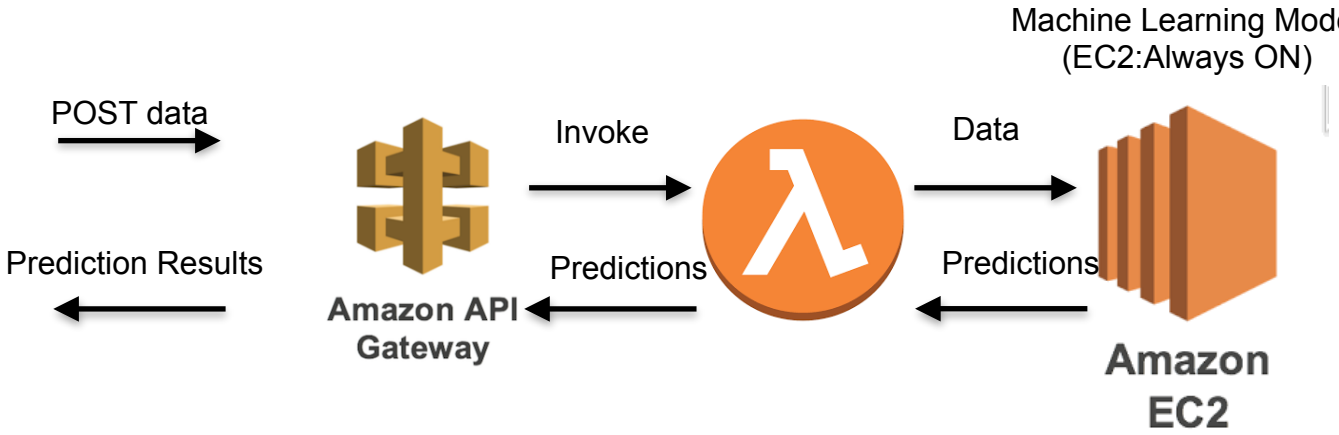
CSCI E-90 Cloud Services, Infrastructure and Computing
Fall 2019
Harvard University Extension School
Instructor: Greg Misicko

Goals

#1 Automated Batch Prediction Service

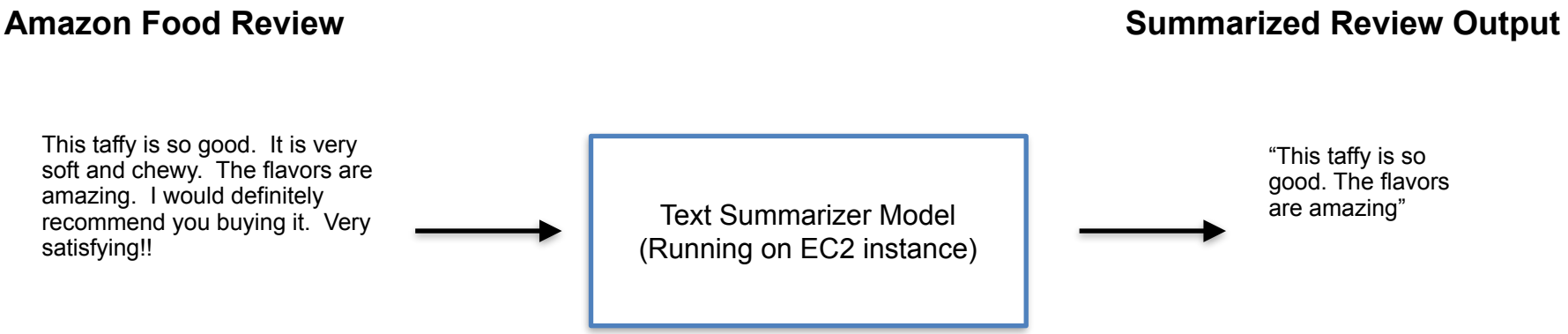


#2 Real-time Prediction Service

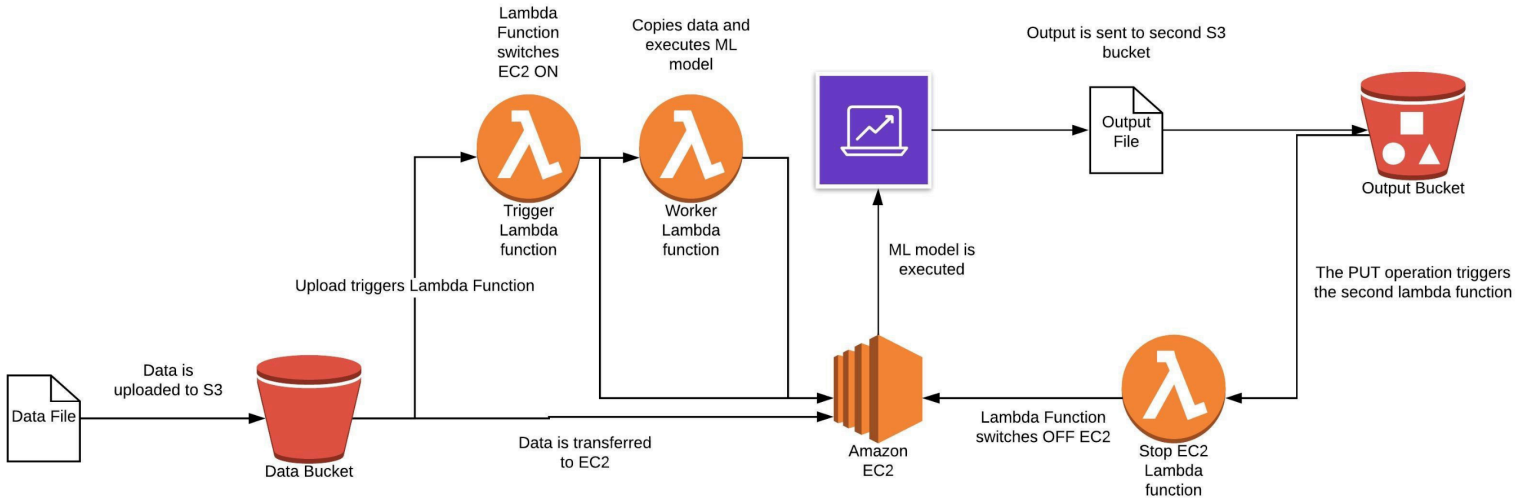


Machine Learning (ML) Model

- EC2 instance is running the “Text Summarizer” ML Model. (Reference: https://towardsdatascience.com/data-scientists-guide-to-summarization-fc0db952e363)
- Effectively the model identifies key sentences in the text.
- The model is using the K-means technique to cluster similar sentences, and pick the closest sentences to the cluster centroid.
- Dataset: Amazon Food Reviews (https://www.kaggle.com/snap/amazon-fine-food-reviews)



Batch Text Summarization



Reference: https://towardsdatascience.com/automating-machine-learning-models-on-aws-bfa183fe4065.

Steps:

1) Batch data/text (Amazon Food Reviews data file) posted to input S3 triggers a Lambda function ("Trigger Lambda") which turns on the EC2 instance.
2) "Trigger Lambda" invokes the "Worker Lambda" function.
3) Worker Lambda orchestrates copying the data from S3 to EC2 instance, running the model and copying the results to S3.
4) "EC2" instance executes the text summarizer model and writes the model to output S3.
5) The output S3 triggers a Lambda function ("Stop EC2 Lambda") to stop the EC2 instance.

Benefits:

1) Cost savings. Lambda functions and EC2 instances run only when data is posted to S3.

Example Worker Lambda Function

Paramiko package was added as a Lambda LAYER (*)

SSH tunnel into EC2 instance

List of commands run on EC2 instance

```
import json
import boto3
import paramiko

def lambda_handler(event, context):

    s3_client = boto3.client('s3')
    #Download private key file from secure S3 bucket
    s3_client.download_file('cscie90-finalproject-virginia','virginia.pem', '/tmp/virginia.pem')

    k = paramiko.RSAKey.from_private_key_file("/tmp/virginia.pem")
    c = paramiko.SSHClient()
    c.set_missing_host_key_policy(paramiko.AutoAddPolicy())

    host=event['IP']
    print ("Connecting to " + host)
    c.connect( hostname = host, username = "ec2-user", pkey = k )
    print ("Connected to " + host)

    commands = [
        "aws s3 cp s3://cscie90-finalproject-inputdata/Reviews.csv --region us-east-1 /home/ec2-user/Reviews.csv",
        "python3 /home/ec2-user/kmeans.py",
        "aws s3 cp /home/ec2-user/top_500_summary.csv s3://cscie90-finalproject-results/summary.csv",
        "rm /home/ec2-user/top_500_summary.csv"
        ]

    for command in commands:
        print ("Executing {}".format(command))
        stdin , stdout, stderr = c.exec_command(command)
        print (stdout.read())
        print (stderr.read())

    return
    {
        'message' : "Script execution completed. See Cloudwatch logs for complete output"
    }
```

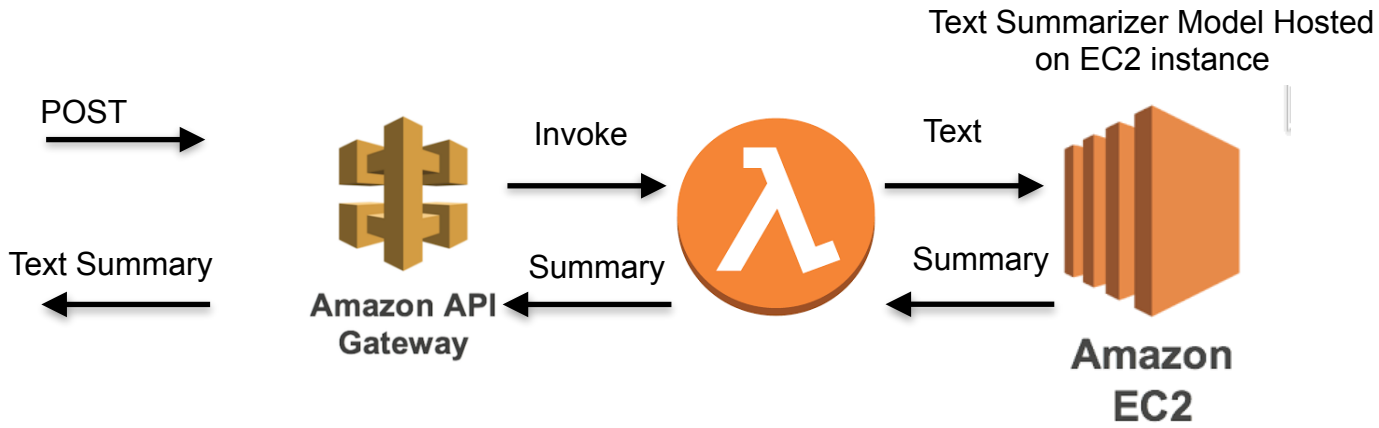
Role: Lambda function was assigned administrative rights to access the EC2 instance.

(*):Steps listed in the following article were followed to install the python paramiko (SSH tunnel) package as a layer.
https://towardsdatascience.com/hosting-your-ml-model-on-aws-lambdas-api-gateway-part-1-9052e6b63b25

Example Batch Results in the output S3 bucket

Review ID	Amazon Food Review Text	Summarized Text
2	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut into tiny squares and then liberally coated with powdered sugar. And it is a tiny mouthful of heaven. Not too chewy, and very flavorful. I highly recommend this yummy treat. If you are familiar with the story of C.S. Lewis' "The Lion, The Witch, and The Wardrobe" - this is the treat that seduces Edmund into selling out his Brother and Sisters to the Witch.	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. Not too chewy, and very flavorful.
5	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelon, root beer, melon, peppermint, grape, etc. My only complaint is there was a bit too much red/black licorice-flavored pieces (just not my particular favorites). Between me, my kids, and my husband, this lasted only two weeks! I would recommend this brand of taffy -- it was a delightful treat.	The taffy was all very enjoyable with many flavors: watermelon, root beer, melon, peppermint, grape, etc. My only complaint is there was a bit too much red/black licorice-flavored pieces (just not my particular favorites). Between me, my kids, and my husband, this lasted only two weeks!
6	This saltwater taffy had great flavors and was very soft and chewy. Each candy was individually wrapped well. None of the candies were stuck together, which did happen in the expensive version, Fralinger's. Would highly recommend this candy! I served it at a beach-themed party and everyone loved it!	This saltwater taffy had great flavors and was very soft and chewy. None of the candies were stuck together, which did happen in the expensive version, Fralinger's. Would highly recommend this candy!
7	This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. Very satisfying!!	This taffy is so good. The flavors are amazing.

Real Time Text Summary

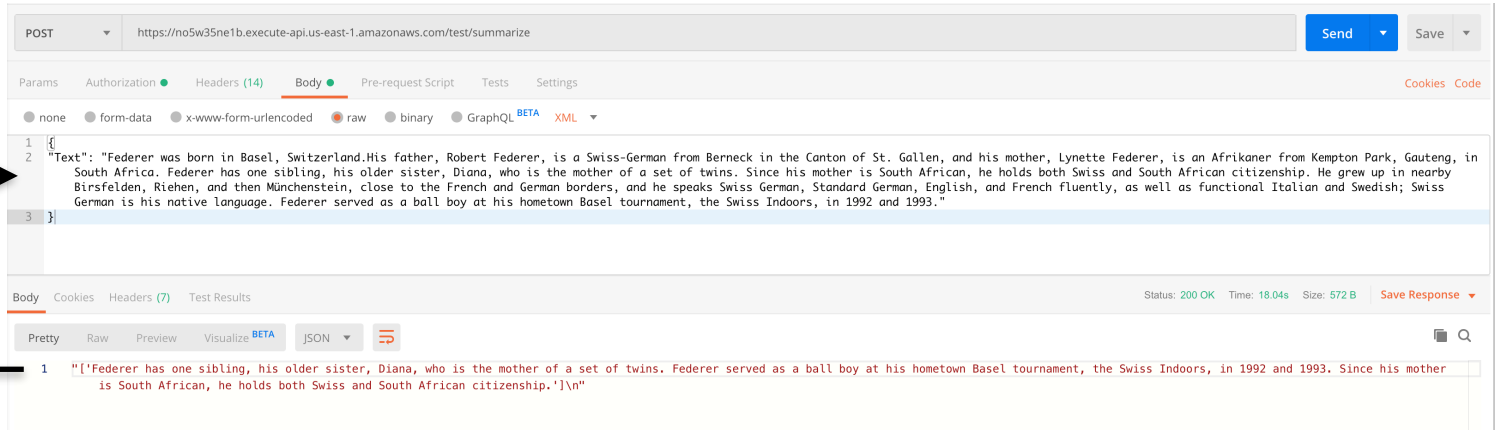


POSTMAN Application

API GATEWAY

POSTED TEXT

REAL TIME SUMMARY



Steps:

1) Sample data taken from Roger Federer Wikipedia was sent to API Gateway using POSTMAN.

2) API Gateway calls the Lambda function. This function orchestrates the execution of text summarizer model on the sample data.

3) “EC2” instance executes the text summarizer model and returns the result to lambda function. (Please note EC2 instance is always on to reduce latency.)

4) The lambda function returns the results back to POSTMAN as seen in the response body.

Summary

- Goals #1 and #2 were met successfully
- Lambda functions can play an important role in the cost savings. One can optimally start and stop EC2 instances, and lambda functions are billed on per invocation.
- Lambda functions have a space limitation. They are unable to accommodate deep learning frameworks like tensor flow. The tensor flow library is 350+ Mb, and Lambda functions support a max size 262Mb in its layers. In such situations EC2 instances have to be used (as shown in the project) to run the model and return predictions.

Resources

- Project Report (pdf):
- Code:https://github.com/amitrgupta27/cscie90-finalproject
- Youtube URL (Project Overview):https://youtu.be/aQPjzQAn_Ys
- Youtube URL (Deep Dive): https://www.youtube.com/watch?v=s8wfpDl0Okk
- Contact:amitrgupta27@gmail.com
- References:
  - https://towardsdatascience.com/hosting-your-ml-model-on-aws-lambdas-api-gateway-part-1-9052e6b63b25
  - https://towardsdatascience.com/hosting-your-ml-model-on-aws-lambdas-api-gateway-part-2-23517609522b
  - https://towardsdatascience.com/automating-machine-learning-models-on-aws-bfa183fe4065
  - https://towardsdatascience.com/data-scientists-guide-to-summarization-fc0db952e363