CS4830: Big Data Lab Assignment-6
Shania Mitra CH18B067

Q1. Count the number of lines in a file uploaded to GCS bucket in real-time by using Google Cloud Functions and Pub/Sub

- **Write a Google cloud Function which gets triggered whenever a file is added to a bucket and publishes the file name to a topic in Pub/Sub.**
- **Write a python file, which acts as a subscriber to this topic and prints out the number of lines in the file in real-time**

main.py

```
def message(data, context):
    from google.cloud import pubsub_v1
    pub_client = pubsub_v1.PublisherClient()
    topic_name = 'projects/graphite-byte-260703/topics/topic_lab6'
    pub = pubsub_v1.PublisherClient()
    pub.create_topic(topic_name)
    out = data['name']
    out = out.encode("utf-8")
    pub_client.publish(topic_name, out)
    print("Message Received")
```

subscription.py

```
from google.cloud import pubsub_v1
from google.cloud import storage

sub = pubsub_v1.SubscriberClient()
topic_name = 'projects/graphite-byte-260703/topics/topic_lab6'
sub_name = 'projects/graphite-byte-260703/subscriptions/subscript'
sub = pubsub_v1.SubscriberClient()
sub.create_subscription(name=sub_name, topic=topic_name)

def callback(package):
    x = package
    print(x.data)
    with open('addresses.csv', 'r') as f:
        count = 0
        for line in f:
            count = count+1
    print('Expected Line Count: ' + str(count))
    package.ack()

future = sub.subscribe(sub_name, callback)

try:
    future.result()
except KeyboardInterrupt:
    future.cancel()
```

Running it on Cloud Shell:

```
way2shania@cloudshell:~ (graphite-byte-260703)$ python3 subscription.py
Message Received
Expected Line Count: 22
way2shania@cloudshell:~ (graphite-byte-260703)$
```

Q2. There are two kinds of subscribers - pull and push subscribers. What are the differences between the two and when would you prefer one over the other?

The decision to push or pull is highly dependent on the situation.

- If everything can be managed from a single location, going with push makes sense because it will be easier to make adjustments to distribution agents.
- When the agent downloads the data rather than sending it, push is also better.
- Data travels from distributor to subscriber, and because agents employ subscribers, pull is preferred when several agents are coming from a single distributor.
- As a result, the preference is determined by the quantity and type of distribution agents as well as the speed with which subscribers can be connected.

Thus, the preference depends on the number and type of distribution agents and how fast the connectivity is to the subscribers.
