DataEng: Data Maintenance In-class Assignment

This week you will construct a data archiver that compresses, encrypts and stores pipelined data into a low-cost, high-capacity GCP Storage Bucket.

**Submit**: Make a copy of this document and use it to record your responses and results (use colored highlighting when recording your responses/results). Store a PDF copy of the document in your git repository along with your code before submitting for this week.

Develop a new python PubSub subscriber similar to the subscribers that you have created multiple times for this class. This new subscriber (archive.py) will receive data from a PubSub topic, compress the data, encrypt the data and store the resulting data into a [GCP Storage Bucket](https://cloud.google.com/storage/docs/creating-buckets).

## A. [MUST] Discussion Questions

When archiving data for a data pipeline we could (a) compress, (b) encrypt and/or (c) reduce the data. Here, “reducing the data” refers to the process of interpolating or aggregating detailed data, such as 5 second breadcrumbs for all buses on all trips, into coarser data. For example, we could aggregate 5-second breadcrumbs into 30-second breadcrumbs.

Under what circumstances might each of these transformations (compress, encrypt, reduce) be desirable for data archival?

**When to Compress:**

Storage Efficiency: When storage capacity is an issue data compression is helpful. It can result in cost savings, particularly when dealing with huge datasets, as it lowers the amount of disk space required to keep the data.

Data Transfer: Compression can help with data transfer over networks by reducing the amount of data that needs to be sent. This results in quicker transfer times and less bandwidth being used.

Backup and Archiving: Compressed data is more effective for long-term storage and backups because it takes up less space.

**When to Encrypt:**

Security and privacy: In order to prevent unwanted access to sensitive or private data, encryption is essential. This covers any proprietary information as well as financial and personal data.  
Regulatory Compliance: A lot of sectors have laws requiring data to be encrypted in order to protect its security and privacy (like GDPR and HIPAA).

**When to reduce:**

Data Volume: By combining finer-grained data (like 5-second breadcrumbs) with coarser data (like 30-second breadcrumbs)we can reduce the amount of data which will make managing and storing it easier and less expensive.

Data analysis: When fine-grained details are not required, aggregated data might be adequate for a variety of research and reporting purposes.

Performance: As less data needs to be processed, query performance can be enhanced with lower data quantities.

Would it make sense to combine these transformations?

Combining compression with encryption can be beneficial because compressed data can be encrypted to provide security and conserve storage. Though encrypted data does not compress well,we compress data before encrypting it.

Reduction and Compression: Data can be further decreased through compression once it has been aggregated. This mixture works good for long-term storage when in-depth investigation is not necessary.

Reduction, Compression, and Encryption: We can compress data to conserve storage space, encrypted to guarantee confidentiality and reduced to lessen its volume for optimal efficiency and security.

## B. [MUST] Create Test Pipeline

Create a new PubSub topic called “archivetest” or something similar. Create a new subscriber program (call it archiver.py) that subscribes to the topic, receives the data and (for now) discards it.

To produce test data, copy/reuse the publisher program used for your class project, and alter it to publish to the new archivetest topic. Run this test publisher manually to gather data (1 day and 100 vehicles) from busdata.cs.pdx.edu and test the archivetest topic and your archiver.py program.

As always, you can/should test your code with smaller data sets first. Try it with just one bus or one trip, and then when everything is working, run it with 100 vehicles.

## C. [MUST] Store Data to GCP Storage Bucket

Modify archiver.py to store all received data to a [GCP Storage Bucket](https://cloud.google.com/storage/docs/creating-buckets). You will need to create and configure a Storage Bucket for this purpose. We recommend using the Nearline Storage class for this assignment though you are free to choose any of the offered classes of service. Be sure to remove the bucket at the end of the week to reduce GCP credit usage.

How much bucket space (in KiBs) does it take to store 1 day of breadcrumbs for 100 vehicles?

71500 Kbs

## D. [SHOULD] Compress

Modify archiver.py to compress the data before it stores the data to the storage bucket. Use [zlib compression](https://stackabuse.com/python-zlib-library-tutorial/) which is provided by default by python. How large is the archived data compared to the original?

Compression ratio: 0.08

The original data is 69,397,232 bytes larger than compressed data.

How much bucket space (in KiBs) does it take to store the compressed data?

5400 KBs

## E. [SHOULD] Encrypt

Modify archiver.py to encrypt the data prior to writing it to the Storage Bucket. Your archive.py program should encrypt after compressing the data. Use RSA encryption as described here: [link](https://www.geeksforgeeks.org/how-to-encrypt-and-decrypt-strings-in-python/) There is no need to manage your private encryption keys securely for this assignment, and you may keep your private key in a file or within your python code.

Be sure to test your archiver by decrypting and decompressing the data stored in the Storage Bucket. We suggest that you create a separate python program for this purpose.

How much bucket space (in KiBs) does it take to store the encrypted, compressed data?

5400 KBs

## F. [ASPIRE] Add Archiving to your class project

Add an archiver to your class project’s pipeline(s). To receive extra credit, mention your archiver when submitting the next part of your project. You should only need one archiver for the entire project, so coordinate with your teammates if you choose to take this step. For the class project, we recommend storing to a Google Storage Bucket and compressing. Encryption is OK too but not necessary.