

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](#).

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?

Compress data to save storage space and reduce bandwidth during transfers, especially when dealing with large datasets. Encrypt data to ensure security and privacy, particularly for sensitive or regulated information, protecting it from unauthorized access. Reduce data granularity to improve storage efficiency and processing speed, suitable when detailed data is not necessary for analysis. Each transformation addresses specific needs: storage and transfer efficiency (compression), security (encryption), and analysis performance (reduction). Combining these transformations can provide comprehensive benefits, such as reducing first, then compressing, and finally encrypting for optimal archival.

Would it make sense to combine these transformations?

Yes, combining these transformations often makes sense. Compressing reduces storage and transfer costs, encrypting ensures data security, and reducing granularity enhances storage and processing efficiency. For instance, reducing data first can make compression more effective, and encrypting after compression secures the compressed data. Together, these methods

optimize storage, maintain security, and improve processing efficiency. Hence, a combined approach can provide comprehensive benefits for data archival.

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.

Created code for publisher, receiver and creating the bucket

C. [MUST] Store Data to GCP Storage Bucket

Modify archiver.py to store all received data to a [GCP Storage Bucket](#). 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?

There were 295852 records for 24JAN2023. 71088 Kibs is the bucket space utilized to store 1 day's breadcrumb of 100 vehicles.

D. [SHOULD] Compress

Modify archiver.py to compress the data before it stores the data to the storage bucket. Use [zlib compression](#) which is provided by default by python. How large is the archived data compared to the original?

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

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](#)

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?

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.