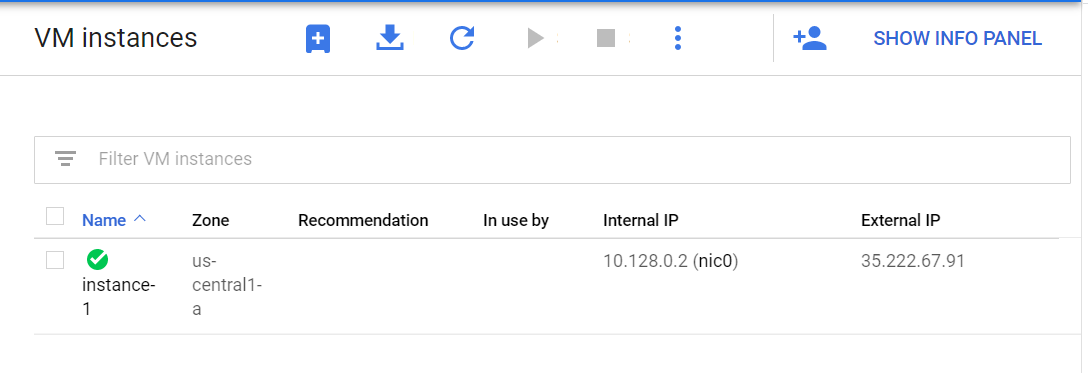
Sreeti Ravi

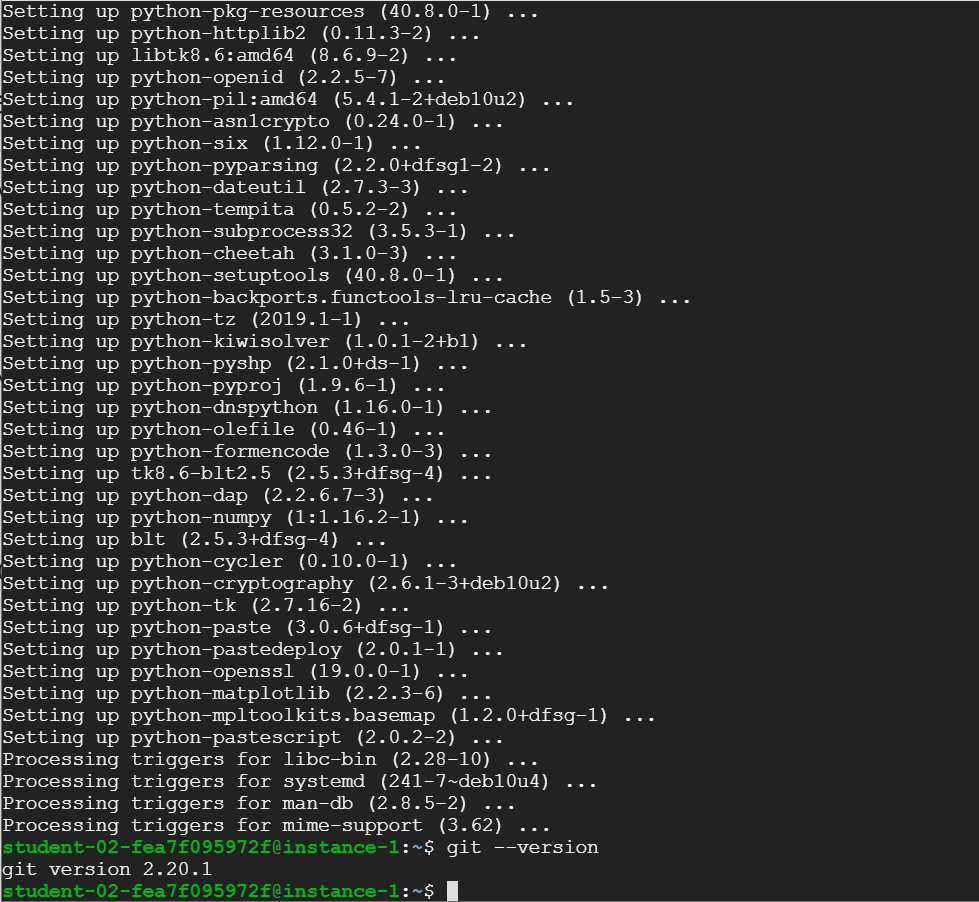
Lifecycle and Pipelines

Rent-A-VM to Process Earthquake Data

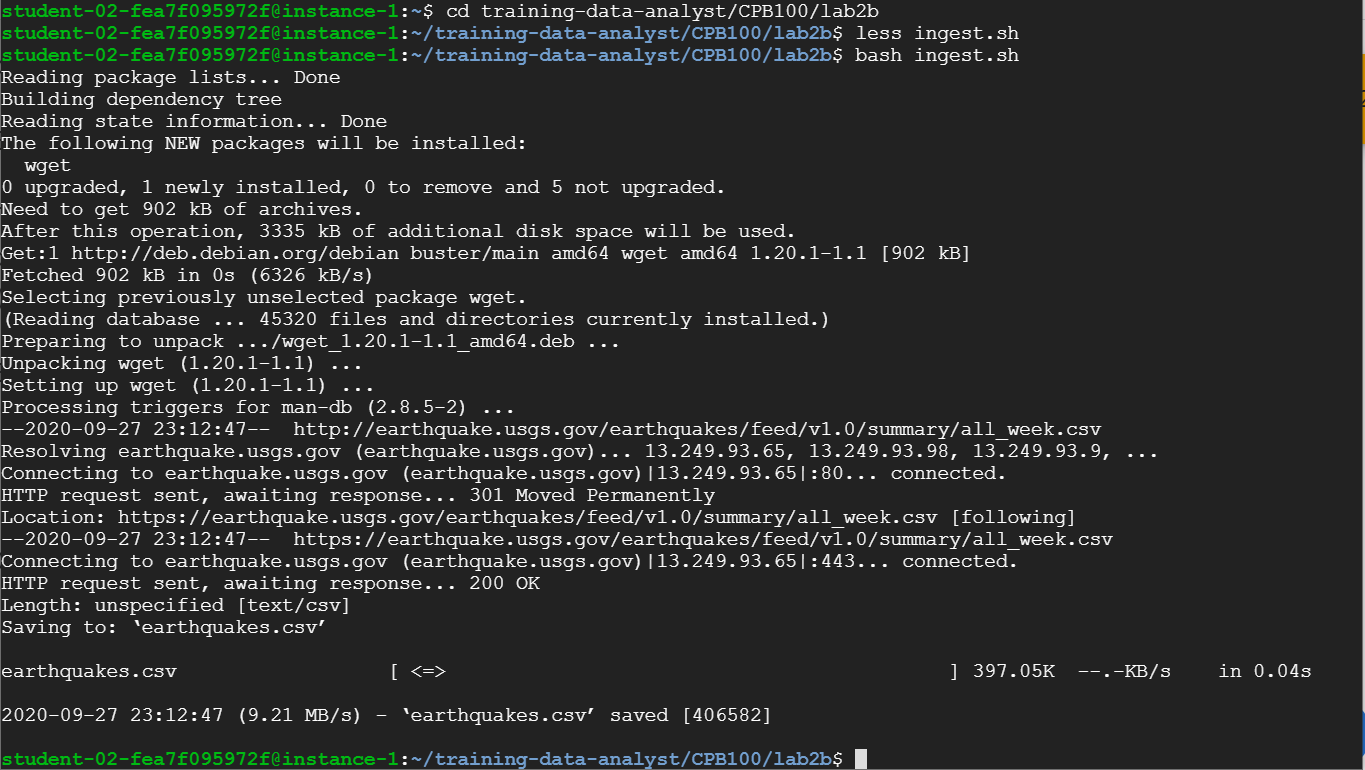
1. Create Compute Engine instance with the necessary API access



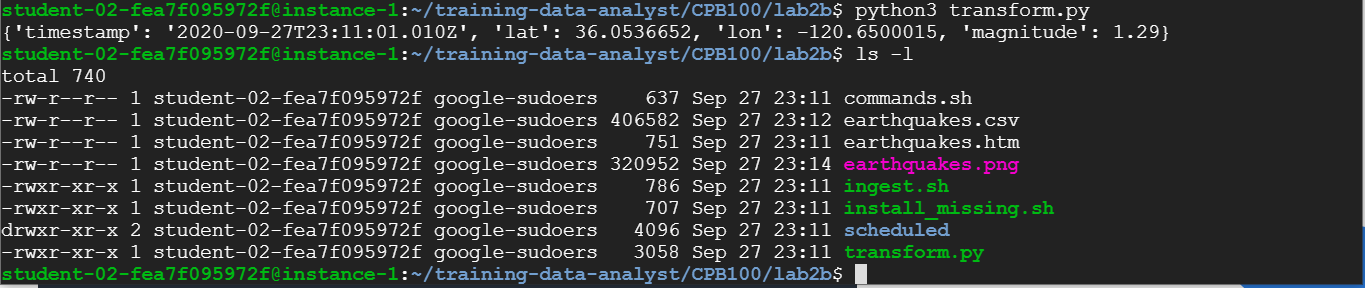
1. SSH into the instance
2. Install software



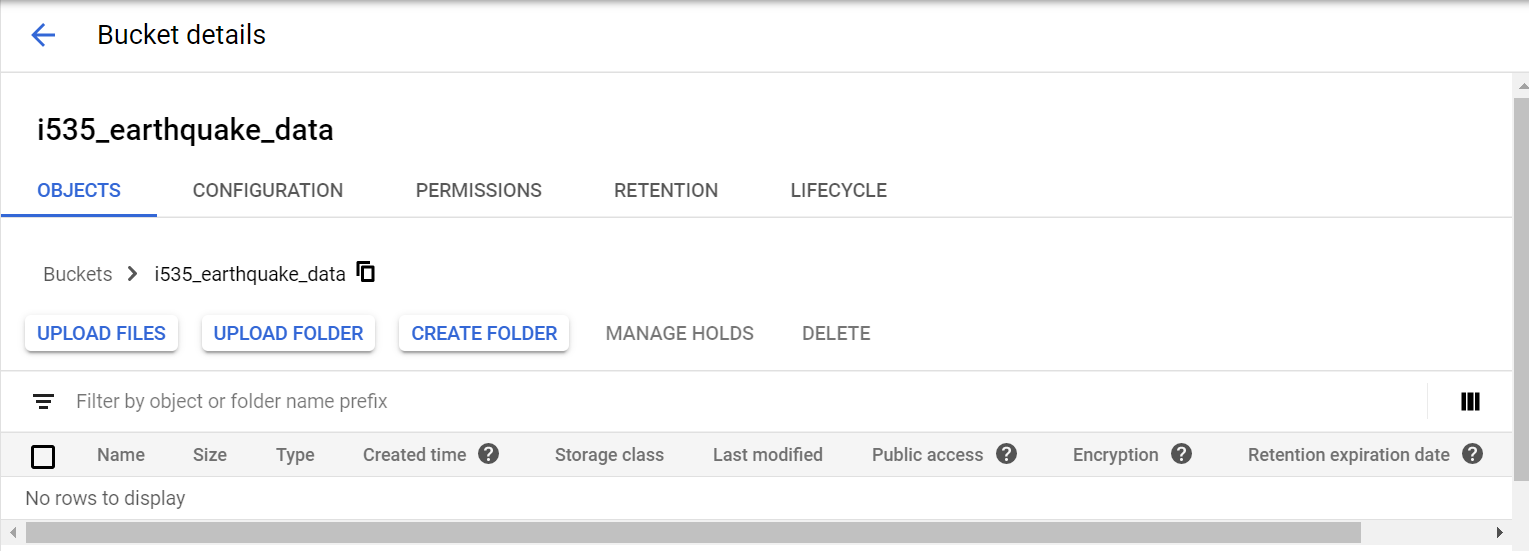
1. Ingest USGS data



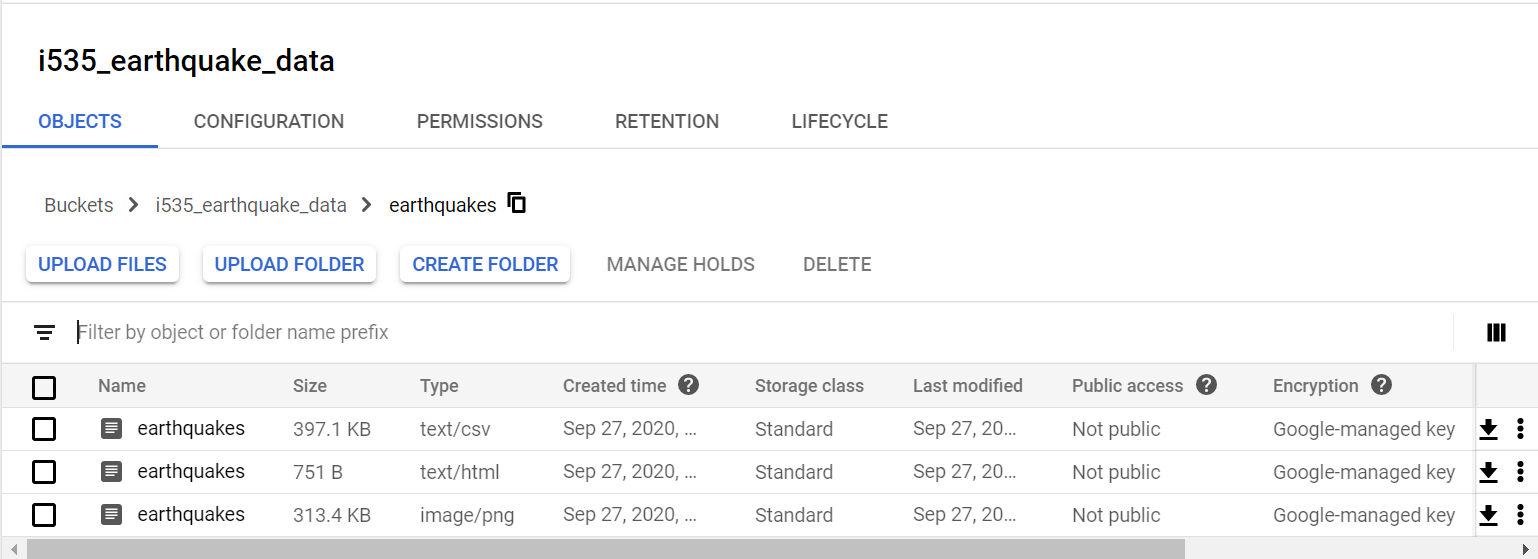
1. Transform the data



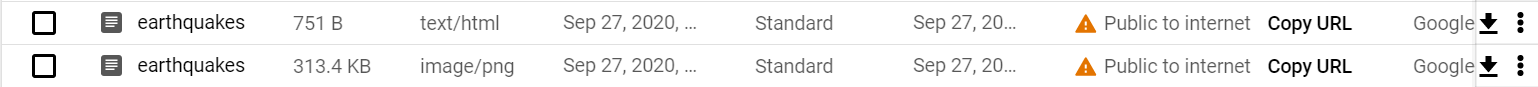
1. Create a Cloud Storage bucket

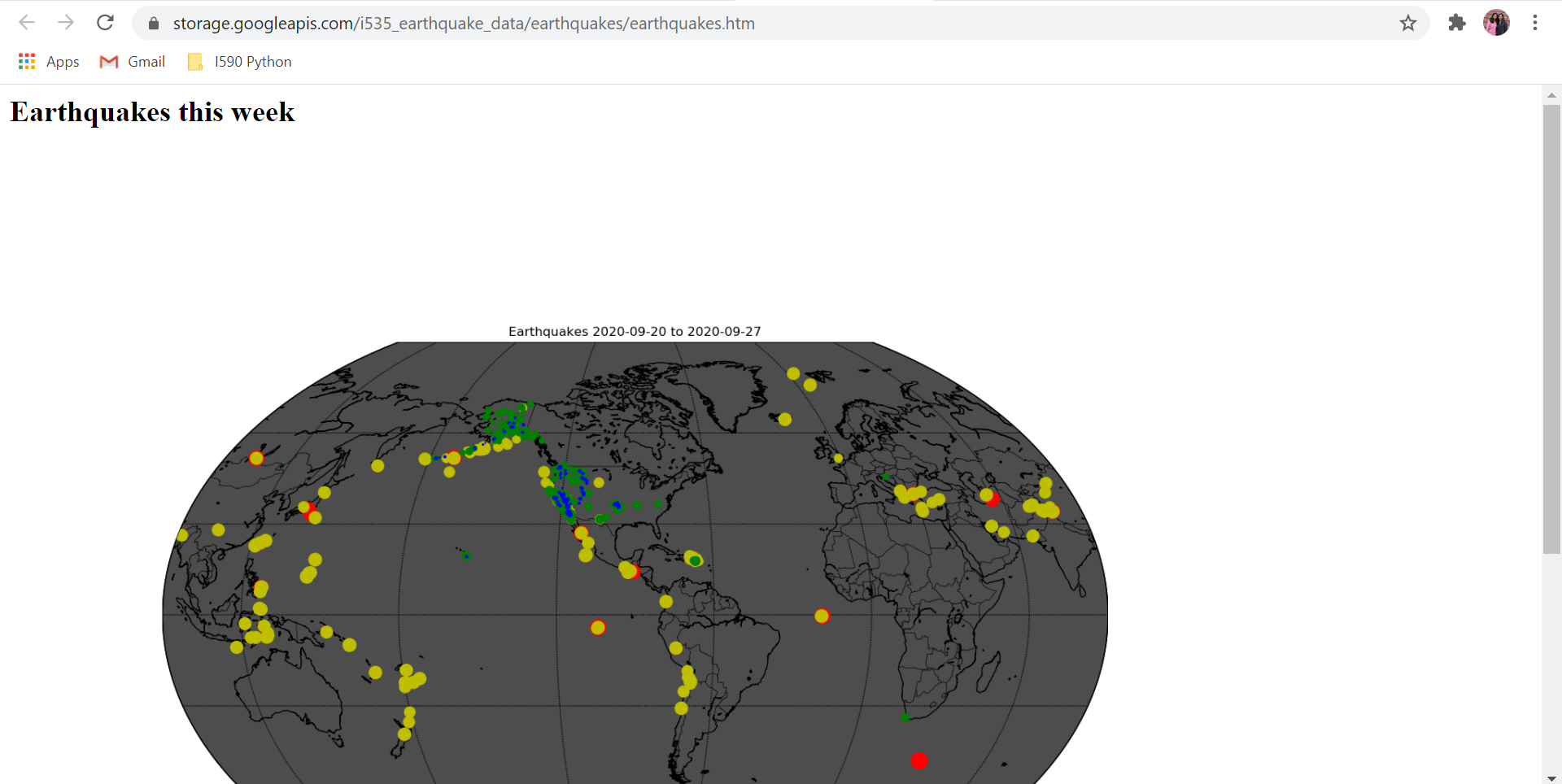


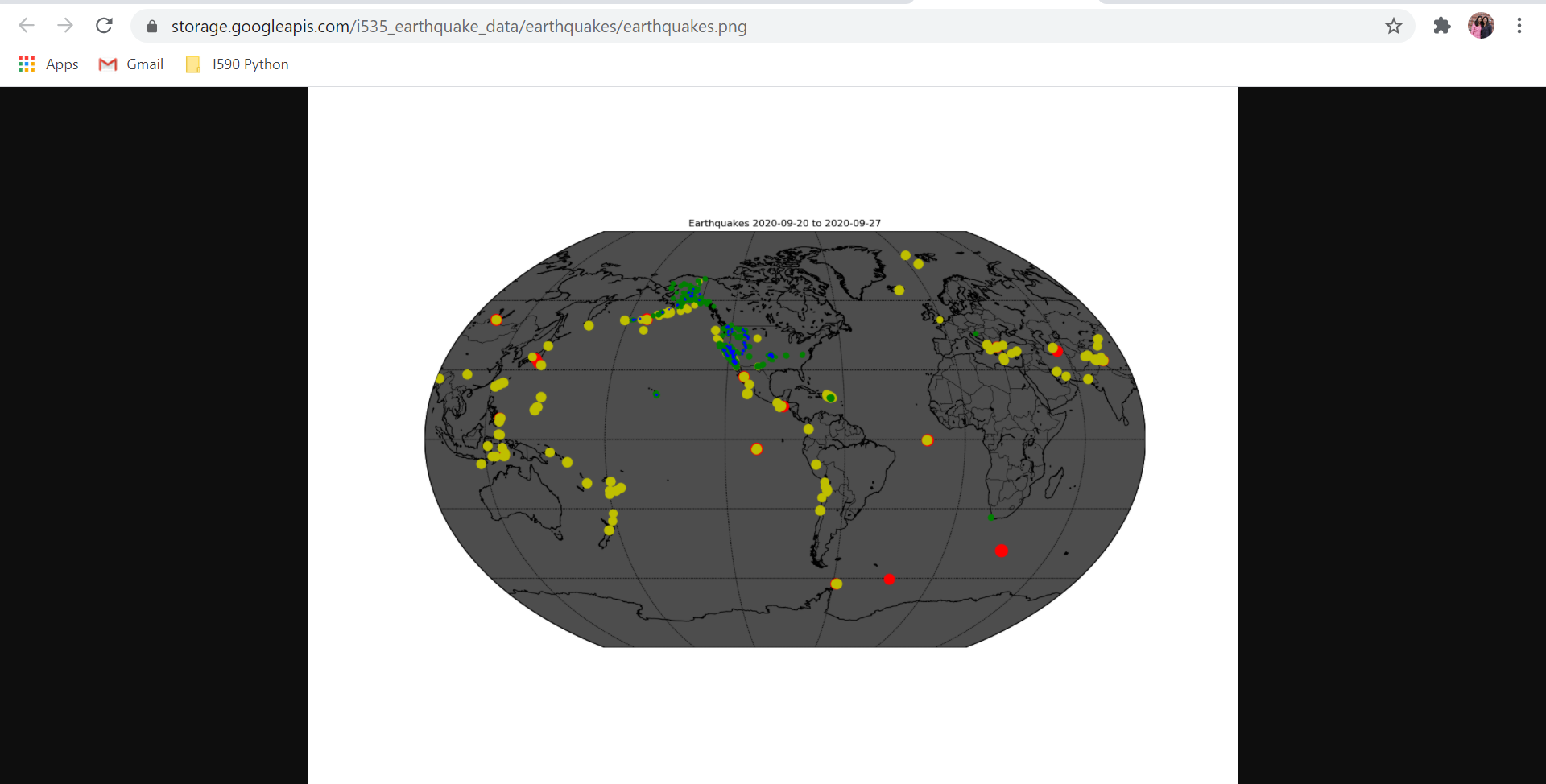
1. Store data



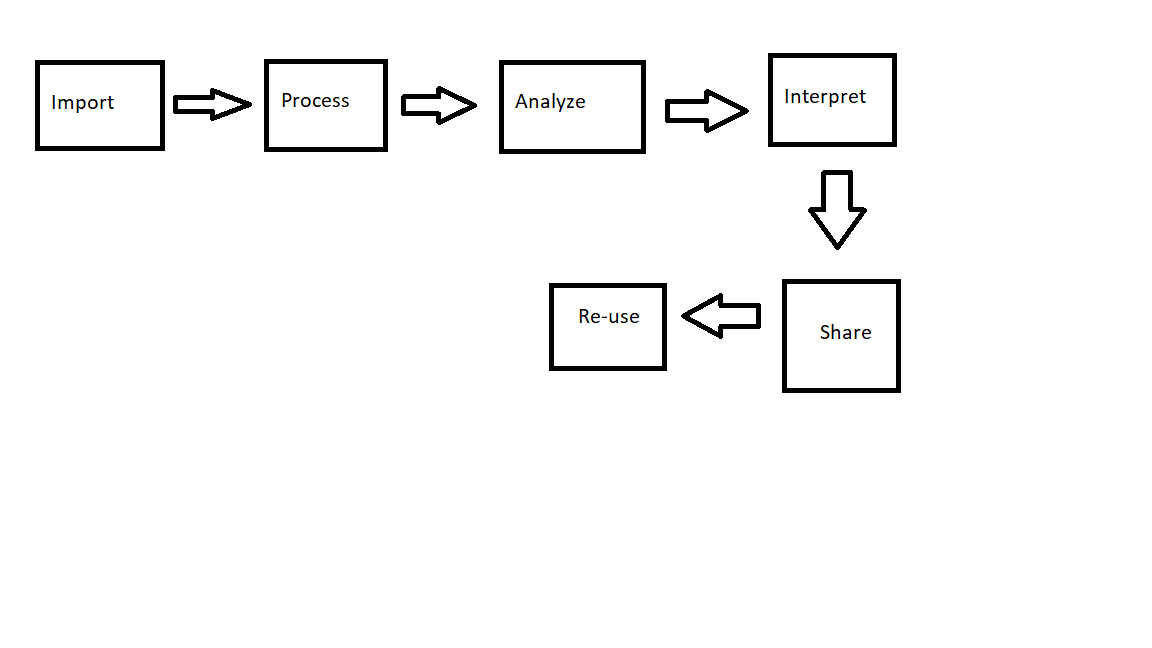
1. Publish Cloud Storage files to web







The data source for this lab was the US Geological Survey. This lab required a download of a dataset of earthquakes in the past 7 days. The original data is from the US Geological Survey website where the data is extracted as a CSV file. In the lab, we copied the files to a bucket we created to store the original and transformed data in Cloud Storage. The earthquake data from USGS is parsed and the data includes timestamp, latitude, longitude, and magnitude. Based on magnitude size, different control marker colors are assigned. After the base map is drawn the stronger earthquakes are plotted first. The larger earthquakes have larger circles and weaker earthquakes are plotted on top of the larger earthquakes. To do all of this, Python libraries, such as numpy, matplotlib and io are used.



This pipeline includes imported data from the USGS website. That data is then transformed with Python. It’s analyzed and interpreted by plotting the graph using basemap. It is stored in the Google Cloud Storage and shared from Google Cloud. I think validation could be added to verify that the data is correct. Validation is an important part of any project that contains data.