1 Introduction

This document details the workflow for automating the retrieval, processing, and storage of NAIP (National Agriculture Imagery Program) data using AWS S3. The process involves downloading relevant TIFF files, handling duplicate images, extracting bounding box data, and storing outputs in an organized manner.

2 Workflow Overview

The workflow follows these steps:

1. Extract relevant NAIP tile information from a shapefile.
2. Identify and download the corresponding TIFF files from AWS S3.
3. Handle duplicate APFONAME tiles by differentiating based on orientation (SE, NE, SW, NW).
4. Extract bounding box information from TIFF images.
5. Save processed files and metadata back into AWS S3.

3 NAIP Data Retrieval

3.1 Extracting Relevant Tiles

The script `NAIP_grab.py` downloads a shapefile from S3 containing city tile boundaries and extracts all unique APFONAME values. These values are saved in a CSV file for reference when filtering TIFF files.

3.2 Downloading TIFF Files

Using the APFONAME values, `NAIP_download.py` searches the NAIP dataset on S3 for matching TIFF files. The script:

- Reads APFONAME values from the CSV.
- Lists available TIFF files in NAIP's S3 bucket.
- Uses multi-threading to efficiently copy the files to the target S3 repository.
- Checks for existing files to prevent unnecessary downloads.

4 Handling Duplicate Tiles

The script `check_files.py` ensures duplicate APFONAME tiles are processed correctly by appending a numbered suffix to duplicate filenames. The differentiation is done using the spatial orientation in the filename (SE, NE, SW, NW).

5 Extracting TIFF Boundaries

The `Extract_outlines.py` script extracts bounding boxes from TIFF images stored in S3. This process:

- Downloads TIFF files in batches to reduce memory usage.
- Uses Rasterio to extract bounding box coordinates.
- Saves bounding boxes as a Shapefile for GIS visualization.

6 AWS S3 Integration

All processing occurs within AWS CloudShell and S3 to prevent local storage constraints. The pipeline:

1. Retrieves data from the NAIP S3 bucket.
2. Processes and renames files before saving them in the target bucket (`tgsp25`).
3. Stores metadata such as bounding box information for easy access in GIS applications.

7 Conclusion

This automated workflow streamlines the processing of high-resolution NAIP imagery. By leveraging AWS S3, multi-threading, and GIS tools, we efficiently extract and manage spatial data for further analysis.