Loading Geospatial data to Google BigQuery
Agenda

- What is Geospatial Data
- Why loading to BigQuery
- How to load
  - Two ways to load
  - Challenges
  - How beam/dataflow helps
- Geospatial Analysis
- Q&A
What is Geospatial Data

- A type of data that describes objects, events, or other features with a location on the earth.
- A powerful tool for understanding the world around us.
- Geospatial data analysis can help organizations make better decisions with a better understanding of the spatial relationships between features.
## Use Cases Enabled by Geospatial Data Analysis

<table>
<thead>
<tr>
<th>Industry</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insurance</td>
<td>Assess risk and analyze property damage claims, eg post severe weather events</td>
</tr>
<tr>
<td>Banking and Finance</td>
<td>Analyze market trends using weather or other geospatial data</td>
</tr>
<tr>
<td></td>
<td>Manage investments with geospatial analytics/Analyze portfolio risk</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>Infrastructure sustainability / Infrastructure location selection</td>
</tr>
<tr>
<td>Sustainability</td>
<td>Monitor and manage natural resources</td>
</tr>
<tr>
<td></td>
<td>Conduct environmental impact assessments/Analyze climate data.</td>
</tr>
<tr>
<td>Utilities</td>
<td>Manage infrastructure and assets.</td>
</tr>
<tr>
<td>Real estate</td>
<td>Analyze property values, conduct site selection, and manage property portfolios</td>
</tr>
<tr>
<td>Public Sector</td>
<td>Land management / Emergency response planning / Environmental monitoring</td>
</tr>
<tr>
<td>Transportation &amp; Logistics</td>
<td>Manage fleets / Improve customer service</td>
</tr>
<tr>
<td>Retail</td>
<td>Analyze customer behavior / Conduct market research / Optimize store locations</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Manage crop yields / Improve resource management</td>
</tr>
<tr>
<td>Healthcare &amp; Public Health</td>
<td>Analyze disease data / Conduct epidemiological research</td>
</tr>
</tbody>
</table>
Why Loading Geospatial Data to BigQuery

- BigQuery as a cloud-based data warehouse, enables low latency and large scale geospatial analysis
- Regardless of geospatial data types, there’s a way to ingest it, store it, analyze it on GCP.
- Enable integration with other data in BigQuery
- Enables geospatial data enabled machine learning models on GCP
BigQuery Spatial Functions

- `ST_CONTAINS(outside_geom, inside_geom)` -- true if outside completely surrounds inside
- `ST_INTERSECTS(a_geom, b_geom)` -- true if intersection of a and b is non-empty
- `ST_DISTANCE(a_geom, b_geom)` -- shortest distance between two geometries
- `ST_AREA(a_geom)` -- area in square meters
- `ST_LENGTH(a_geom)` -- length of line or perimeter geometry

https://cloud.google.com/bigquery/docs/reference/standard-sql/geography_functions
Types of Geospatial Data

Vector (Tabular)

- Point
- Linestring
- Polygon
- Multi-polygon
- Collections

Format: shape, kml, geojson, WKT ...

Raster (Imagery)

Format: GeoTiff ...
How to load - Challenges

Polygonize the high resolution image with many pixels in it is very compute intensive.

A ~15 meter ground resolution one degree tile image could have more than 10 million pixels.
Geotiff to Polygon
How to Load - 2 ways to load

- **Dataflow with GeoBeam library**
  - GeoBeam load shapefiles very effectively.
  - Works well with small to medium size images (< ~a few hundreds MB).
  - Great examples
  - Runs fine in the local runner

- **Dataflow running gdal commands**
  - Can handle complex data processing (retile raster, reproject, polygonizer etc)
  - Better parallelism and memory consumption


https://gdal.org/programs/gdal_polygonize.html
Steps of Loading Shape File with Geobeam

1. Read zip file
2. Geobeam shapefile source
3. Orient polygon
4. Format record
5. Data transformation
6. Enrich metadata
7. Write to BigQuery
Steps of Loading GeoTiff Files with GDAL

- `gdal_polygonize.py nass.tif -b 1 -f "GeoJSON" nass.geojson`
- `jq -c '.features[]' nass.geojson > nass-nl.geojson`
- `bq load --source_format=NEWLINE_DELIMITED_JSON --json_extension=GEOJSON -autodetect=true geodata.ny_nass nass-nl.geojson`
How beam/dataflow helps

- GeoBeam, built on top of Apache Beam
  - Provides a set of FileBasedSource classes that make it easy to read, process, and write geospatial data
  - Provides a set of helpful Apache Beam transforms and utilities that make it easier to process GIS data in Dataflow pipelines.
- Dataflow customized container worker for
  - Easy deployment of 3rd party libraries as you can specify them in build files and dataflow will handle the deployment.
- Use Dataflow as compute and run command lines allows to focus on data handling and not underneath beam complexity
- Easy parallel processing with auto scaling, composer, and Thread Pool Executor while converting geotiff to geojson
- Integrated with GCP for monitoring logging
Reference Architecture

GCP Data Ingestion

Cloud Storage

Raster Files

Vector Files

Ingestion
Cloud Dataproc

BigQuery

Data Curation

SQL

Curated Layer
Views /

Visualization

Orchestration - Cloud Composer

Logging & Monitoring - Cloud Logs + BigQuery
Demo

Layers

- POLYGON crop land
- POLYGON ny tree density
- POLYGON zipcode
- TILES ny tree census
QUESTIONS?