Key Considerations for Storing Data on the Cloud

Tom Augspurger

NOAA EDMW 2022

September 2022

Cloud-native Geospatial Principals

- Using principals from a series of blogposts by Chris Holmes
- You have access to *all* the data (along with everyone else)
- You have access to scalable compute *located next to the data*
- Data can be easily visualized on a map

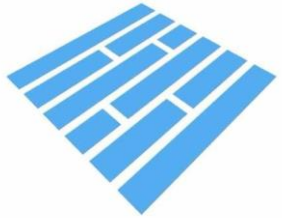
https://medium.com/planet-stories/cloud-native-geospatial-part-1-basic-assumptions-and-workflows-aa67b6156b53

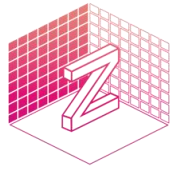
Cloud-native Geospatial Consequences

- You have access to *all* the data (along with everyone else)
  - No single hard drive can store all the data
  - Use a service like Azure Blob Storage, S3, GCS
  - No need to create “dark replicas” of these datasets
  - Need cloud-optimized file formats
- You have access to scalable compute *located next to the data*
  - Very valuable to “move the compute to the data”
  - Compute on the data in-place

Cloud-friendly file formats

- Metadata compactly in a convenient place
- Somehow support chunking / tiling
- Combine with HTTP range requests
  - Give me bytes 100-200 from the file at https://...





Cloud-friendly *clients*

> One of the quiet secrets of the “cloud optimized” geospatial world is that, while all the attention is placed on the formats, the actual **really really hard** part is writing the clients that can efficiently make use of the carefully organized bytes.

- Paul Ramsey (http://blog.cleverelephant.ca/2022/04/coshp.html via Pete Gadomski)

Thanks!

https://planetarycomputer.microsoft.com

https://github.com/TomAugspurger/noaa-edmw-2022

taugspurger@microsoft.com