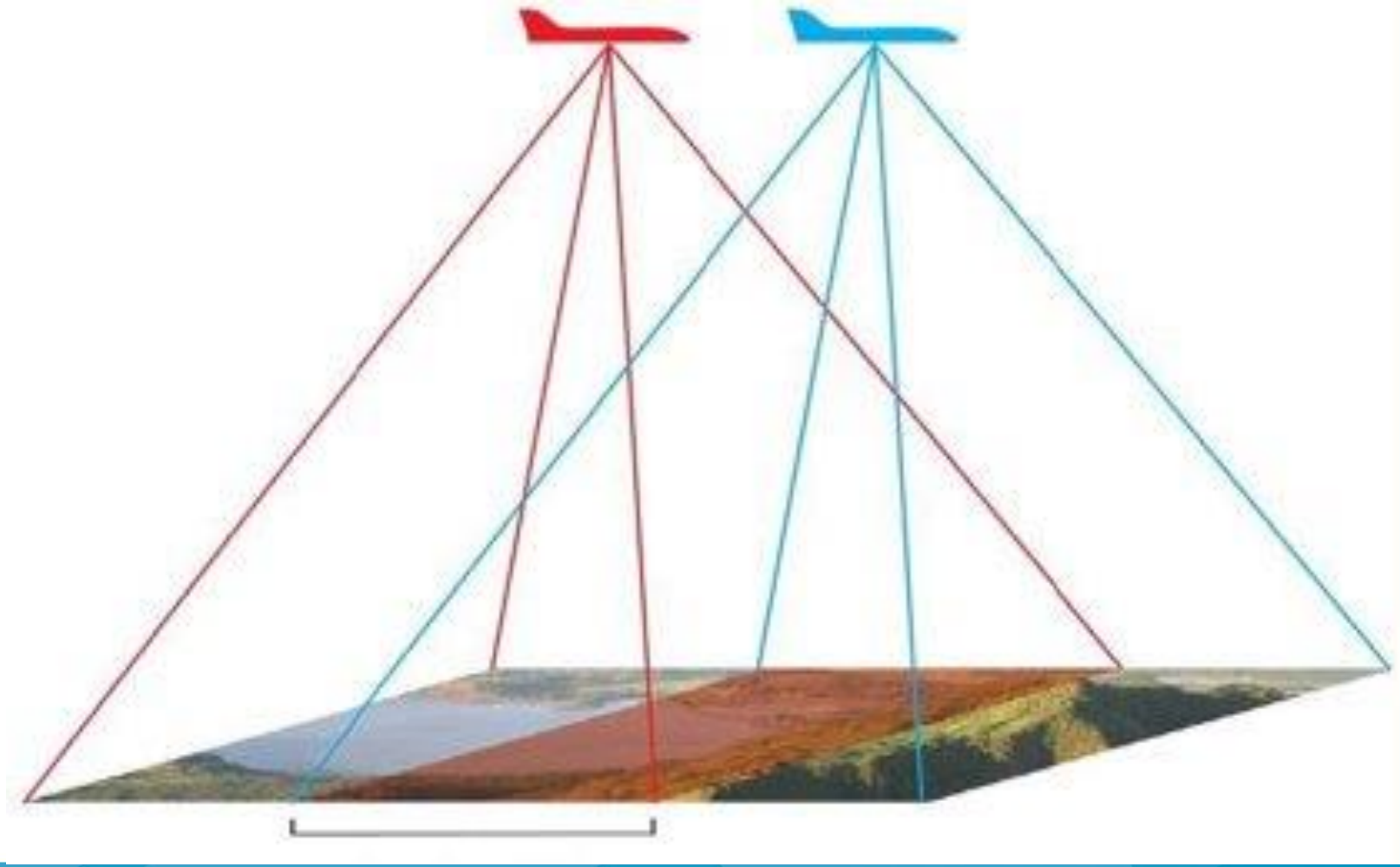
kadaster

Managing large datasets in the cloud




_05891 : M:\DP513\Images_input\69\233_05891.tif
_05892 : M:\DP513\Images_input\69\233_05892.tif

Some statistics

Product	Bron
Total storage used	2.5 Petabyte~ eg, 2560 Terrabyte~
Yearly growth	250TB~
Webservices requests	600 requests per second~

Note

Everything I tell here is not GIS specific and is generally applicable.

Old approach

- Store files of a single year on a NAS
- Everything else on loose harddrives
  - Finding archived images takes time
  - Harddrives eventually expire

Cloud migration

- Use objectstorage
- Ability to temporary rent machines to process data
- Preferably with the same provider, storage close to compute

Our approach

- Generic data portal for managing data
  - Allows for sending and receiving files via a 'wetransfer' like website
  - Collect metadata on each file
  - Start your favorite workflow managment software to manipulate the data
  - Make a distinction between temporary and permanent data
  - As for the the permanent data...
    - Set archiving policies
    - Backup policies
    - Storage tiering policies

Archiving Policies

- How long must I retain the data
- Which laws are applicable to the data
- Notification for removal

Backup Policies

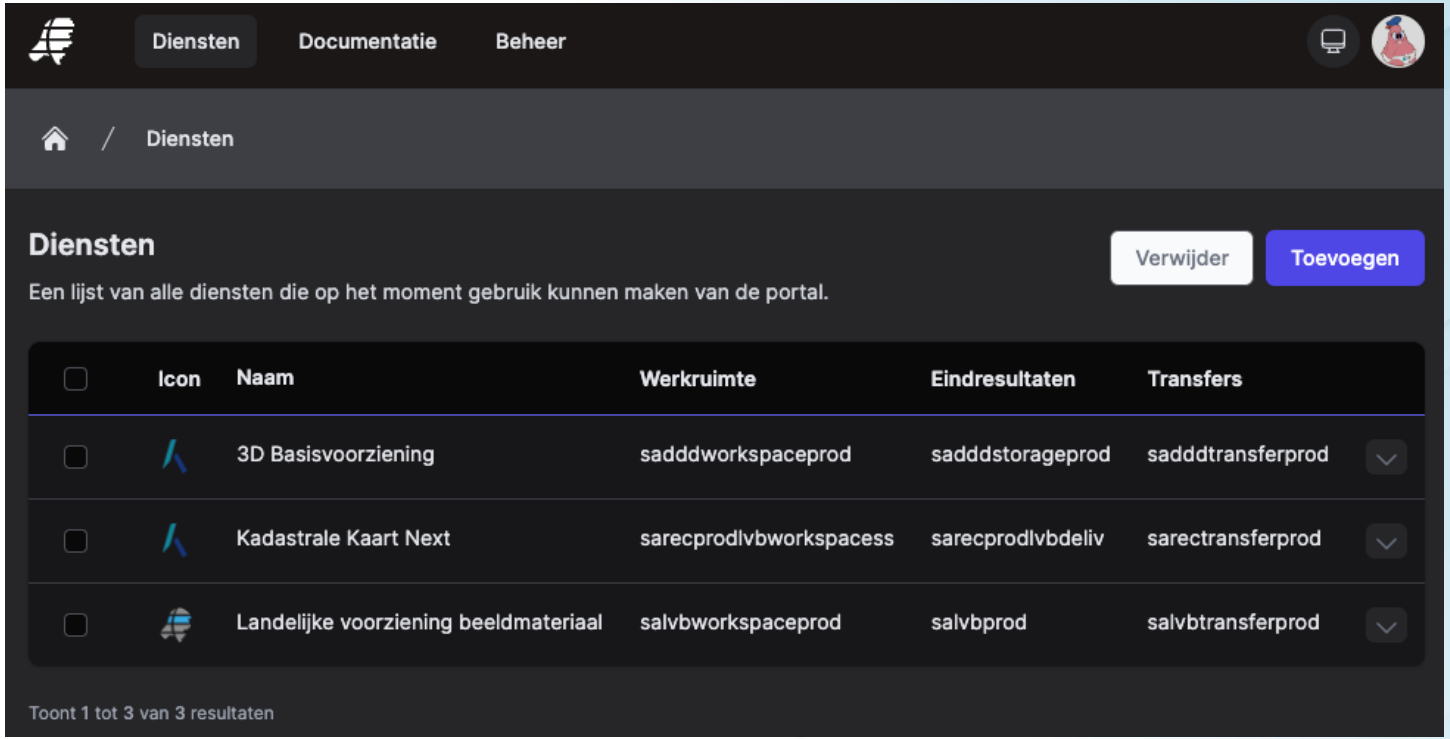
- Make the data immutable.
- Copy data to a different object storage provider.
- Copies on both sides will remain the same, we do not use versioning or incremental backups.
- Data must include metadata.
- Offsite backup storage provider must be fully independent

Storage tiering policies

- Most object storage providers allow for storage tiering (Using Azure blob as example)

Tier	Storage costs	Reading costs
Hot	Expensive	Cheap
Cool	Moderate	Moderate
Archive	Cheap	Expensive

Small demonstration



kadaster


Bonduelle