Dataset production

An agent turns a dataset request into a reviewed recipe, the cluster runs it in parallel, and checks against the live catalog decide whether it merges.

REQUEST TO PUBLISHED COLLECTION

1 GITHUB ISSUE

Dataset request

States the source, extent, H3 resolution and target bucket.

`geo-workflows`

2 AGENT ALWAYS ON

Recipe from geo-ingest

The agent reads the source and the catalog, then generates the Job YAML with geo-ingest.

`geo-ingest`

3 PULL REQUEST

Reviewed recipe

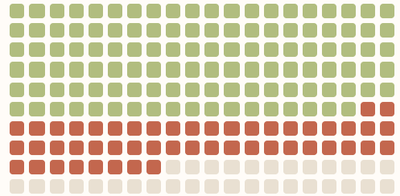
One PR per build, reviewed before anything runs on the cluster.

`geo-workflows`

4 PARALLEL JOBS

Overnight build

Indexed Jobs or Armada, one pod per chunk.



done running queued

5 BUCKETS

STAC and S3

GeoParquet, PMTiles, COG and H3 hex parquet, with a STAC collection.

`public-<name>`

6 · verify-stac checks the live catalog through geo-query · a failure goes back to the agent · a pass merges

The always-on agent

A headless agent works through the request queue and opens the PRs. It runs as its own ServiceAccount that can manage Jobs and read their logs, and nothing else.

Scheduled updates

CronJobs re-run builds for live feeds daily, monthly or yearly.

`coral-reef-watch ×3`

Backups

Weekly copies off NRP to MinIO, and a licence-checked public mirror on source.coop.

`minio-sync` `source-sync`