GSKY – a scalable Geospatial data server
Major Challenges for users

- Datasets are too large/cumbersome to move, costly to download, but more convenient to analyse on desktop
- Very large number of files in major spatial timeseries datasets are very difficult to wrangle
- Data files generally need to be converted/transformed to interoperate to compare/analyse
- Both HPC, data and coding skills required for manipulating the data are barriers to analyzing data
- Geospatial users would prefer to use standard tools and known protocols for important reference datasets

Number of files growing rapidly:
3x every 2 years

For managed Datasets, number of files:
2018 - 330M
2020 – 1T
GSKY: A Scalable, Distributed Geospatial Data Server

**Request**

**GSKY**

**Response**

**Features**

- Distributed
- Scalable
- Concurrent

**INPUT**

Geospatial request

**OUTPUT**

User’s client

© NCI Australia 2019
The Apollo moon landings were supported by a computer interface known as DSKY. Astronauts could input data and commands into the keypad and see the results returned on an electronic display. In much the same way, GSKY is an interface that allows human manipulation of deeply buried geospatial data. Using GSKY, a user can make complex requests and see the results in their web browser in near real-time.

**GSKY** cannot navigate its users to the moon – it can, however, help us understand it.
GSKY responds to Open Geospatial Consortium (OGC) API over http protocol:
• Web Map Service (for displaying the images on the map server)
• Web Coverage Service (for delivering the actual data as “coverages” - independent of the underlying storage format or files)
• Web Processing service

GSKY allows:
• Performant aggregations, subsetting, subsampling, polygon/pencil/pixel drills
• Execution of on-the-fly data transformations, re-projections and other algorithms
• Deal with very large files

GSKY is implemented using
• Rich metadata server for data query e.g., spatial, temporal, other physical variables
• Clustered backend workers – high performance I/O and scale-out server-side compute
**MAS**: Integrated with GSKY and provides a backend abstraction over the data

- It identifies individual data objects (datasets, variables, spatial and temporal extents).
- Can be shared by different geospatial collections or by splitting collections into non-overlapping geographical extents.

**MAS is able to process queries in milliseconds, even for the ones comprising large spatial areas or temporal ranges, which often results in thousands of data objects being identified.**

MAS is also used by other projects (e.g., our climate data services)
GSKY High Level Architecture

- Based on “flow-based programming”.
- Data transformed by connected processes, forming a Directed Acyclic Graph (DAG).

OGC Request

GSKY + MAS + Data + ...

OGC Response
GSKY High Level Architecture

MAS

Geospatial Index

GeoCrawl

/g/data

Worker 1

Worker 2

...

Worker n

OWS

GSKY

© NCI Australia 2019
GSKY file handling of common large Earth Observation datasets
Amount of data is scaling up with more modern satellite instruments

**MODIS (CSIRO & TERN)**
MODIS data from 2001 – present, with 4 timestamps per month and 1k files per timestamp.
GSKY is handling approx. 900k MODIS files

**Landsat-8 dataset (DEA)**
Landsat 8 dataset has data from 2013 – present, approximately 3k files per timestamp.
GSKY is handling approx. 7 million files.

**Sentinel 2 ARD (DEA)**
Approx 1400 available days (2015-present) and about 12k files per day for this dataset.
GSKY is handling approx. 17 million files.
ASTER geophysics data
Problem handling large variables
- 62 GBytes

Using:
• default TDS -> times out
• Tuned TDS -> 90s
• GSKY -> 1s

Secret: down-sampling chunked data I/O
**Digital Earth Australia (DEA) Geoscience Earth Observations**, which include the following products of the Landsat 5, 7 and 8 satellite missions:

- Surface reflectance (NBAR/NBART true and false colour)
- Terrain corrected surface reflectance geometric median (geomedian)
- Intertidal Extents Models (ITEM)
- High and Low Tide Composites (HLTC)
- Water Observations from Space (WOfS)
- Sentinel 2 Analysis Ready Data (Beta)
- Blended service (landsat + sentinel)

**GEOGLAM**, the GEO Global Agricultural Monitoring initiative, which include the following products:

- MODIS Total Vegetation Cover v3.1 (8-day and Monthly)
- MODIS Total Vegetation Cover Anomaly v3.1 (Monthly)
- MODIS Total Vegetation Cover Decile v3.1 (Monthly)
- MODIS Vegetation Fractional Cover 8-day v3.1 (8-day and Monthly)
- CHIRPS Precipitation v2.0 (Monthly)
You can browse and search NCI’s full collection by going to the [Terria Map](http://terriamap.com) or [National Map](http://nationalmap.com) websites.

To view the DEA or GEOGLAM collections, click on Add Data -> My Data -> Add Web Data and enter the following URLs respectively:

http://gsky.nci.org.au/ows/geoglam

We are continually adding new datasets to GSKY.
Summary: NCI’s GSKY – A Scalable Geospatial Data Server

GSKY:
• Currently WMS, WCS, WPS for earth obs, and geophysics datasets
• Open Source: http://github.com/nci/gsky

For more information, see:
http://gsky.nci.org.au/
https://gsky.readthedocs.io/

Quick look:
Add data -> My Data -> Add Web Data
URLs: http://gsky.nci.org.au/ows/dea
http://gsky.nci.org.au/ows/geoglam

© NCI Australia 2019
What can you do with GSKY to start with?

Live Demo:

1. Add geospatial data as a single layer.
2. Provide access to data subsets (e.g. groups of variables of interest) as data layer.
3. Serve multiple layers simultaneously overlaying on top of each other for comparative analysis.
4. Provide the same layer for different points of time allowing cross-sectional analysis.
5. Allow users to download the raw data corresponding to a user-supplied bounding box.

**NationalMap** is an online map-based tool to allow easy access to spatial data from Australian government agencies. It was an initiative of the Department of Communications and the Arts (DCA) now currently managed by the Digital Transformation Agency (DTA) and the software has been developed by Data61 working closely with the DCA, Geoscience Australia and other government agencies.


Landsat8 - Lake Eyre, 05/05/2019
[https://www.nationalmap.gov.au/#share=s-2KVUKCRx57sQOjTjlhEOkmvyYDL](https://www.nationalmap.gov.au/#share=s-2KVUKCRx57sQOjTjlhEOkmvyYDL)
Lon = 137.14 Lat=-28.5
location = Lake Eyre
same layer (WOfS), different style

HLTC, Sydney, 1/11/2007 - now (optional)
location = Sydney
same layer, different date
TerriaJS is an open-source framework for web-based geospatial catalog explorers.

http://terria-cube.terria.io

Cambridge Gulf, 1/11/2016
http://terria-cube.terria.io/#share=s-788ZXugGlzQ6XtJj20GnZAk4G2

location = Cambridge Gulf
different layer, same date
High Tide Composition, and Low Tide Composite
bushfire example, 12 - 13 Sep, 2019
http://terria-cube.terria.io/#share=s-IbRTW2mmLsclY4mMtA2tHk1SSoB

Lon = 152.35 Lat=-30.08
location = Timbarra National Park
layer = Multi-sensor (Landsat and Sentinel 2) surface reflectance false colour (Beta)

different date: 12/09/2019 and 13/09/2019
Load Global rainfall data and doing analysis using “analysis tools”.

http://map.geo-rapp.org

You can follow this document: https://gsky.readthedocs.io/en/latest/_userguide/gsky_through_the_web.html