CSIRO and the Open Data Cube

Dr Robert Woodcock , Matt Paget, Peter Wang, Alex Held

CSIRO
www.csiro.au

Overview

- The challenge – The Earth Observation Data Deluge
  - Integrated science needs
  - Data volume, rate of growth and variety
  - User expectations
- Meeting the challenge, responding to change
  - The EOI ecosystem
  - Community and capacity building
- The Australian Geoscience Data Cube
  - What it is / how it works

Australian Government
Geoscience Australia

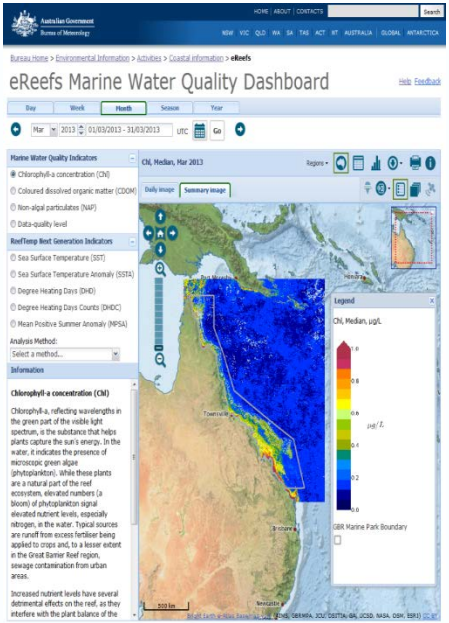
CSIRO

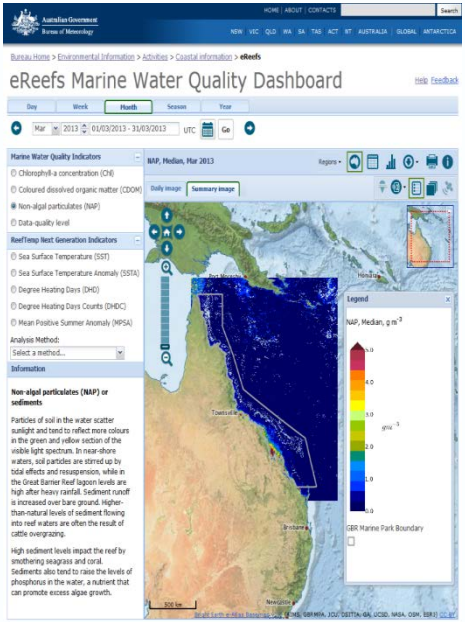
NCI
NATIONAL COMPUTATIONAL INFRASTRUCTURE

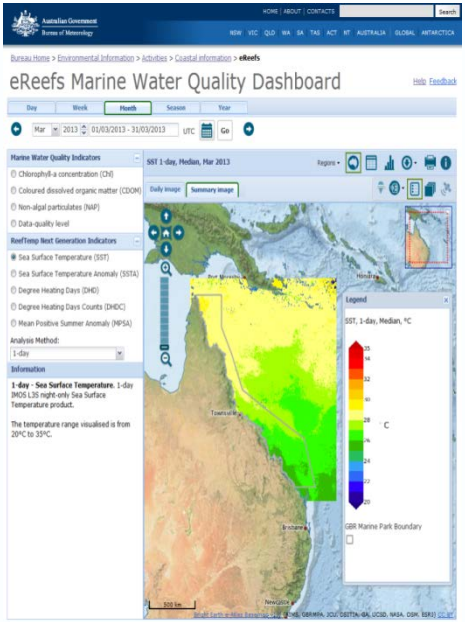
CSIRO

eReefs Marine Water Quality Dashboard

http://www.bom.gov.au/marinewaterquality/

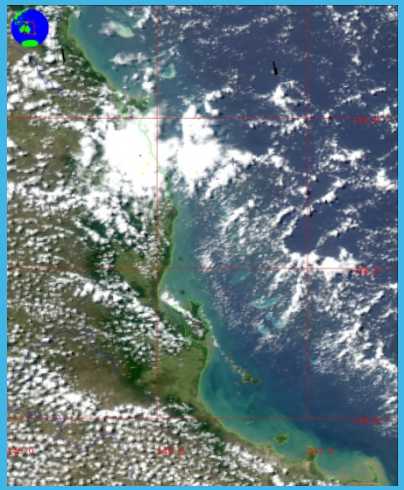


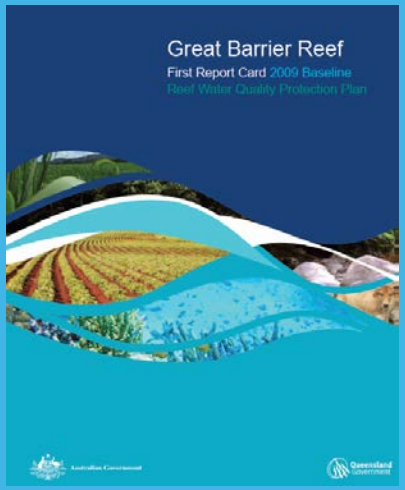




eReefs is a collaboration between:
GREAT BARRIER REEF foundation
Australian Government Bureau of Meteorology
CSIRO
Australian Government
AUSTRALIAN INSTITUTE OF MARINE SCIENCE
Queensland Government

Supported by funding from:
Australian Government
Queensland Government
SCIENCE AND INDUSTRY ENDOWMENT FUND
bhpbilliton resourcing the future
BMA BHP Billiton Mitsubishi Alliance





Collaborators:

Thomas Schroeder, Arnold Dekker, David Blondeau-Patissier, Kadija Oubelkheir, Nagur Cherukuru, Lesley Clementson, Paul Daniel, Janet Anstee, Britta Schaffelke, Michelle Devlin

Australian Government

Queensland Government

Reef Water Quality Protection Plan

You are here:
Home>Measuring success>Report cards>Report Card 2012 and 2013

Report Card 2012 and 2013

This report card measures progress from the 2009 baseline (http://www.reefplan.qld.gov.au/measuring-success/report-cards/first-report-card.aspx) towards Reef Water Quality Protection Plan 2009 (PDF, 2.39 MB) (http://www.reefplan.qld.gov.au/about/assets/reefplan-2009.pdf) (Reef Plan) targets. It assesses the combined results of all Reef Plan actions up to June 2013.

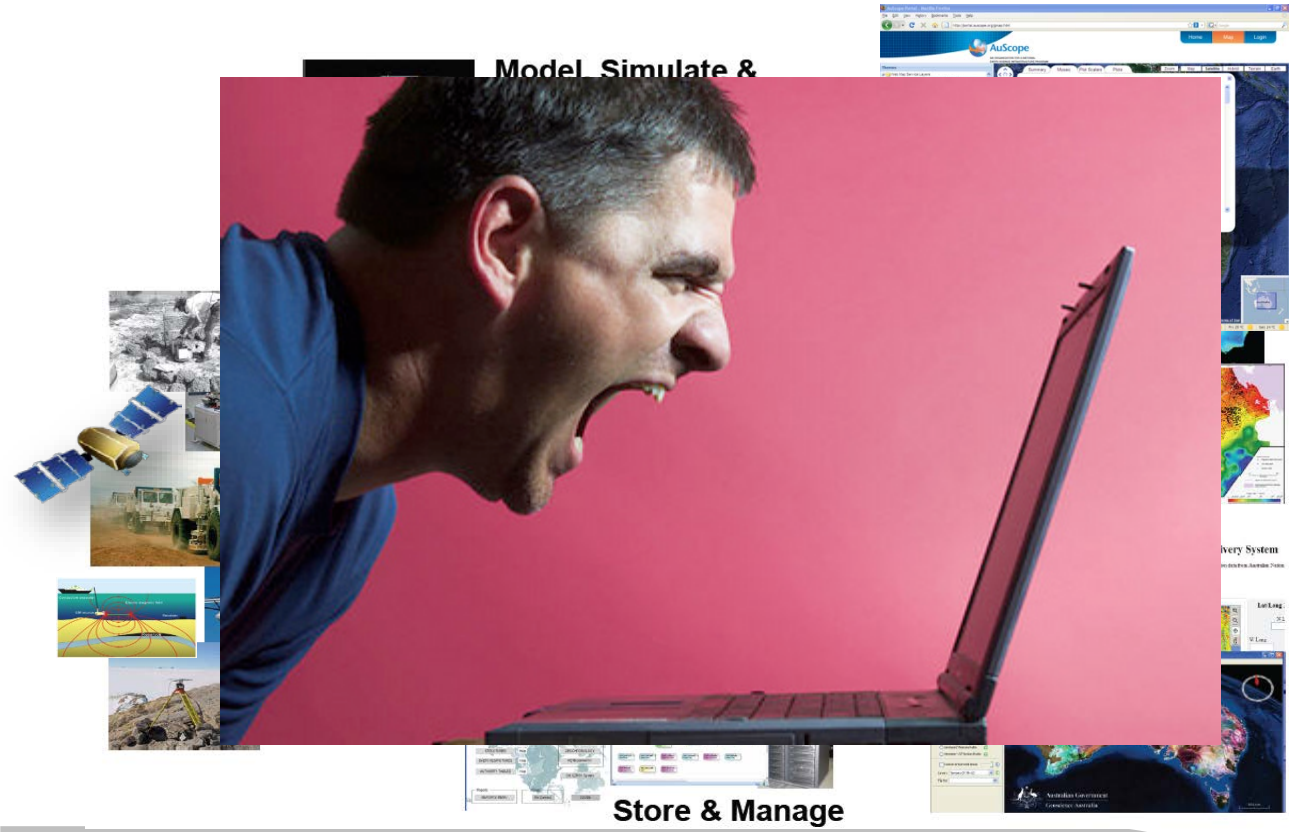
Key findings

- Results show modelled annual average pollutant loads entering the reef have significantly reduced, indicating the immediate 2013 goal of halting and reversing the decline in the quality of water entering the Great Barrier Reef has been met.
- The adoption of improved land management practices and resulting water quality improvements are an encouraging sign of progress towards the long-term goal of ensuring that by 2020 the quality of water entering the reef from adjacent catchments has no detrimental impact on the health and resilience of the Great Barrier Reef.
- Landholders have made major progress in adopting improved land management practices across the Great Barrier Reef catchment. Forty-nine per cent of sugarcane growers, 59 per cent of horticulture producers and 30 per cent of graziers adopted improved management practices by June 2013. The Burdekin and Burnett Mary regions recorded the highest levels of adoption (55 per cent) in the sugarcane industry. Two regions exceeded the grazing target of 50 per cent adoption — Mackay Whitsunday (69 per cent) and Burdekin (54 per cent).
- Progress towards the sediment target was rated very good, with the estimated annual average sediment load reducing by 11 per cent overall. The greatest reduction was in the Burdekin

...estimated annual average sediment load reducing by 11 per cent overall

CSIRO

Recognise the complete picture....



The growing expectations of users

- Increasing need for:
  - richer content
  - personalised tools
  - better filtering
  - diverse content from multiple sources
  - Anywhere, anytime, any device, on-demand

How do I connect

my tools to

your content?

Acquisition to Products

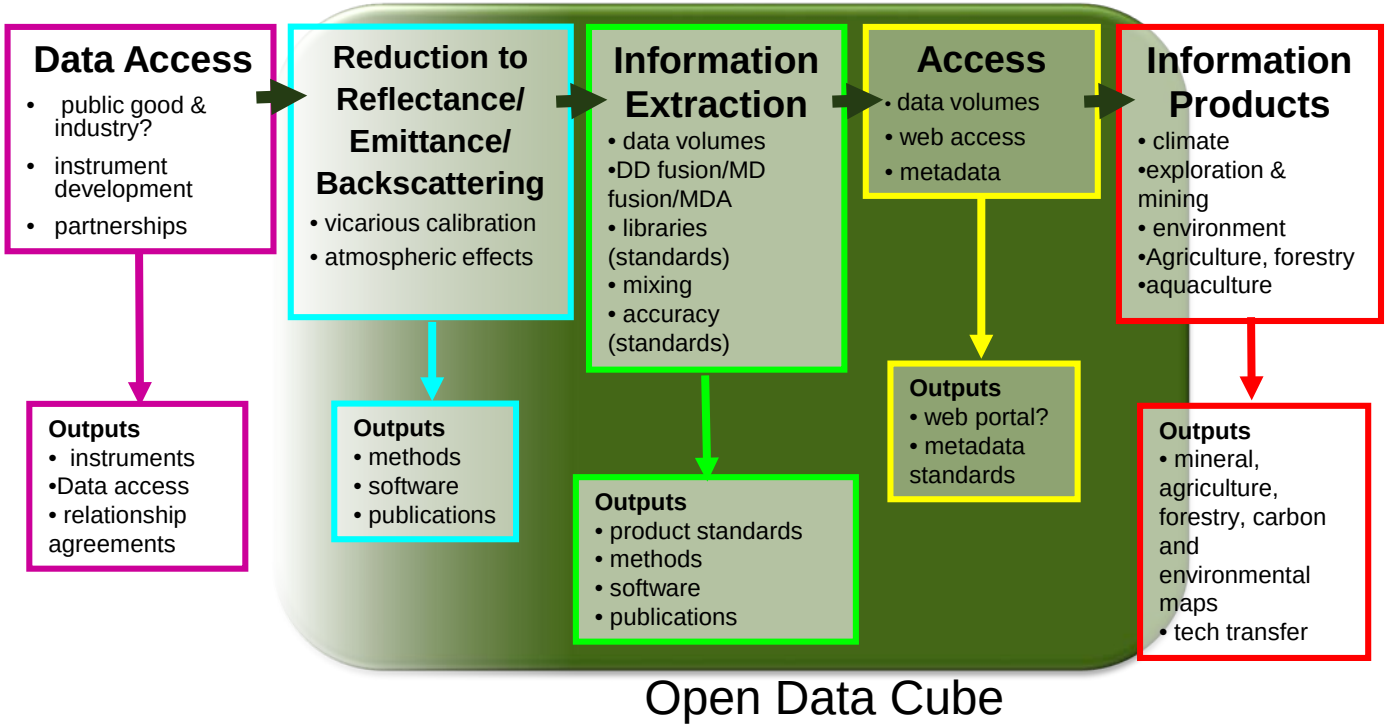
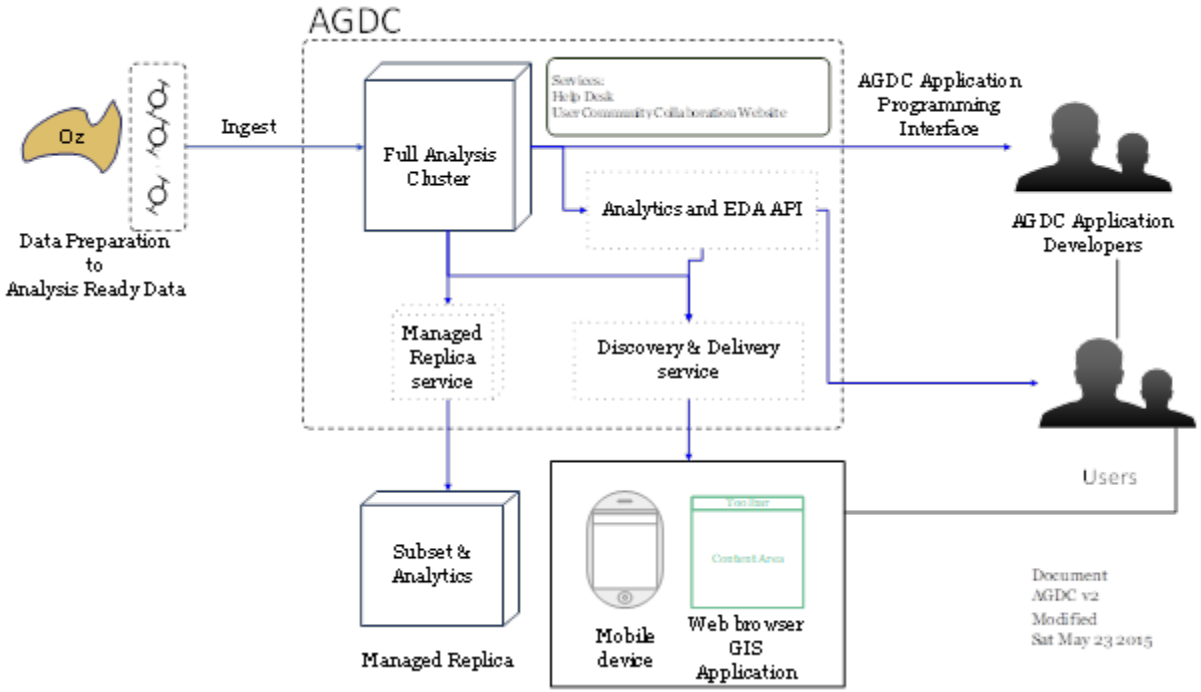
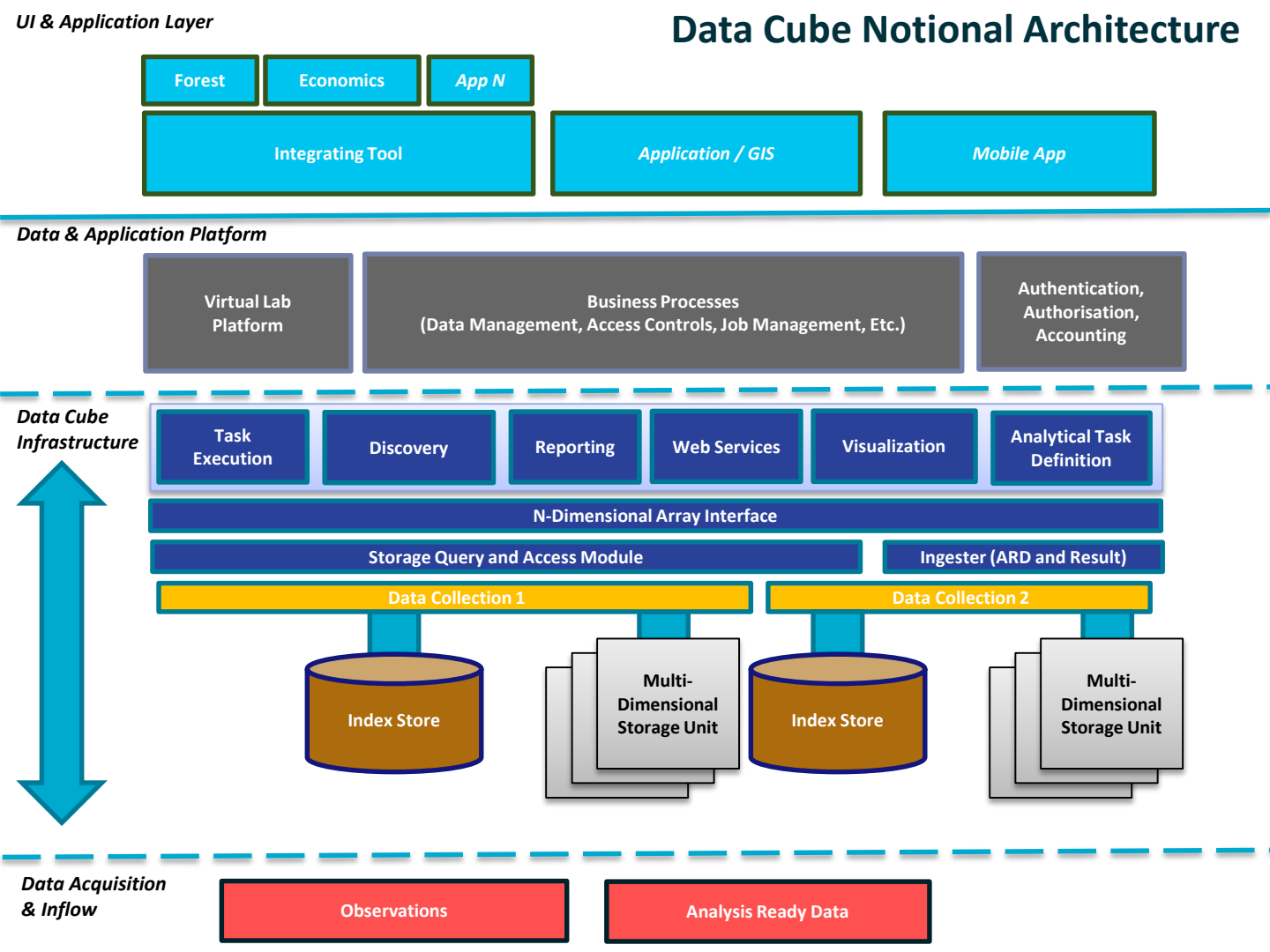


Diagram inspired by Dr Cindy Ong

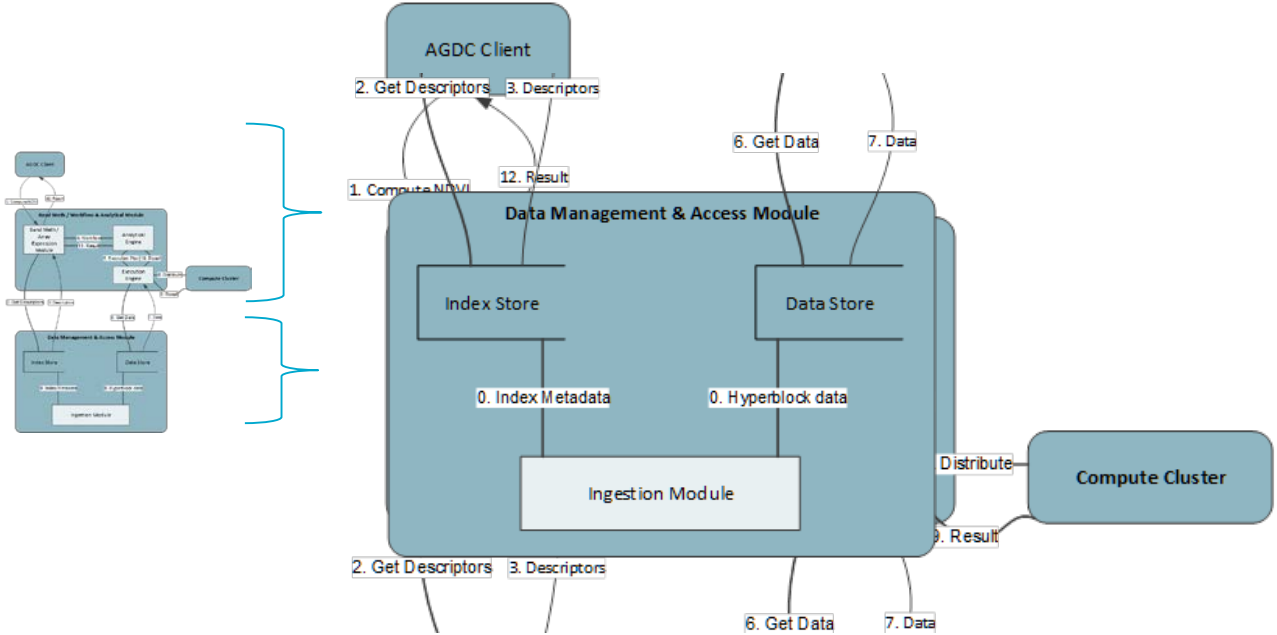
Data Cube - External Interfaces (Concept)



Data Cube Notional Architecture



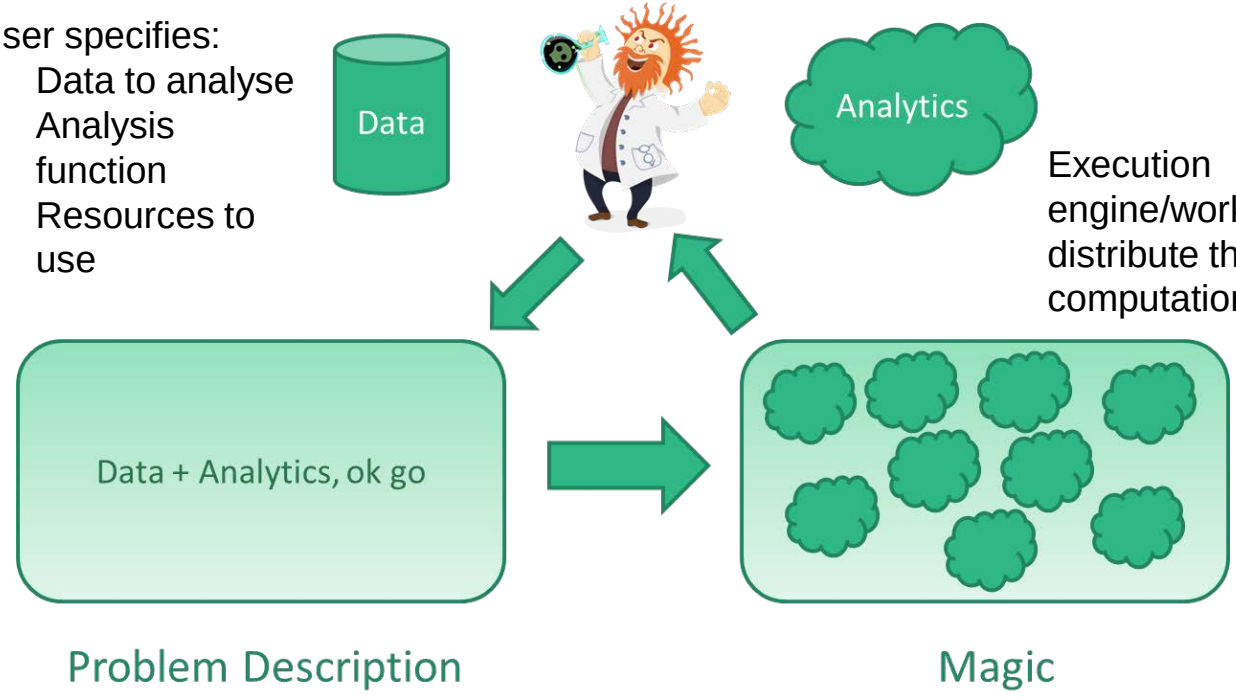
Flow Diagram - Prototype



N-dimensional Array Interface
Python, Numpy, xarray…

User specifies:

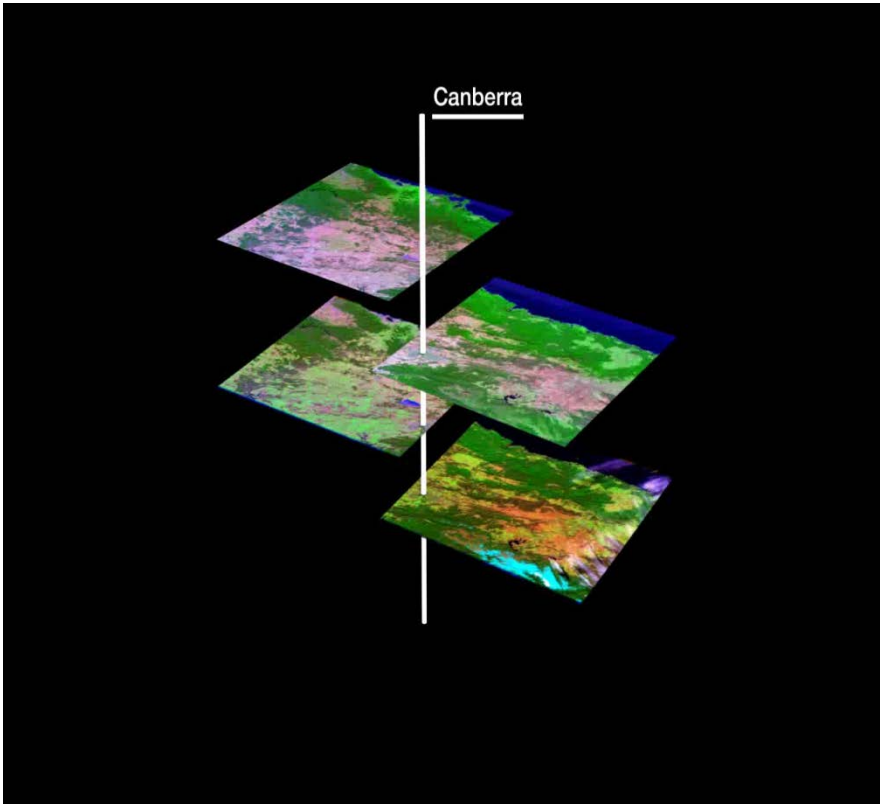
- Data to analyse
- Analysis function
- Resources to use



Execution engine/workflows distribute the computation

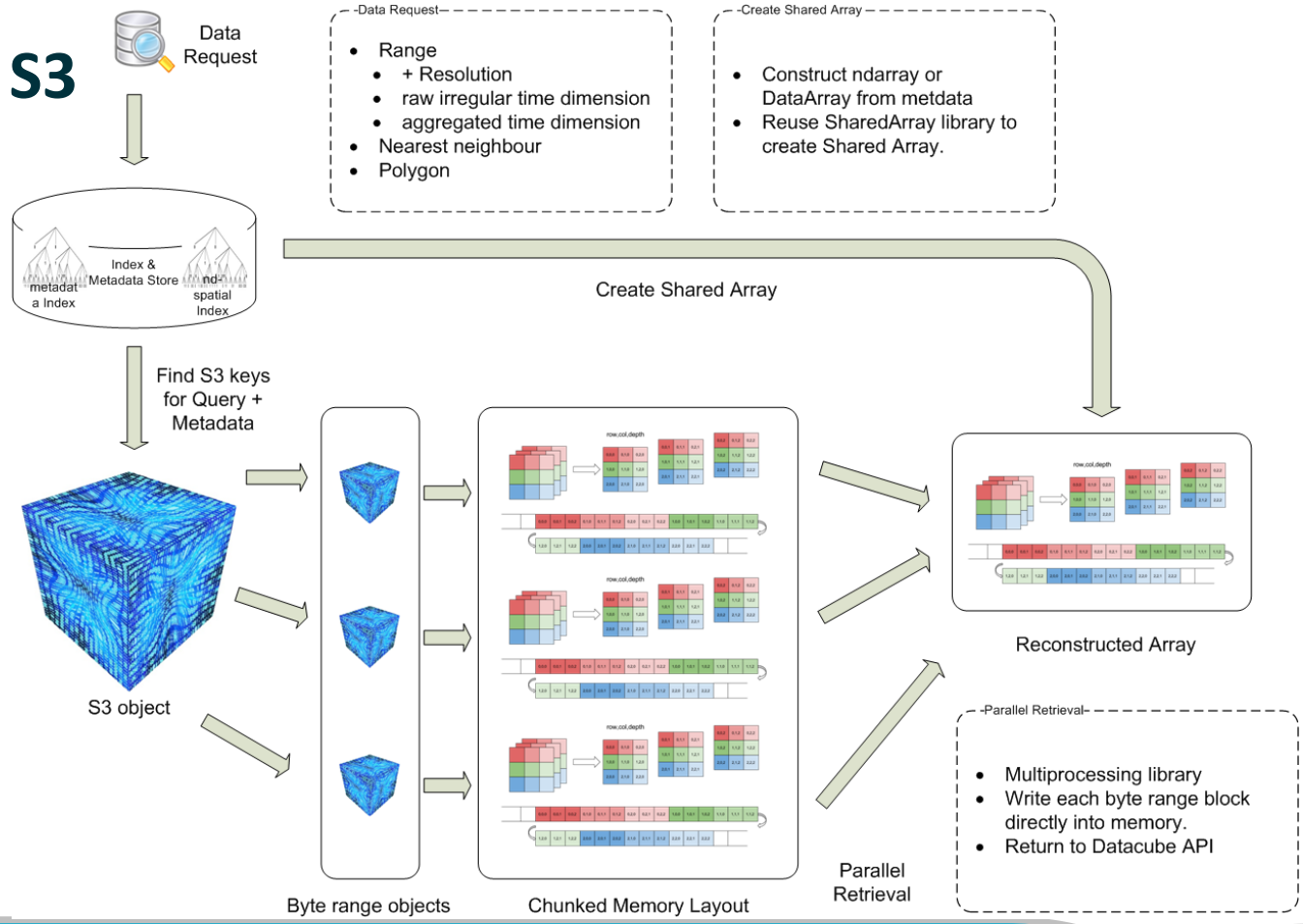
Simple data structures and analysis

- ‘Dice and Stack’
  - Can reproject on demand as well by indexing existing files
- Calibrated to surface reflectance observations (CARD4L)
- Spatial alignment and consistent calibration makes analysis *much* simpler
- Every unique observation is kept and included for analysis creating dense time-series



Storage - Files, Formats, Databases, Objects…

- ODC supports:
  - Files
    - Multiple file formats via GDAL (eg. GeoTif, NetCDF)
    - Can be on a file system or at the end of a HTTP (e.g. AWS S3 end point)
    - Multi-dimensional files (for those that support it – NetCDF)
      - Space, Time, Bands, Masks, attributes… - configurable
      - Tends towards single Time per file but can (have!) run into limits on file system at scale
  - Native S3 driver - objects on AWS
    - No files -> objects are sparse arrays
    - No limit to number of files (there is a limit but it is astronomical)
    - More cost effective on AWS than EBS (normal file system) storage
    - Scales like S3 does
    - Parallel IO


S3
Data Request
Data Request
Range
+ Resolution
raw irregular time dimension
aggregated time dimension
Nearest neighbour
Polygon
Create Shared Array
Construct ndarray or DataArray from metdata
Reuse SharedArray library to create Shared Array.
Index & Metadata Store
metadat a Index
nd-spatial Index
Create Shared Array
Find S3 keys for Query + Metadata
S3 object
Byte range objects
Chunked Memory Layout
Parallel Retrieval
Reconstructed Array
Parallel Retrieval
Multiprocessing library
Write each byte range block directly into memory.
Return to Datacube API

Deployment to HPC, Cloud, PC

- Raijin @ National Computational Infrastructure
- 57,472 cores (2.6 GHz) in 3592 compute nodes;
- 160 TBytes (approx.) of main memory;
- 10 PBytes (approx.) of usable fast file system (for short-term scratch

- Cloud – AWS
- EC2 and S3
  - S3 native IO (no files)
  - Parallel IO
  - Elastically scalable
  - Serverless? - AWS Lambda
- CSIRO Earth Analytics Industry Innovation Hub

amazon web services™

Data-Intensive Quantitative EO Science

The Open Data Cube (ODC) :

- supports the management and quantitative analysis of massive volumes of Earth observation (EO) and other geoscientific data.
- EO data are:
  - calibrated to surface reflectance observations,
  - organised as regular geographic tiles rather than scenes or images,
  - Co-located
    - high performance data (HPD) and high performance compute (HPC) at national facility for continental/Global scale
    - Cloud and PC portable for alternate needs (industry, elasticity)

This approach positions the ODC to become a sensor-independent system for management, analysis and sharing of EO data from workbench science to continental and production scale analysis

Australian Government
Geoscience Australia

CSIRO

NCI
NATIONAL COMPUTATIONAL INFRASTRUCTURE

CSIRO