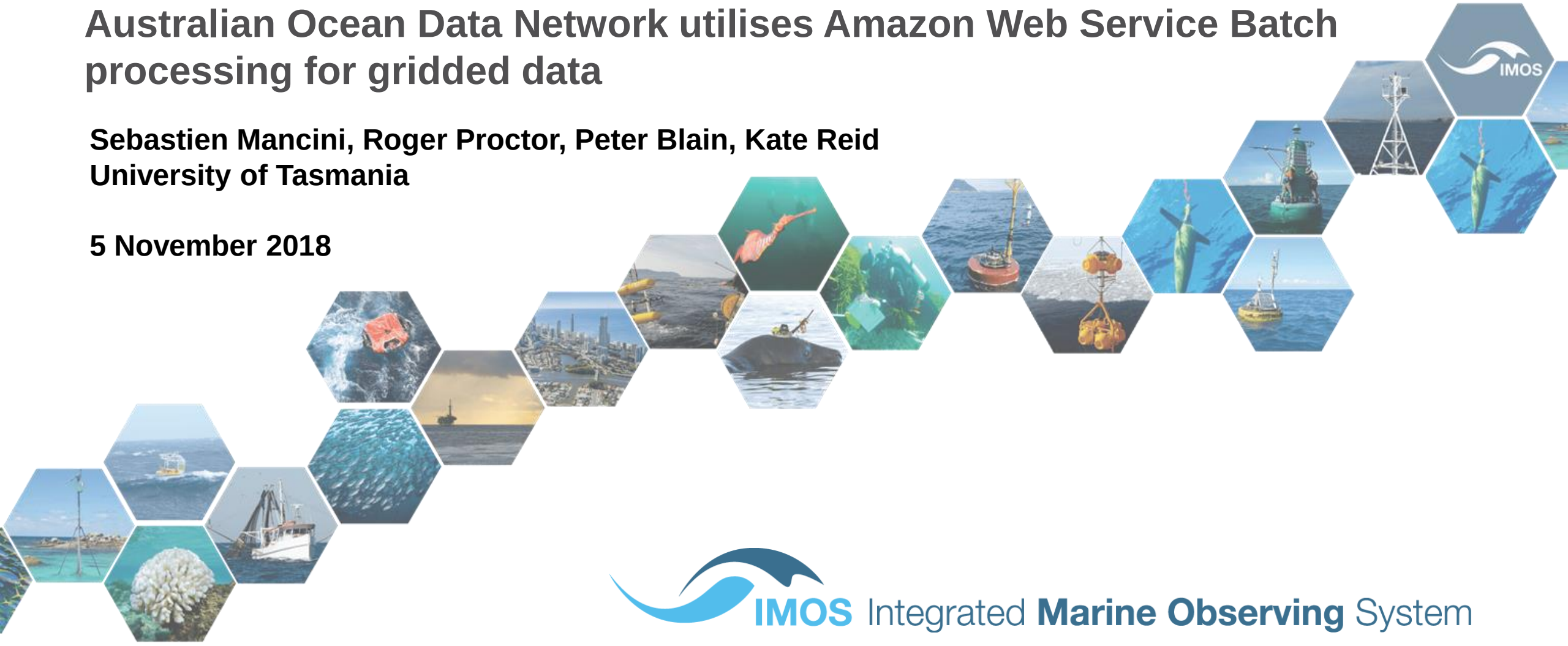
Australia's Integrated Marine Observing System (IMOS)

Australian Ocean Data Network utilises Amazon Web Service Batch processing for gridded data

Sebastien Mancini, Roger Proctor, Peter Blain, Kate Reid
University of Tasmania

5 November 2018



IMOS Integrated **Marine Observing** System

Outline of the talk

12 minutes, plus 3 minutes for questions

1. About IMOS
2. Gridded datasets
3. Cloud computing and AWS Batch
4. Helpdesk and user statistics
5. Future improvements

IMOS Facilities

AUSTRALIAN OCEAN DATA NETWORK

SATELLITE REMOTE SENSING

WIRELESS SENSOR NETWORKS

OCEAN RADAR

SHIPS OF OPPORTUNITY

AUTONOMOUS UNDERWATER VEHICLES

NATIONAL MOORING NETWORK

ANIMAL TRACKING #2 (ACOUSTIC TELEMETRY)

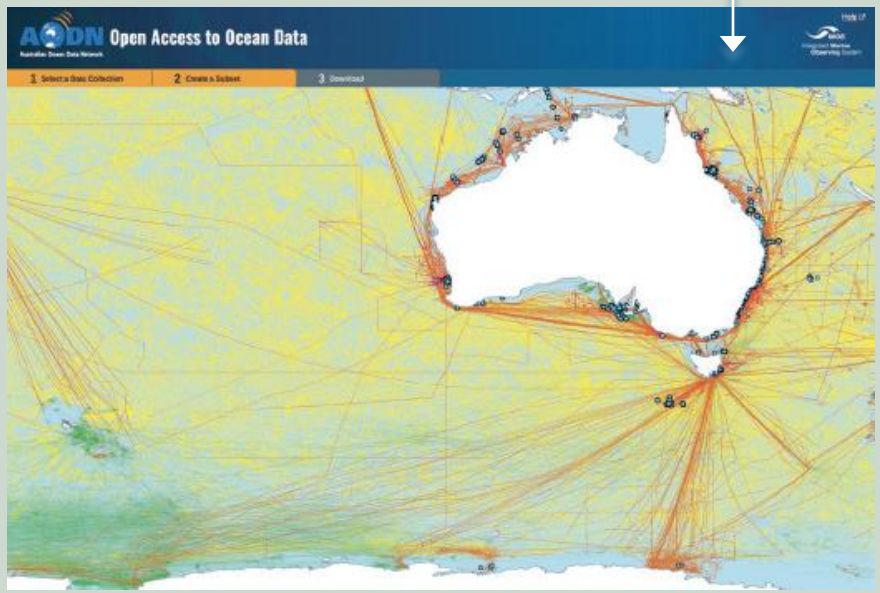
DEEP WATER MOORINGS

ARGO FLOATS

OCEAN GLIDERS

ANIMAL TRACKING #1 (TAGGING)

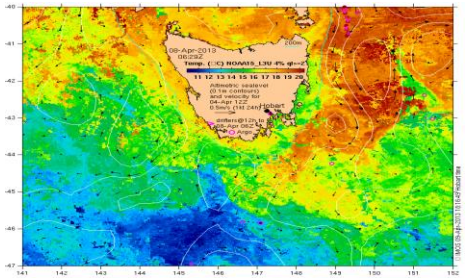
All data discoverable, accessible, usable and reusable

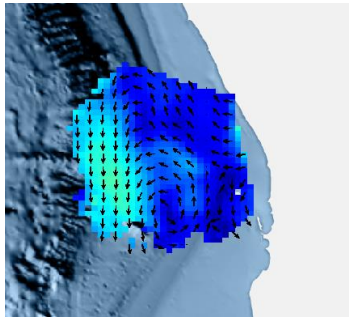


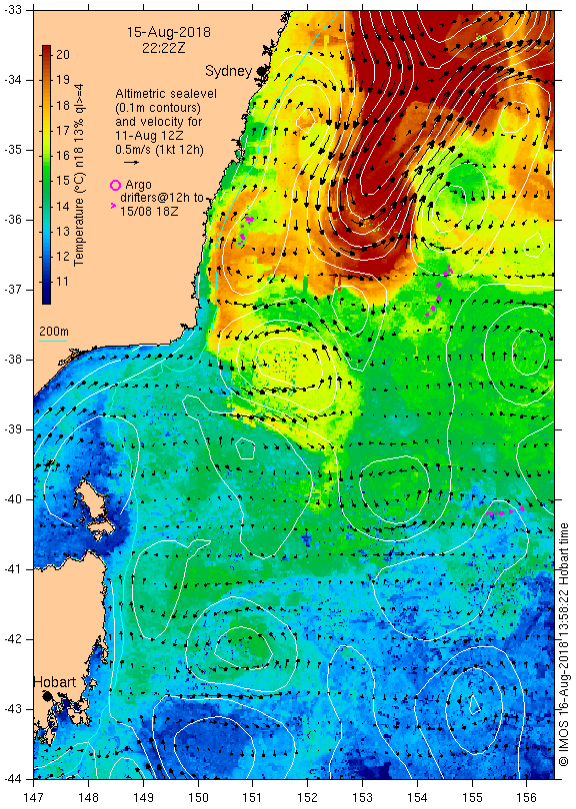
https://portal.aodn.org.au

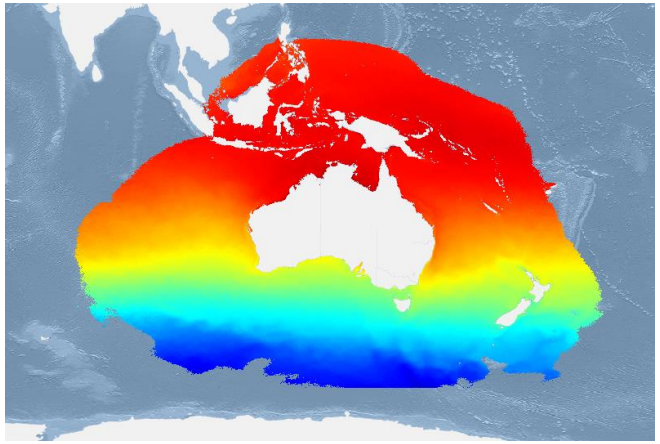
IMOS gridded datasets

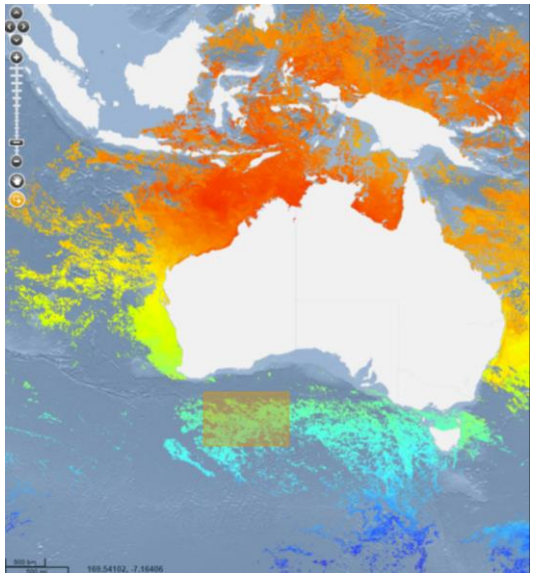
- Sea Surface Temperature (SST) products
- Ocean Colour products
- Satellite altimetry products
- Coastal radar products
- Climatology
- Bathymetry





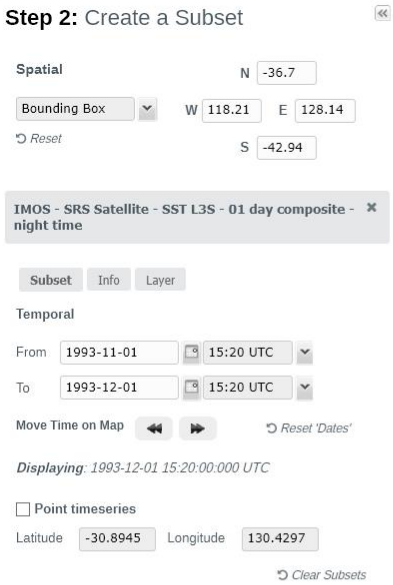


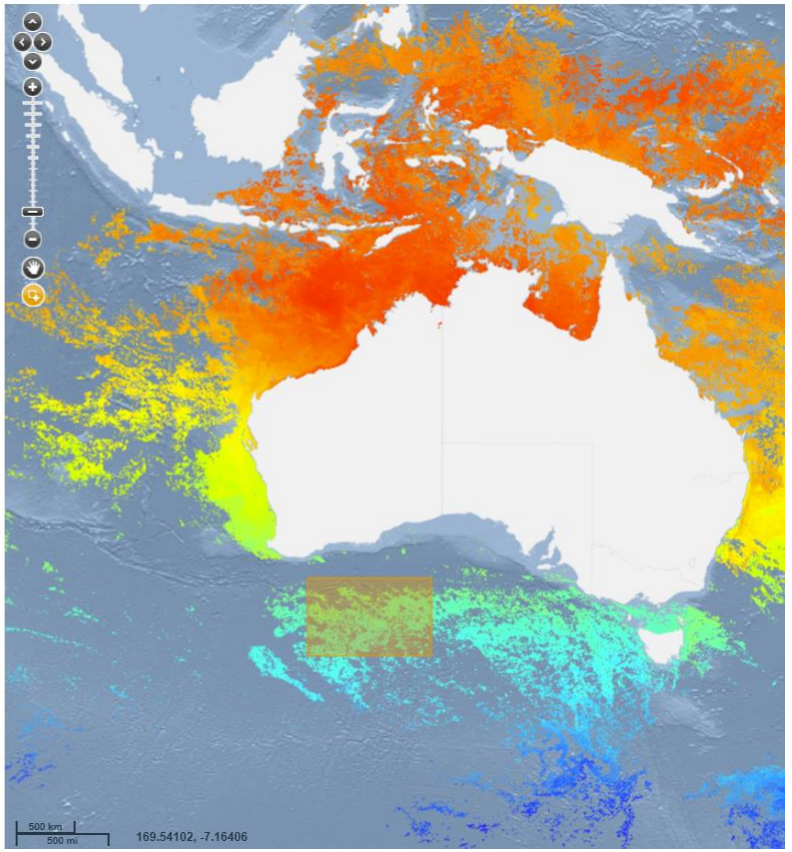




What does the user want?

- Visualise data before downloading
- Retrieve a timeseries at a particular location and download data in CSV format
- Subset and aggregate data and generate output file in netCDF format
- Most of them do not want to access data on THREDDS





What does the user want?

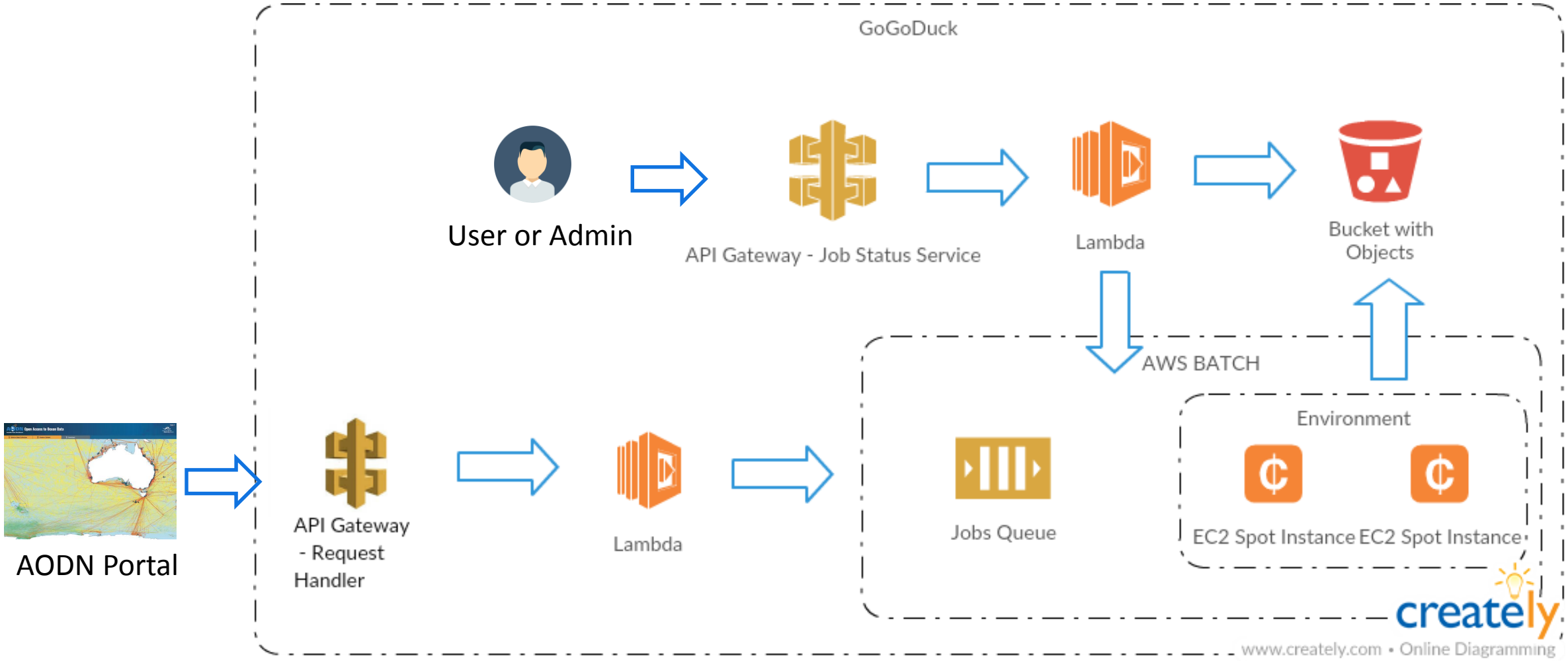
Most downloaded datasets on the AODN Portal for the period 01/2016 to 04/2018 using Google Analytics

Rank	Event Label	Total Events
1	IMOS - Australian National Mooring Network (ANMN) Facility - Current velocity time-series	944
2	IMOS - Argo Profiles	853
3	IMOS - Australian National Mooring Network (ANMN) Facility - Temperature and salinity time-series	776
4	IMOS - SRS Satellite - SST L3S - 01 day composite - night time	765
5	IMOS - SRS - MODIS - 01 day - Chlorophyll-a concentration (OC3 model)	679
6	IMOS - Australian National Facility for Ocean Gliders (ANFOG) - delayed mode glider deployments	501
7	IMOS - SRS Satellite - SST L3S - 06 day composite - day time	458
8	IMOS - OceanCurrent - Gridded sea level anomaly - Delayed mode	439
9	IMOS National Reference Station (NRS) - Salinity, Carbon, Alkalinity, Oxygen and Nutrients (Silicate, Ammonium, Nitrite/Nitrate, Phosphate)	420
10	IMOS - SRS Satellite - SST L3S - 1 month composite - day and night time composite	404
11	IMOS - SRS SATELLITE - SST L3S - 01 day composite - day and night time composite	376
12	IMOS - OceanCurrent - Gridded sea level anomaly - Near real time	368
15	IMOS - SRS - MODIS - 01 day - Ocean Colour - SST	324
17	IMOS - SRS SATELLITE - SST L3S - 03 day composite - night time	280
19	IMOS - SRS Satellite - SST L3S - 03 day composite - day and night time composite	227
20	IMOS - SRS Satellite - SST L3S - 01 day composite - day time	209
22	IMOS - SRS Satellite - SST L3S - 06 day composite - day and night time composite	169
23	IMOS - SRS - MODIS - 01 day - Chlorophyll-a concentration (GSM model)	157
26	MARVL3 - Australian shelf temperature data atlas	144
27	IMOS - SRS Satellite - SST L3S - 1 month composite - day time	141

Serverless architecture

- Server details get abstracted away
- Servers only run when needed
- Leave server management to a company that does that as their bread and butter
- Focus on what matters instead - the code and data

Architecture overview of AWS batch



Helpdesk: Queue Monitoring

https://processes.aodn.org.au/wps/jobStatus?format=queue

Search

Australian Ocean Data Network

WPS QUEUE CONTENTS: arn:aws:batch:ap-southeast-2:104044260116:job-queue/JavaDuckExecuteQueue-aws-wps-prod

Queued Jobs

No jobs are currently queued.

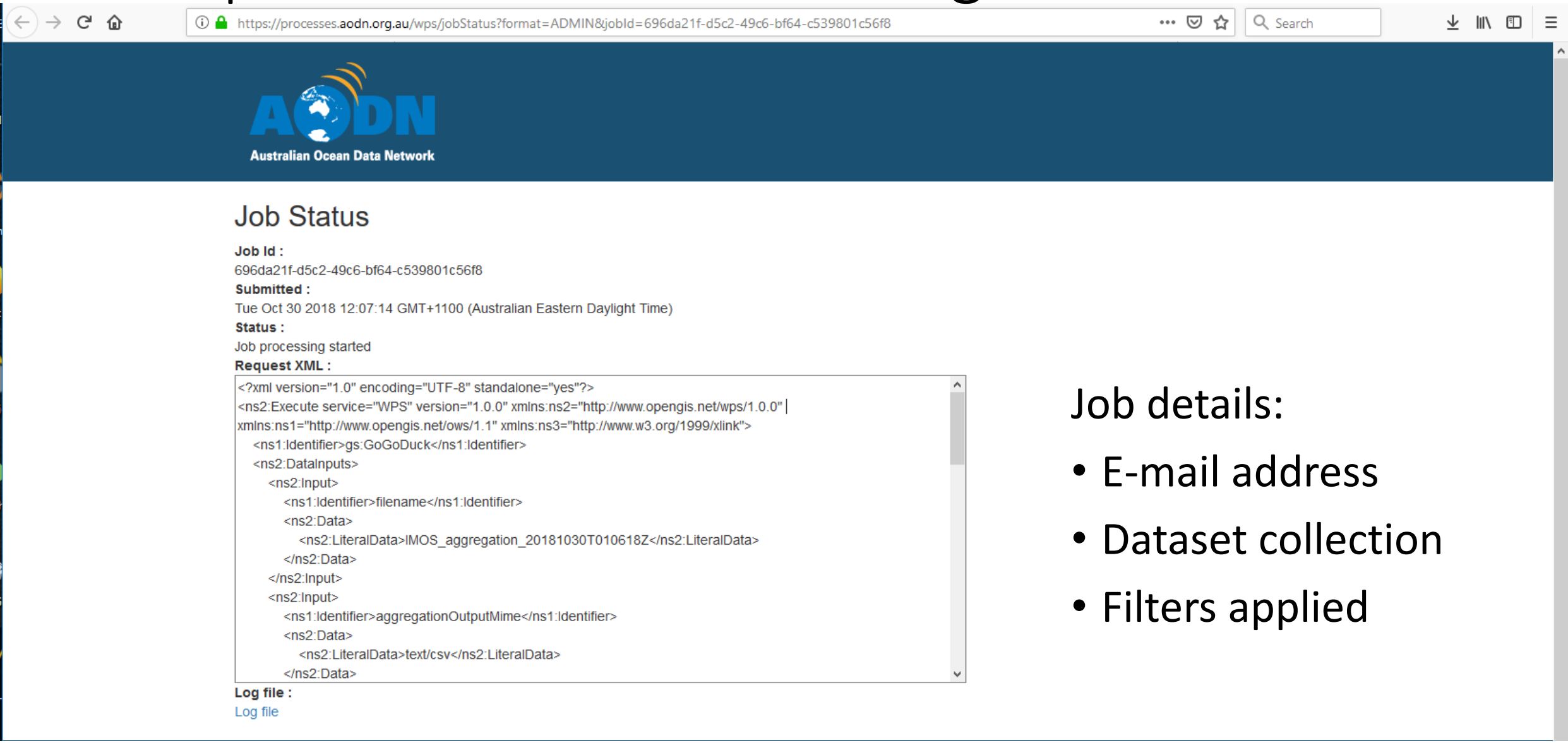
Running Jobs

Job ID	Submitted	Started	Status	Log File
696da21f-d5c2-49c6-bf64-c539801c56f8	30/10/2018 12:07:13	30/10/2018 12:17:14	RUNNING	Log
27d55f97-1461-4245-a285-03b31e74f50c	30/10/2018 12:06:16	30/10/2018 12:36:00	RUNNING	Log
ff0d36ac-8159-4cfe-9954-a97a7ffa0e69	30/10/2018 12:05:10	30/10/2018 12:17:16	RUNNING	Log
d4d8f5b5-d488-4c7a-b915-fe03c671a177	30/10/2018 12:04:12	30/10/2018 12:06:10	RUNNING	Log

Completed Jobs

Job ID	Submitted	Started	Completed	Job Status	Aggregation Result	Log File
79140e25-5e53-4723-ac09-e1f4065649ec	30/10/2018 11:26:45	30/10/2018 11:29:13	30/10/2018 12:35:41	SUCCEEDED	Download ready	Log
7a1735e8-450a-4717-b6b2-8c1059df214f	30/10/2018 05:18:22	30/10/2018 05:21:26	30/10/2018 05:22:55	SUCCEEDED	Download ready	Log
b27cc9f9-47b1-44d4-9512-4fc0c045b04b	29/10/2018 17:16:15	30/10/2018 04:39:23	30/10/2018 04:42:28	SUCCEEDED	Download ready	Log
81faec2d-df96-4a97-9c01-b1f5471e52a6	29/10/2018 16:05:26	29/10/2018 17:06:24	30/10/2018 05:51:02	SUCCEEDED	Download ready	Log
a838446a-d6bf-4e10-b21f-23acc463fad0	29/10/2018 16:02:13	29/10/2018 16:06:55	30/10/2018 04:39:08	SUCCEEDED	Download ready	Log
cba2aee4-50ff-49fd-97bd-acf42beef30b	29/10/2018 16:00:47	29/10/2018 16:06:53	30/10/2018 04:57:07	SUCCEEDED	Download ready	Log
97968351-fc79-44c6-a4f9-4b33c675da27	29/10/2018 15:59:48	29/10/2018 16:06:51	30/10/2018 05:04:51	SUCCEEDED	Download ready	Log

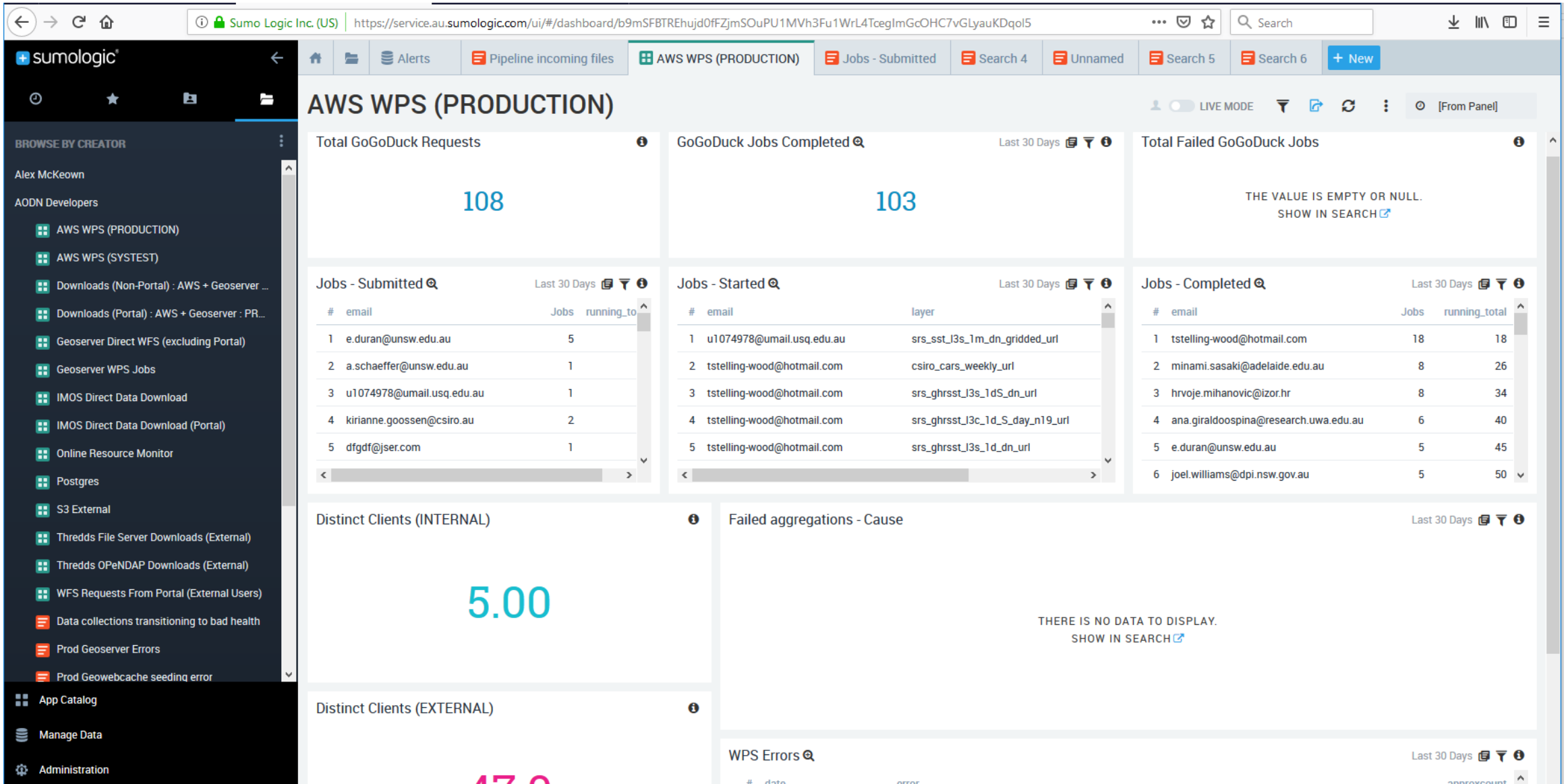
Helpdesk: Queue Monitoring



Job details:

- E-mail address
- Dataset collection
- Filters applied

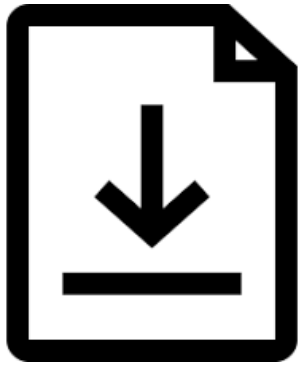
User statistics: Sumologic dashboard



User statistics: February to September 2018

240 unique users

42 users per month



1600 downloads over the entire period

350 downloads in June

32 failed downloads over the entire period

Process time of < 8 min for a 1000 jobs
10% jobs have an averaged processed time of 900 min

Queue time of < 7 min for a 1000 jobs
10% jobs have an averaged queued time of 550 min

Cost comparison:

AWS Batch	vs	Normal approach
80$ per month		**350$** per month
EC2 SPOT instance + local storage + Lambda		EC2 instance (if reserved) + Local storage

Additional cost for both approaches:

- Storage for output file
- S3 operations (data access)

Vendor Lock-in

+ majority of aggregation code written as a generic library/utility

- request handler Lambda has some AWS specific code, but could be refactored to a more generic API handler

- status service Lambda is quite specific to AWS Batch/S3 currently, but could be made more abstract

- the components are glued together by CloudFormation and do make assumptions about running on Lambda, however the majority of the classes are generic, so if running outside of AWS was a requirement, the project could be refactored to make concepts like storage and APIs more generic

Future improvements

- Implement cloud computing to other elements of our infrastructure:
  - Development of our application stack (AODN Portal, Geonetwork, Geoserver, Geowebcache, ncWMS …)
- Improve subsetting/aggregating code:
  - Download multiple point timeseries at the same time
  - Improve efficiency to produce result quicker
- Use of system analytics to improve queue design
  - Multiple queues depending on size of jobs, type of users …
- Apply AWS Batch for other type of data downloads:
  - Large CSV downloads using WFS requests from Geoserver
  - Subset Geotiff (e.g. Bathymetry data)

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

IMOS is a national collaborative research infrastructure, supported by Australian Government. It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent. **www.imos.org.au**

PRINCIPAL PARTICIPANTS

University of Tasmania
(Lead Agent)

Australian Government
Australian Institute of Marine Science
Australian Government
Bureau of Meteorology
CSIRO

Government of South Australia
SARDI
The University of Western Australia
sims sydney institute of marine science
UTS
The University of Sydney
Macquarie University
UNSW Sydney

SIMS is a partnership involving four Universities.

ASSOCIATE PARTICIPANTS

Curtin University

Australian Government
Department of the Environment and Energy
Australian Antarctic Division