ADC-C1

Build efficient, cross-Regional, I/O-intensive workloads with Dask on AWS

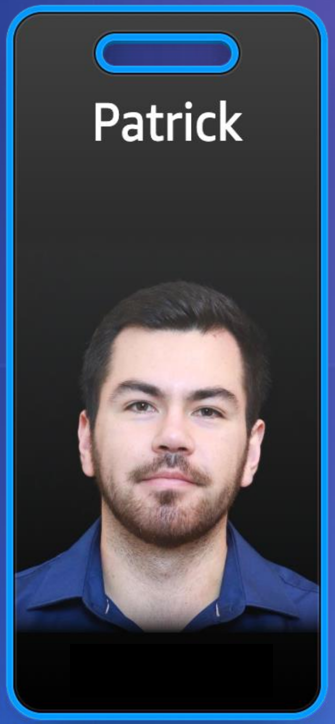
UK Meteorological Office

Patrick O'Connor

Prototyping Engineer
AWS

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Your Presenter



Patrick O’Connor
Specialist SA, WW Prototyping

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data is a **strategic** asset

Companies realize the potency of information and data in today's technological landscape.

99%

of businesses want to be data driven

Harvard Business Review,
[Why Is It So Hard to Become a Data-Driven Company]

← gap →

26.5%

of businesses have been successful

Harvard Business Review,
[Why Becoming a Data-Driven Company Is So Hard]

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customers want more value from their data

Growing exponentially

From new sources

Increasingly diverse

Used by many people

Analysed by many applications

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data strategy on AWS

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sustainability solutions are powered by data

Data is diverse, growing exponentially, and used by many applications

AWS storage and analytics services and data programs can help

Open Data Sponsorship Program

Amazon Sustainability Data Initiative (ASDI)

AWS Data Exchange

Image from Landsat 8 satellite, courtesy of the U.S. Geological Survey

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Open Data Sponsorship Program covers the cost to store and distribute the world's most valuable, impactful data

We work with data providers and data users who seek to:

Democratize access to data by making it available for analysis on AWS

Encourage the development of **communities** that benefit from access to shared datasets

Develop new cloud-native techniques, formats, and tools that **lower the cost** of working with data

Learn more at opendata.aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ASDI: Making access to data faster, cheaper, and easier

Registry of Open Data on AWS

Amazon Sustainability Data Initiative

The Amazon Sustainability Data Initiative (ASDI) seeks to accelerate sustainability research and innovation by minimizing the cost and time required to acquire and analyze large sustainability datasets. These datasets are publicly available to anyone. In addition, ASDI provides cloud grants (pdf link) to those interested in exploring the use of AWS' technology and scalable infrastructure to solve big, long-term sustainability challenges with this data. The dual-pronged approach allows sustainability researchers to analyze massive amounts of data in mere minutes, regardless of where they are in the world or how much local storage space or computing capacity they can access. Learn more about ASDI here.

Categories: weather, climate, water, agriculture, satellite imagery, elevation, air quality, energy, ecosystems, disaster response, oceans

Search datasets (currently 84 matching datasets)

Search datasets

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

ENERGY

ARPA-E PERFORM Forecast data

Managed by National Renewable Energy Laboratory

The ARPA-E PERFORM Program is an ARPA-E funded program that aim to use time-coincident power and load seeks to develop innovative management systems that represent the relative delivery risk of each asset and balance the collective risk of all assets across the grid. A risk-driven paradigm allows operators to: (i) fully understand the true likelihood of maintaining a supply-demand balance and system reliability, (ii) optimally manage the system, and (iii) assess the true value of essential reliability services. This paradigm shift is critical for all power systems and is essential for grids wi...

Department of Energy's Open Energy Data Initiative (OEDI)

Managed by National Renewable Energy Laboratory

Data released under the Department of Energy's Open Energy Data Initiative (DOE). The Open Energy Data Initiative (OEDI) aims to improve and automate access of high-value energy data sets across the U.S. Department of Energy's (DOE's) programs, offices, and national laboratories. OEDI aims to make data actionable and discoverable by researchers and industry to accelerate analysis and advance innovation.

NREL National Solar Radiation Database

Managed by National Renewable Energy Laboratory

Released to the public as part of the Department of Energy's Open Energy Data Initiative, the National Solar Radiation Database (NSRDB) is a serially complete collection of hourly and half-hourly values of the three most common measurements of solar radiation – global horizontal, direct normal, and diffuse horizontal irradiance — and meteorological data. These data have been collected at a sufficient number of locations and temporal and spatial scales to accurately represent regional solar radiation climates.

NREL Wind Integration National Dataset

Managed by National Renewable Energy Laboratory

Released to the public as part of the Department of Energy's Open Energy Data Initiative, the Wind Integration National Dataset (WIND) is an update and expansion of the Eastern Wind Integration Data Set and Western Wind Integration Data Set. It supports the next generation of wind integration studies.

ASDI helps researchers, scientists, and innovators around the world advance their work on sustainability-related research by providing publicly available, free access to important scientific data.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The UK Meteorological Office

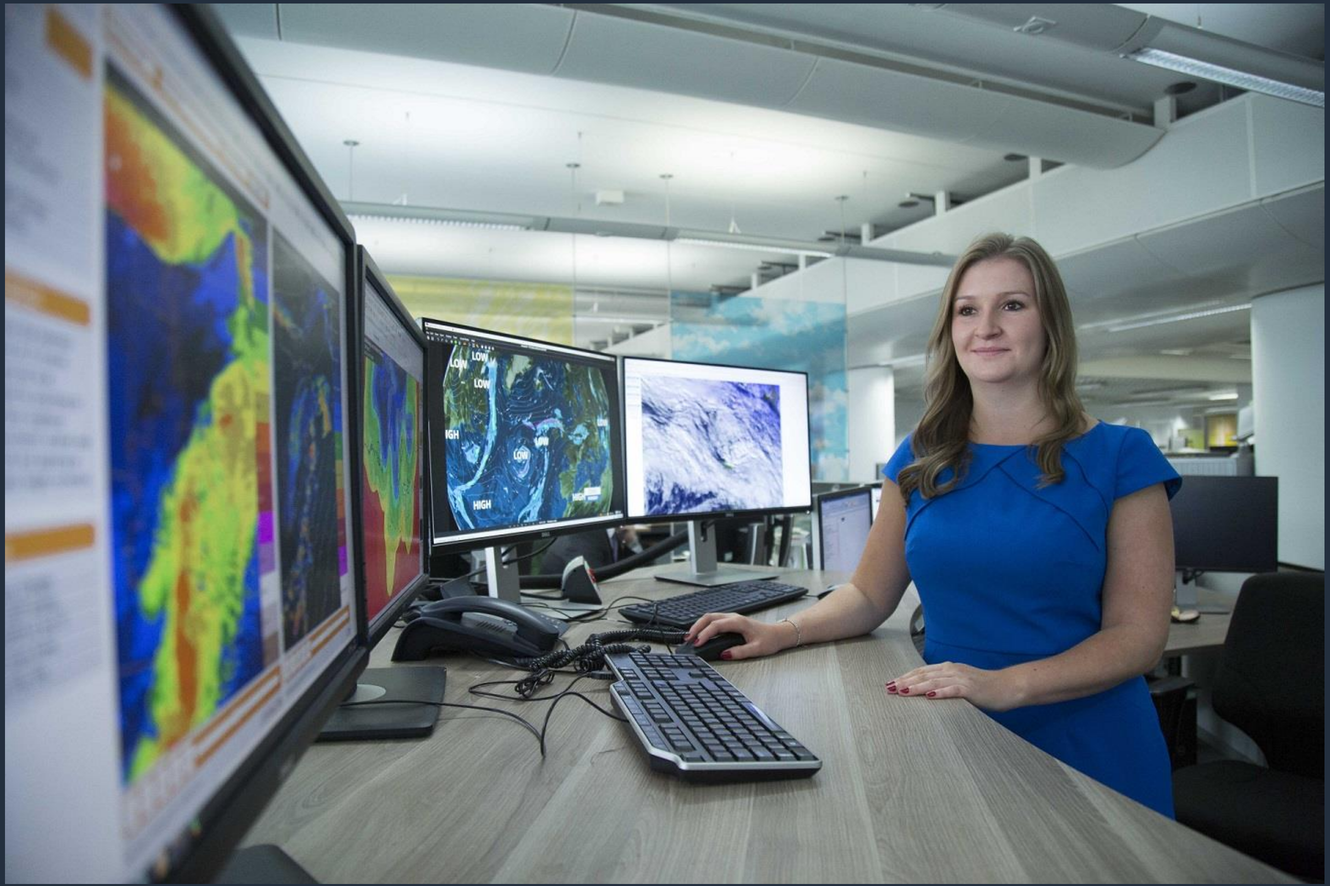
The Met Office was founded in 1854 and is the national meteorological service for the UK. They provide weather and climate forecasts to help you make better decisions to stay safe and thrive.

They collect, create, and make sense of massive amounts of data every day, using cutting-edge technology for the benefit of mankind - and our planet.

They co-operate with and support businesses, agencies and governments in making short and long-term decisions, making the world a safer and more resilient place tomorrow, and for the years - and decades - to come.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Weather Data



https://www.youtube.com/watch?v=tls9h2q7QlY

Key Stats

168
Years of operation

300 TB
per day of weather data

2.3m users
during Storm Eunice

3.4m
impressions a day on Twitter at times of severe weather

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Met Office

Challenge with datasets across the globe

Challenge

Data is sparsely located

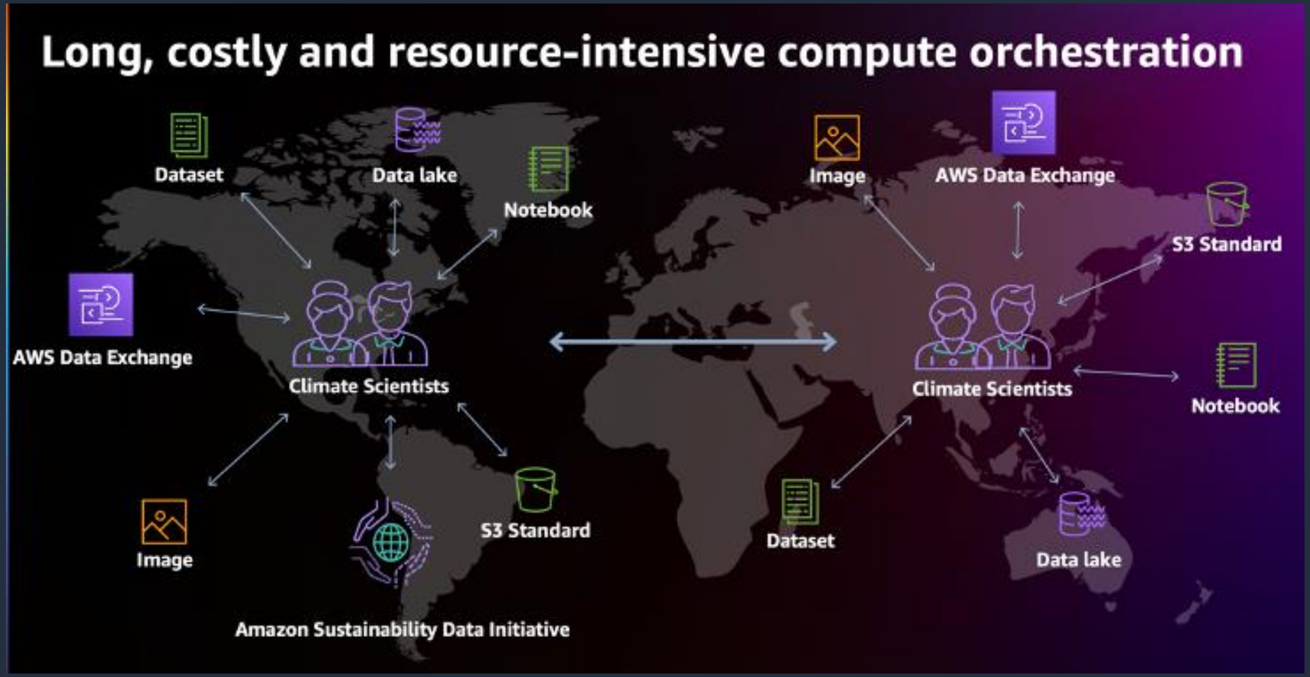
How can we combine cross regional data?

Data volumes into the petabyte scale

Different types of data sources

How can users interact consistently with data?

How can we scale?



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tomorrows' science needs new platforms

If scientists' questions are constrained by tooling
they are encouraged to confirm results they expect

The most important scientific results are unexpected
We need tools which allow scientists to explore and discover with data

We need to give scientists back their "flow"
By giving them tools and platforms which give them a modern user-experience

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

12

Innovating for sustainability

Identify

Identify needs
Start with the customer

↓

Generate ideas
Brainstorm

↓

Evaluate opportunity
Answer the 5 questions

Innovate

Discover solutions
Create PR FAQ, visuals

↓

Select solution

↓

Build Prototype
Refine & Iterate

Implement

Finalize solution
Continually revisit PR FAQ, visuals

↓

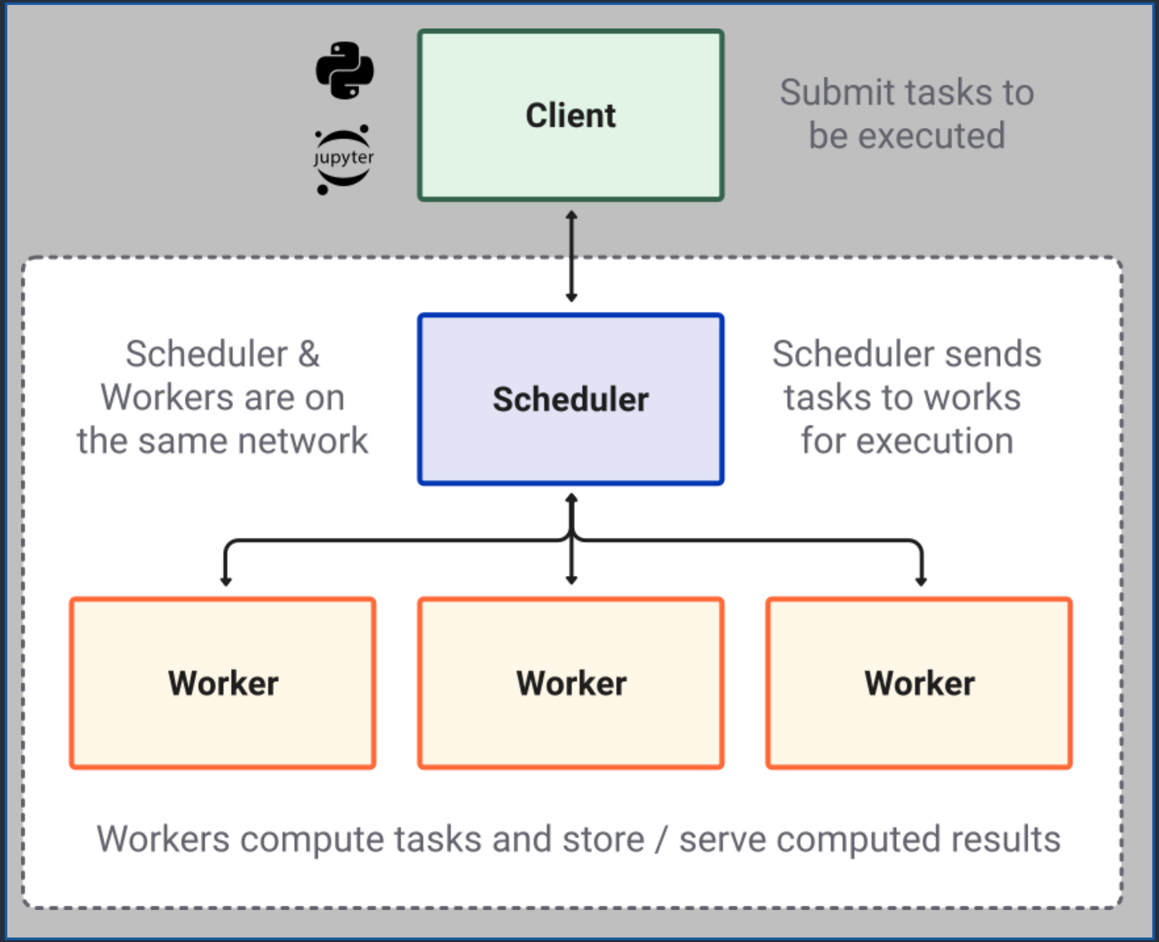
Production & distribution

↓

Transition to operations & scale

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is Dask?



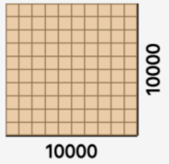
Create Random array

This creates a 10000×10000 array of random numbers, represented as many numpy arrays of size 1000×1000 (or smaller if the array cannot be divided evenly). In this case there are 100 (10×10) numpy arrays of size 1000×1000.

```
[2]: import dask.array as da
     x = da.random.random((10000, 10000), chunks=(1000, 1000))
     x
```

[2]:

	Array	Chunk
Bytes	762.94 MiB	7.63 MiB
Shape	(10000, 10000)	(1000, 1000)
Count	100 Tasks	100 Chunks
Type	float64	numpy.ndarray

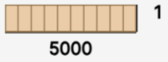


Use NumPy syntax as usual

```
[3]: y = x + x.T
     z = y[::2, 5000:].mean(axis=1)
     z
```

[3]:

	Array	Chunk
Bytes	39.06 kiB	3.91 kiB
Shape	(5000,)	(500,)
Count	430 Tasks	10 Chunks
Type	float64	numpy.ndarray



Call `.compute()` when you want your result as a NumPy array.

If you started `Client()` above then you may want to watch the status page during computation.

```
[4]: z.compute()
[4]: array([1.00226063, 1.01066798, 1.00353892, ..., 1.00020978, 1.00972641,
            0.99609573])
```

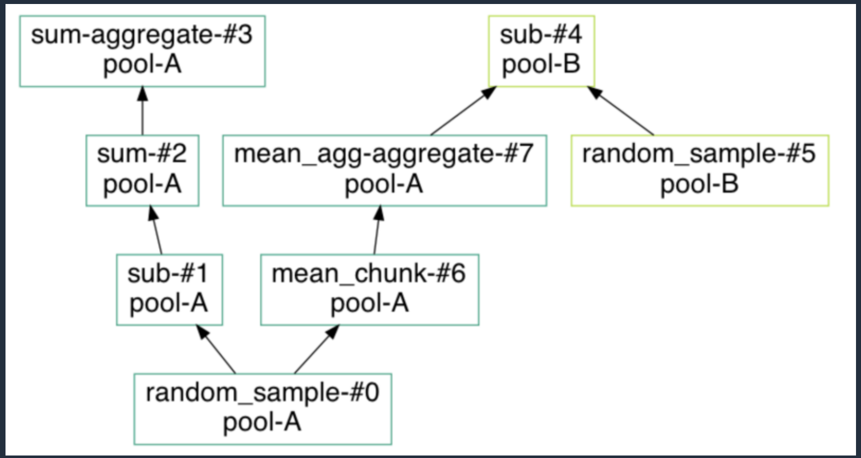
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Additional Technologies

Numpy

Dask-worker-pools

Xarray

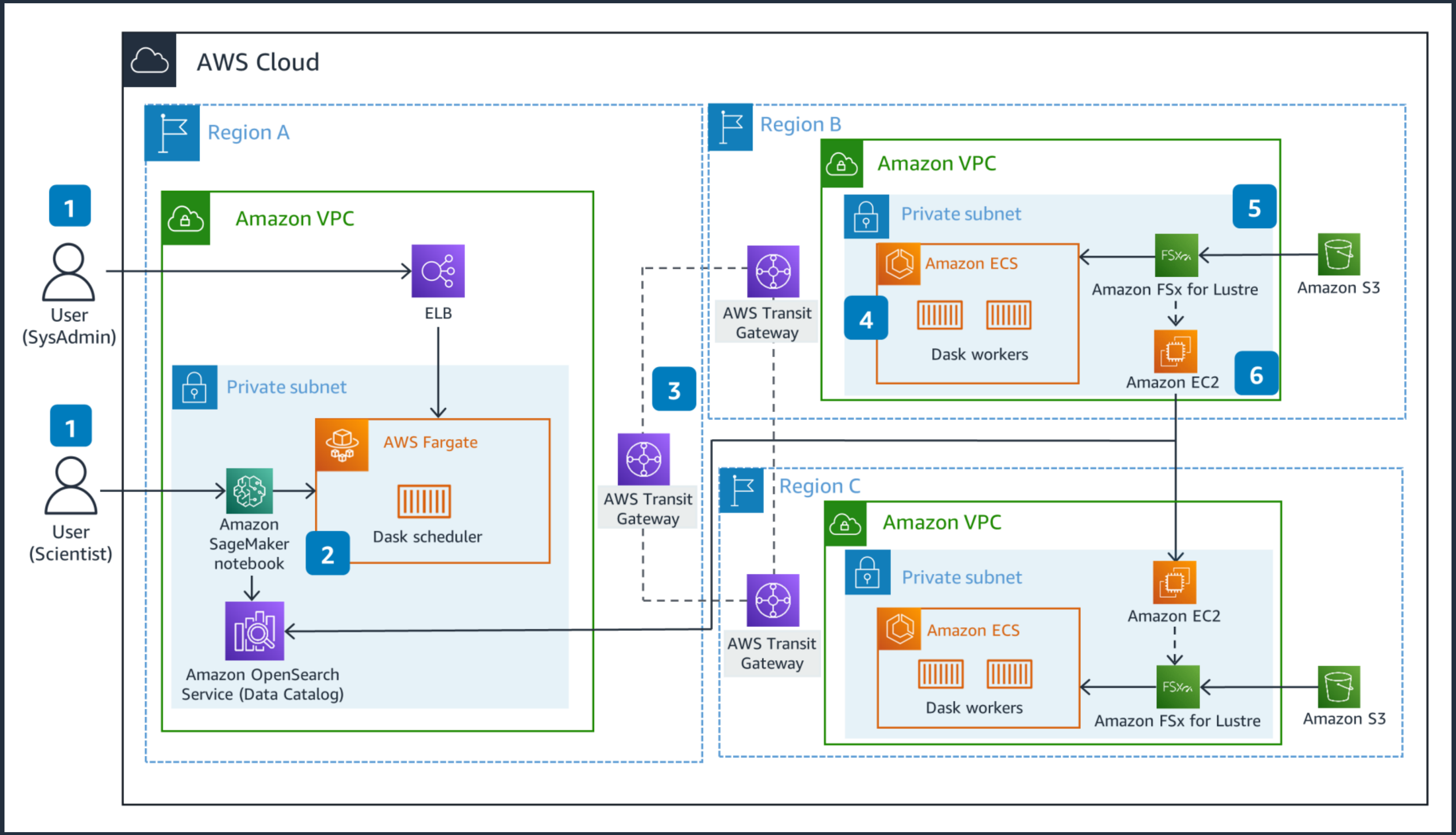



```
import dask.array as da
from dask_worker_pools import pool, propagate_pools, visualize_pools


with pool("A"):
    # Only pool-A workers can access this proprietary random data!
    a = da.random.random((10, 10))

with pool("B"):
    # Only pool-B workers can access this proprietary random data!
    b = da.random.random(10)

run_in_a = (a - 1).sum()
# ^ Want this to run only in A (transferring A data to B is expensive)

run_in_b = b - a.mean()
# ^ Want this to run in B, because `a.mean()` is smaller to transfer than all of `b`


with propagate_pools():
    # ^ Automatically propagates pool restrictions forward
    dask.compute(run_in_a, run_in_b)

visualize_pools(run_in_a, run_in_b, filename="pools.png")
```

https://github.com/gjoseph92/dask-worker-pools

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Orchestrate petabyte-scale computing across AWS Regions



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

User Interface

Data specification exploration

View and select the desired inputs for the data

The cells below will show you some of the available data, and allow you to specify variables as well as filter down lists to make it easy for you to further specify what data you want.

```
[36]: # Display and save variable categories (e.g. temperature)
%run get_variables.ipynb
```

Below are the options for the data type.

Data type: ○ predictive
○ historical
◉ both

Below are available variable categories

```
temperature
wind
wave
precipitation
snow
air_pressure
```

Which of the above would you like to select? temperature

Below are the options for the variable.

variable: air_temperature_at_2_metres

Below are the start and end dates.

Start Date 01/01/2015

End Date 12/31/2020

Data exploration and analysis

We can now do analysis on the data set with the data as an xarray. The cells below calculate the mean and regrided value for the data set

```
[42]: import numpy as np
from dask_worker_pools import pool, propagate_pools
```

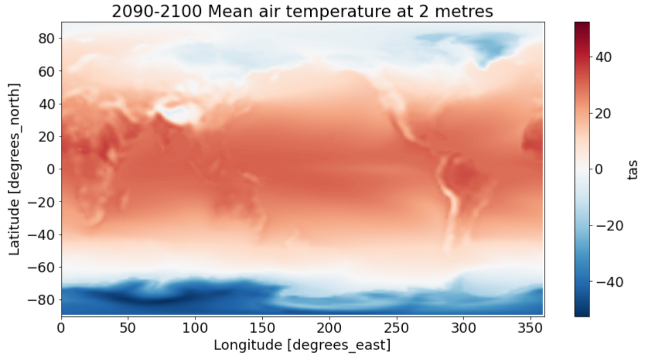
```
[43]: #some paramters for regridding
regrid_lon=np.arange(0.0,360.0,0.1)
regrid_lat=np.arange(-90.0,90.0,0.1)
regrid_method='slinear'
```

Predictive

Retrieve calculated data set

```
[44]: %%time
if (data_type != "historical"):
    %store -r prediction_pool_region predictive_data_set
    with pool(prediction_pool_region):
        predicted_tas_mean = (predictive_data_set['tas']-273).mean(dim='time') #mean
        predicted_tas_regridded = predicted_tas_mean.interp(lon=regrid_lon,lat=regrid_lat,method=regrid_method) #dodgy regridding

    with pool(prediction_pool_region):
        predicted_tas_regridded.compute() #explict compute so you can see where it happens (could just do .plot() but it would be hidden)
        predicted_tas_regridded
        predicted_tas_regridded.plot(figsize=(14,7))
        plt.title(f'2090-2100 Mean {desired_attribute.replace("_"," ")}')
else:
    print("Did not run predictive calculations for historical data")
```

```
CPU times: user 775 ms, sys: 212 ms, total: 987 ms
Wall time: 7.51 s
```



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Performance Metrics

Dataset	Variables	Disk Size	Xarray Dataset Size	Region
ERA5	2011–2020 (120 netcdf files)	53.5GB	364.1 GB	us-east-1

CMIP6	variable_ids = ['tas'] # tas is air temperature at 2m above surface
	table_id = 'Amon' # Monthly data from Atmosphere
	grid = 'gn' experiment_id = 'ssp245' activity_ids = ['ScenarioMIP', 'CMIP']
	institution_id = 'MOHC'
		1.13GB	0.11 GB	us-west-2

Number of Workers					
Compute	Region	2(CMIP) + 2(ERA)	2(CMIP) + 4(ERA)	2(CMIP) + 8(ERA)	2(CMIP) + 12(ERA)
CMIP6 (predicted_tas_regridded)	us-west-2	11.8	11.5	11.2	11.6
ERA5 (historic_temp_regridded)	us-east-1	1512	711	427	202
Difference (propogated pool)	us-west-2 & us-east-1	1527	906	469	251

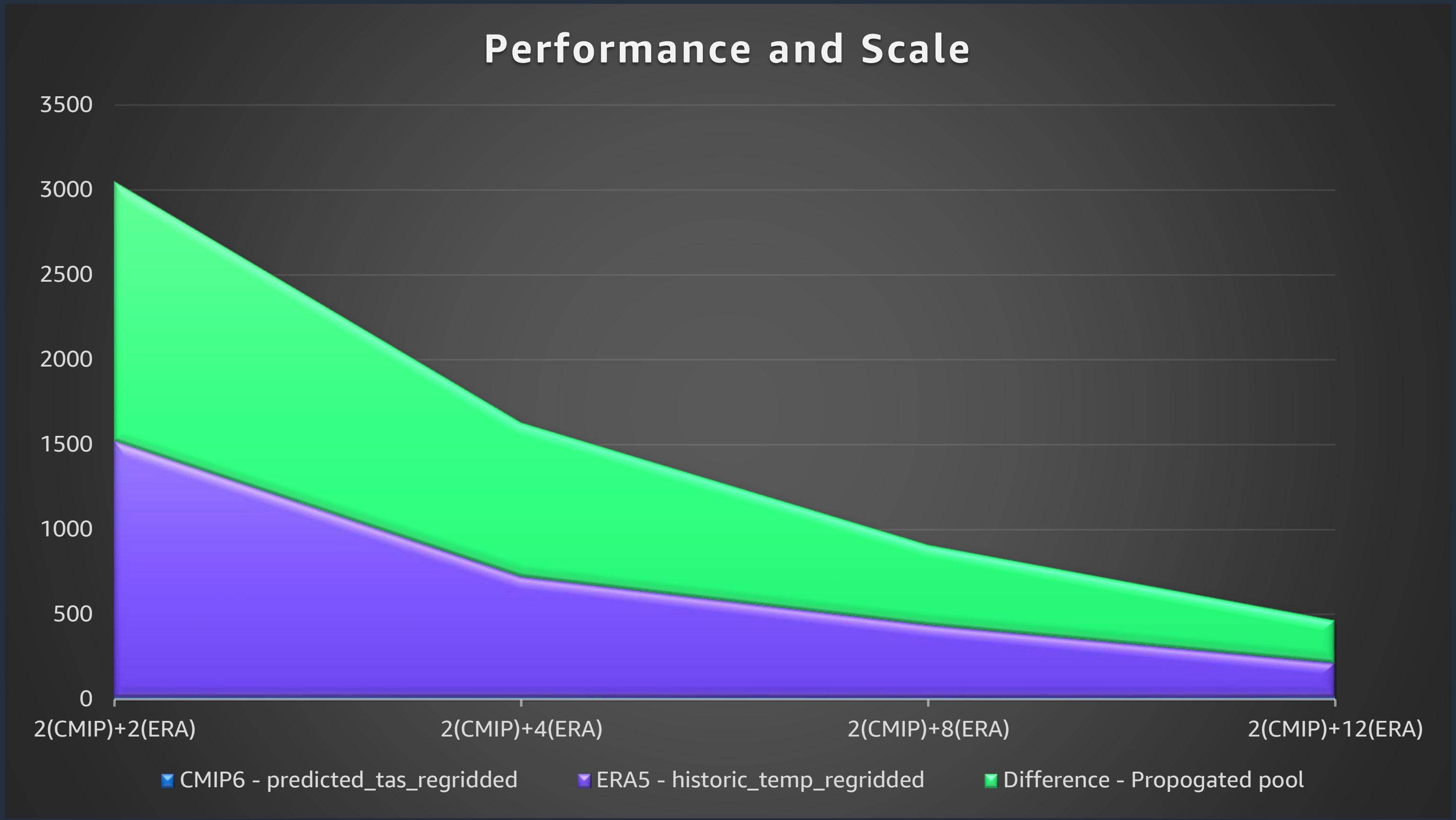
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling Performance

Workload decreases

HPC computation

Optimised Compute



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

"~15 seconds to compute this 20-year index. Subsequent thresholding is near instantaneous, and plotting is pretty quick too"

Richard Hattersley

Lead Technical Architect, UK Met Office

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Outcomes

Functional Outcomes

- Improves data discovery and loading
- Automates distributed compute
- Automates efficient orchestration
- Scientists spend more time exploring data

Technical Features

Non-Functional outcomes

- Enabling customers on AWS
- Project is Opensource (CDK deployment)
- Public Solutions Guidance

Opportunities that enable customers

Next Steps

- Community involvement
- CICD development

A path forward to encourage the adoption of the project

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits of Data Proximate Compute and Amazon FSx for Lustre

Climate Science

Climate data users can interact with big geospatial datasets, discovering new results, today made difficult because of slow time-to-insight.

Time

Estimated 65% time saving. If Met Office weather data was accessed using this architecture, up to 64 days of computing time could be saved every year compared to traditional approaches to accessing object stores.

Power

If this practice was adopted by users of Met Office data, the equivalent of 40 homes daily power consumption could be saved every day compared to traditional approaches to accessing object stores.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.