Met Office INFORMATICS LAB

Why Pangeo?

What is it and why we need it

Theo McCaie

hello@informaticslab.co.uk @informatics_lab informaticslab.co.uk

Met Office INFORMATICS LAB

Met Office Informatics Lab

Grains of sand

1	2	4	8	16	32	64	128
256	512	1024	2048	4096	8192	16384	32768
65536	131K	262K	524K	1M	2M	4M	8M
16M	33M	67M	134M	268M	536M	1G	2G
4G	8G	17G	34G	68G	137G	274G	549G
1T	2T	4T	8T	17T	35T	70T	140T
281T	562T	1P	2P	4P	9P	18P	36P
72P	144P	288P	576P	1E	2E	4E	9E

A mug full
(25th square)

The volume of a person
(33rd square)

5th largest container ship in the world
(53rd square)

A teaspoon full
(16th square)

In a cargo container
(39th square)

2x the worldwide container ship capacity
(64th square)

Generating more data in a week than in total 7 years ago

Increasingly big, increasingly complicated

Probabilistic / ensambles

Many formats

High dimensionality

Short use by date

Mechanistic / statistical

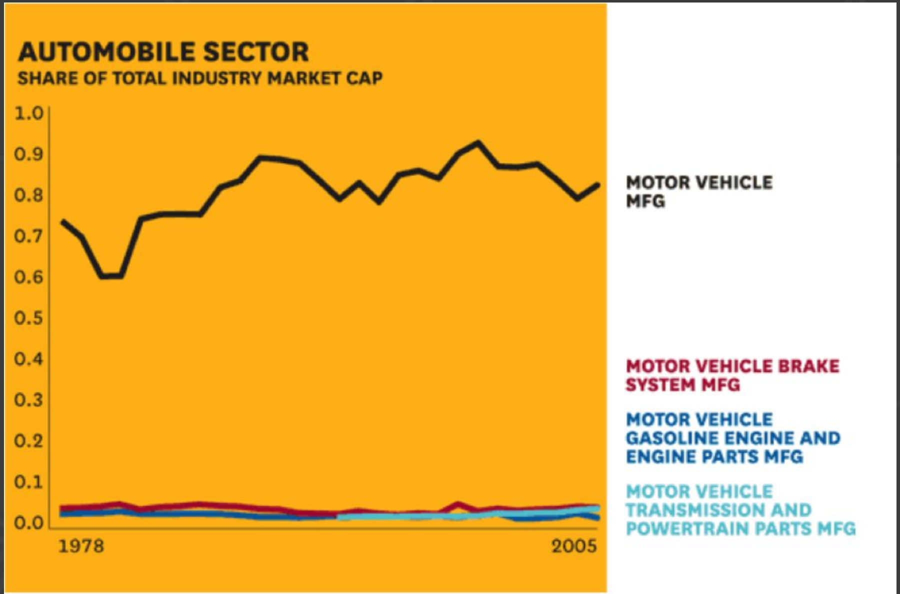
High volume

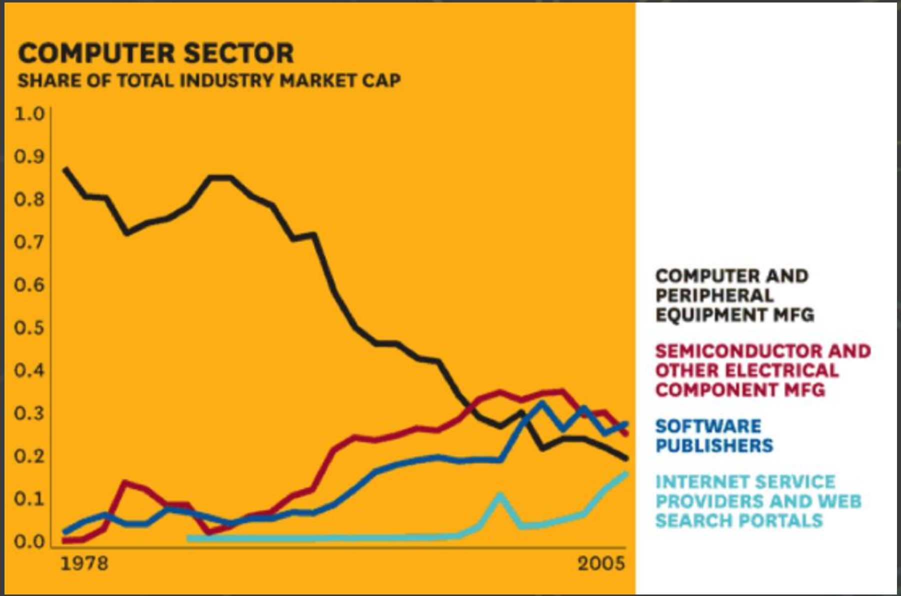
Inconsistent metadata

Hundreds of variables

High velocity

Total market capitalisation: Cars Vs Computers

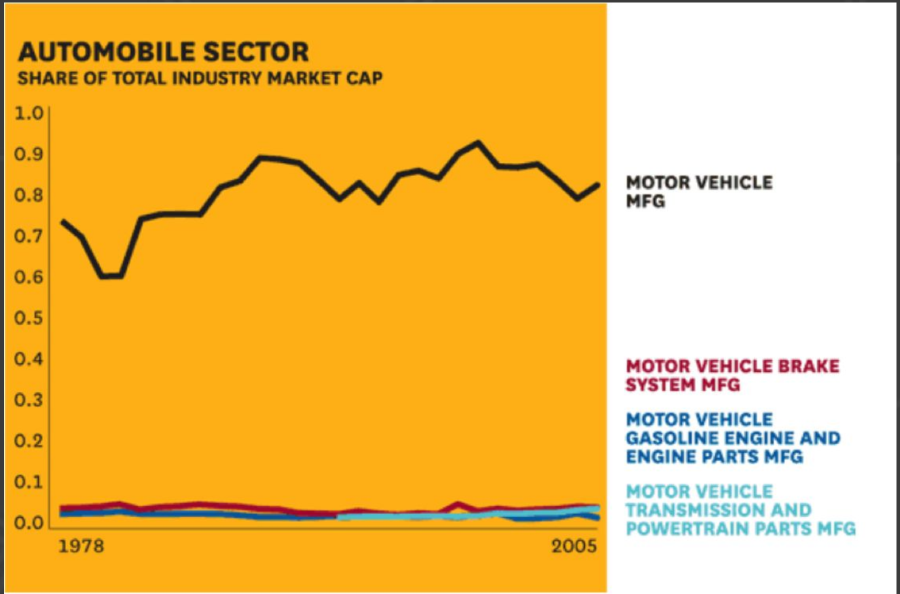


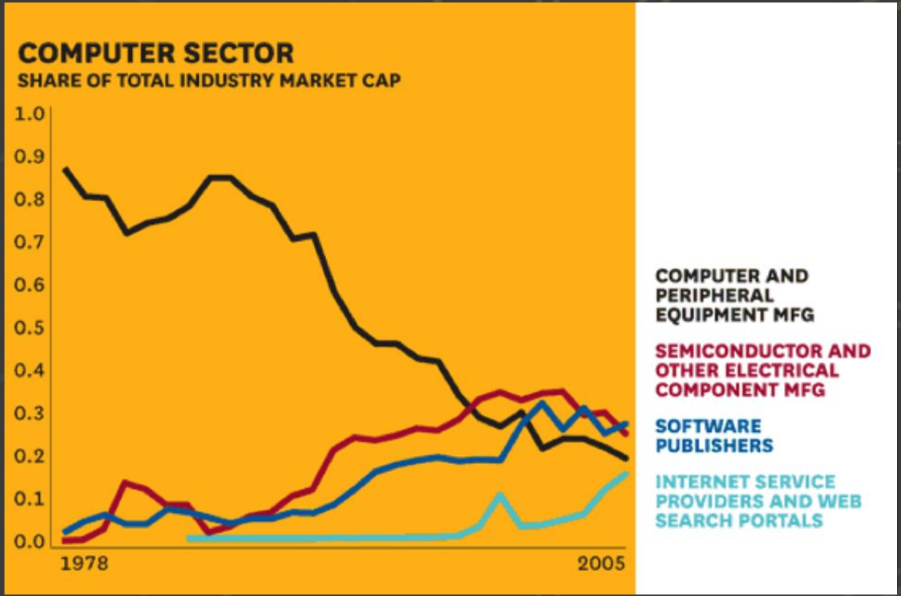


https://hbr.org/2013/07/how-to-drive-value-your-way

Met Office INFORMATICS LAB

Which one is weather forecasting?





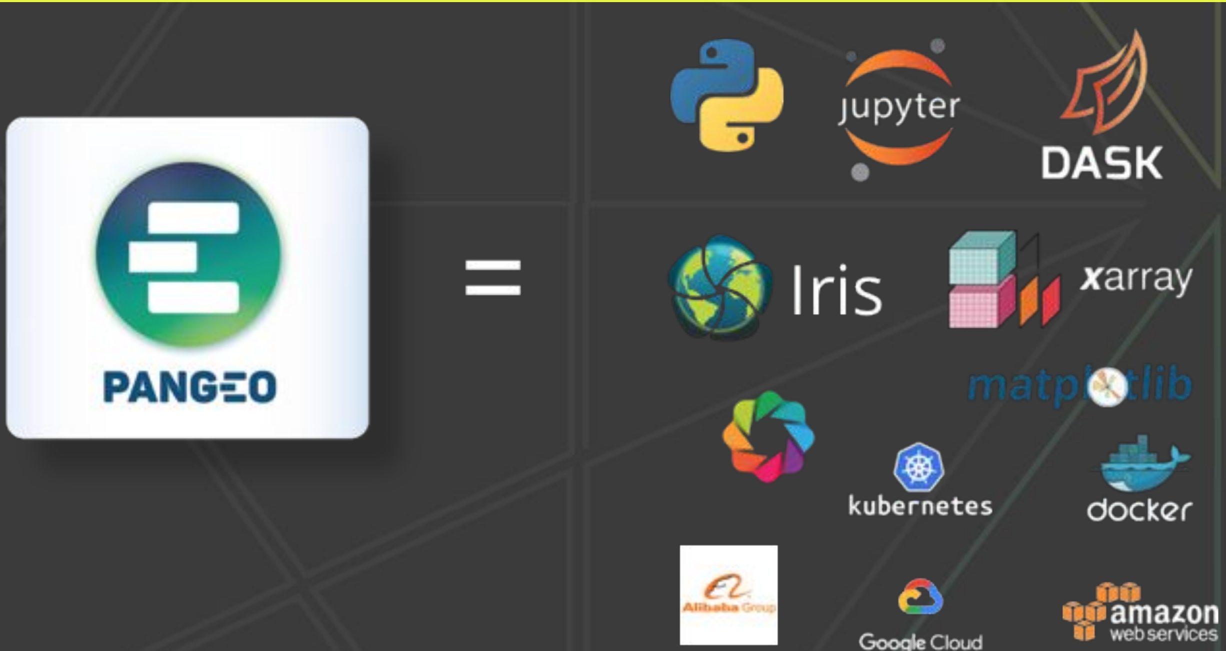
https://hbr.org/2013/07/how-to-drive-value-your-way

Met Office INFORMATICS LAB

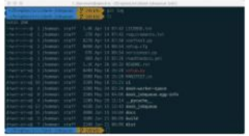
We are looking for a way to:

- Help people make better decisions based on complex data?
- Allow bespoke solutions to domain specific problems?
- Empower our ever more data-literate consumers?
- Unlock all the value in hundreds of PB of data?
- Provide improved accuracy without ever more flops?
- Empower analysts to do what they want not what they can?

Pangeo


PANGEO
=
jupyter
DASK
Iris
xarray
matplotlib
kubernetes
docker
Alibaba Group
Google Cloud
amazon web services

BUILD YOUR OWN PANGEO

Storage Formats	HDF	OPeNDAP	Cloud Optimized COG/Zarr/Parquet/etc.
ND-Arrays	NumPy	DASK	More coming…
Data Models	xarray	Iris	pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
Processing Mode	jupyter Interactive	Batch 	Serverless
Compute Platform	HPC	Cloud Google Cloud Platform	Local

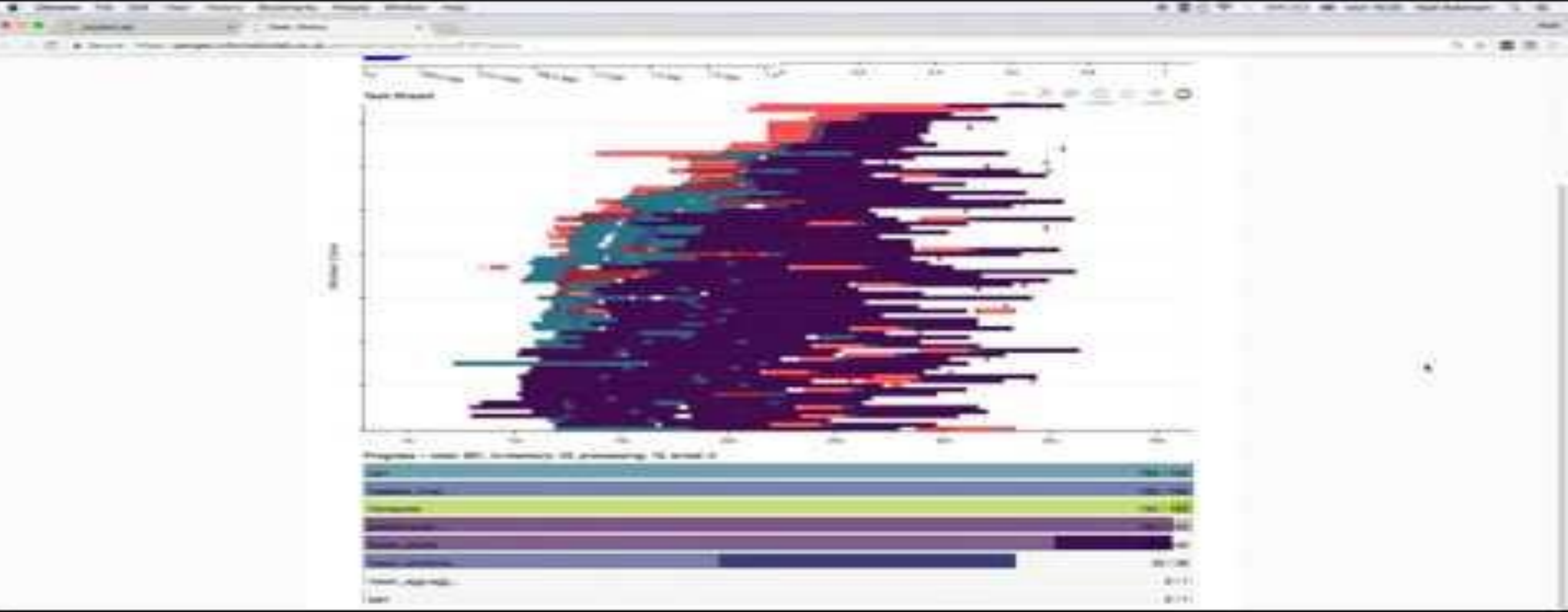
Iris

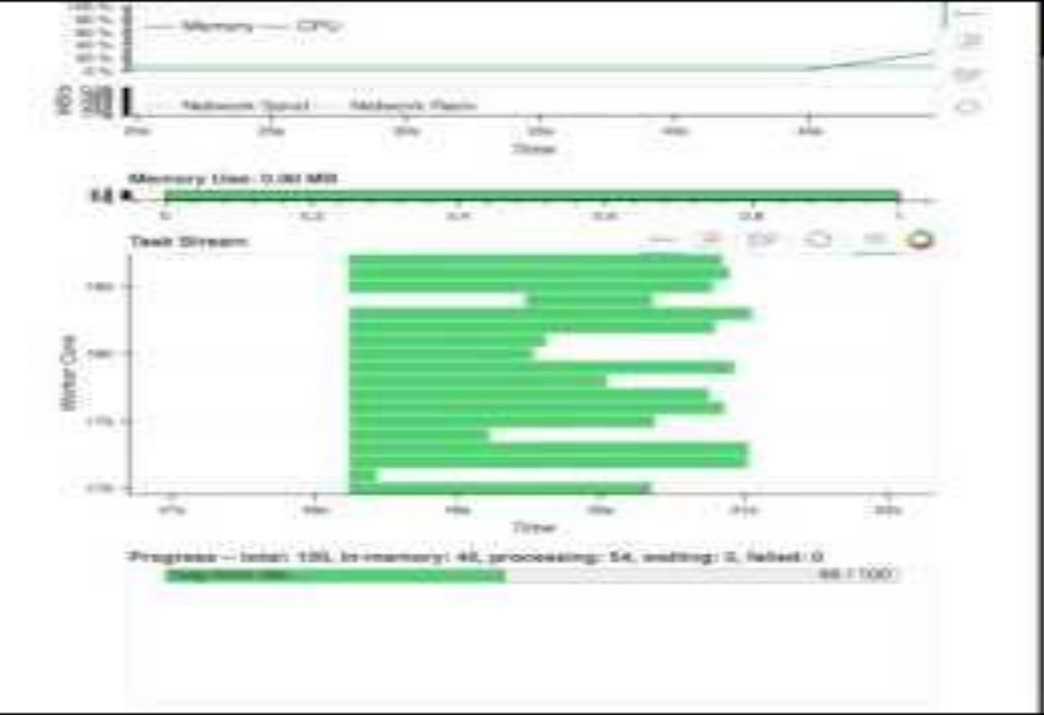
- Open Source (BSD license)
- Encapsulates Dask for distributed calculations
- Re-gridding/projection/interpolation
- Unit conversion
- Reads/converts various formats (Grib, NetCDF, fieldsfiles...)
- Automatic plotting via matplotlib and holoviews
- 8 FTEs at the Met Office working on it!

Pangeo

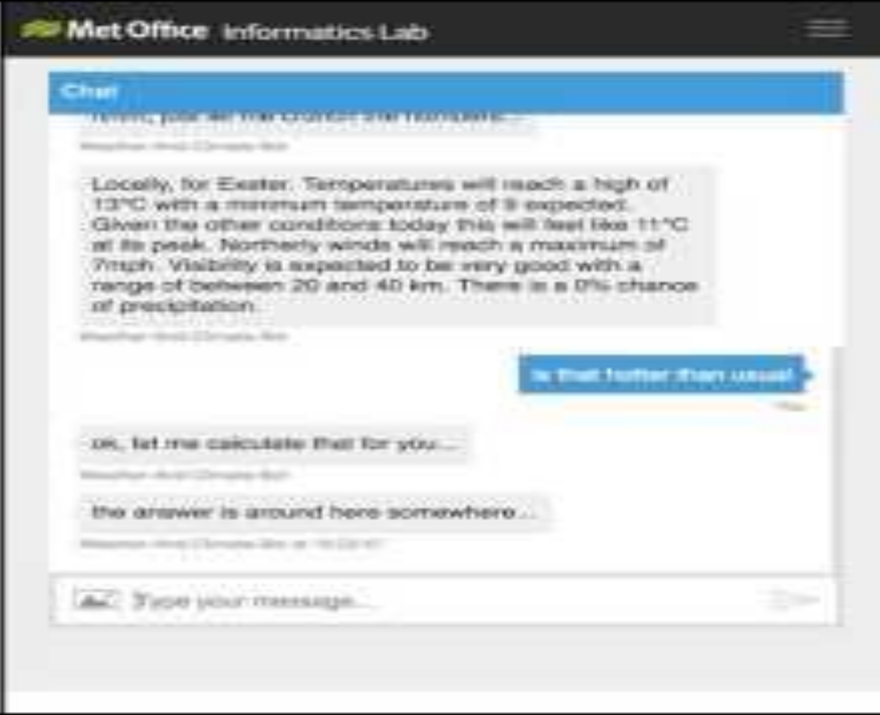
PANGEO =

- Responding to demand elastically
- Interactive analysis to encourage "flow", not fire-and-forget batch jobs
- Laziness/just-in-time
- Thin web client views to interact with data
- Agile, bespoke, product creation


Task Stream
Progress



Progress -- total: 150, in-memory: 46, processing: 54, waiting: 3, failed: 0



hmm, just let me crunch the numbers...

Weather And Climate Bot

Locally, for Exeter. Temperatures will reach a high of 13°C with a minimum temperature of 9 expected. Given the other conditions today this will feel like 11°C at its peak. Northerly winds will reach a maximum of 7mph. Visibility is expected to be very good with a range of between 20 and 40 km. There is a 0% chance of precipitation.

Weather And Climate Bot

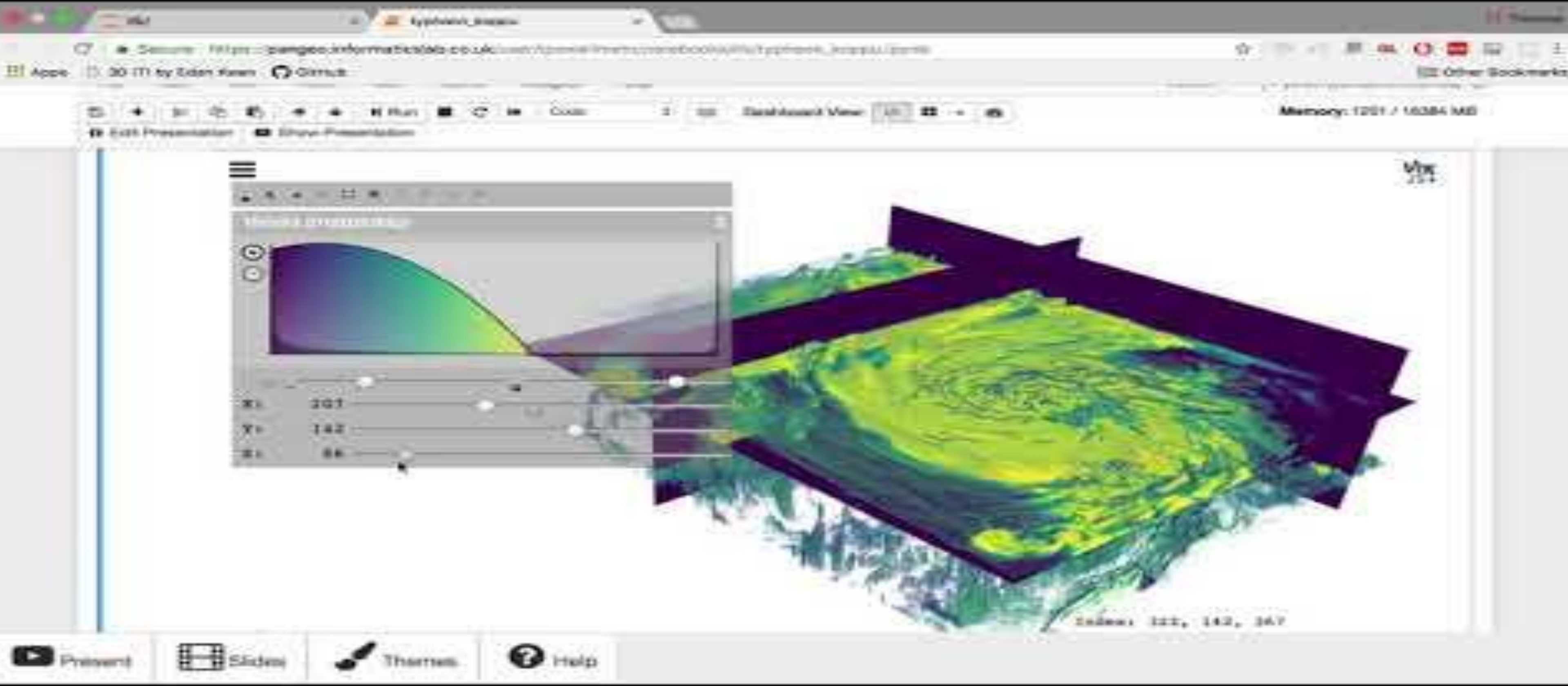
is that hotter than usual

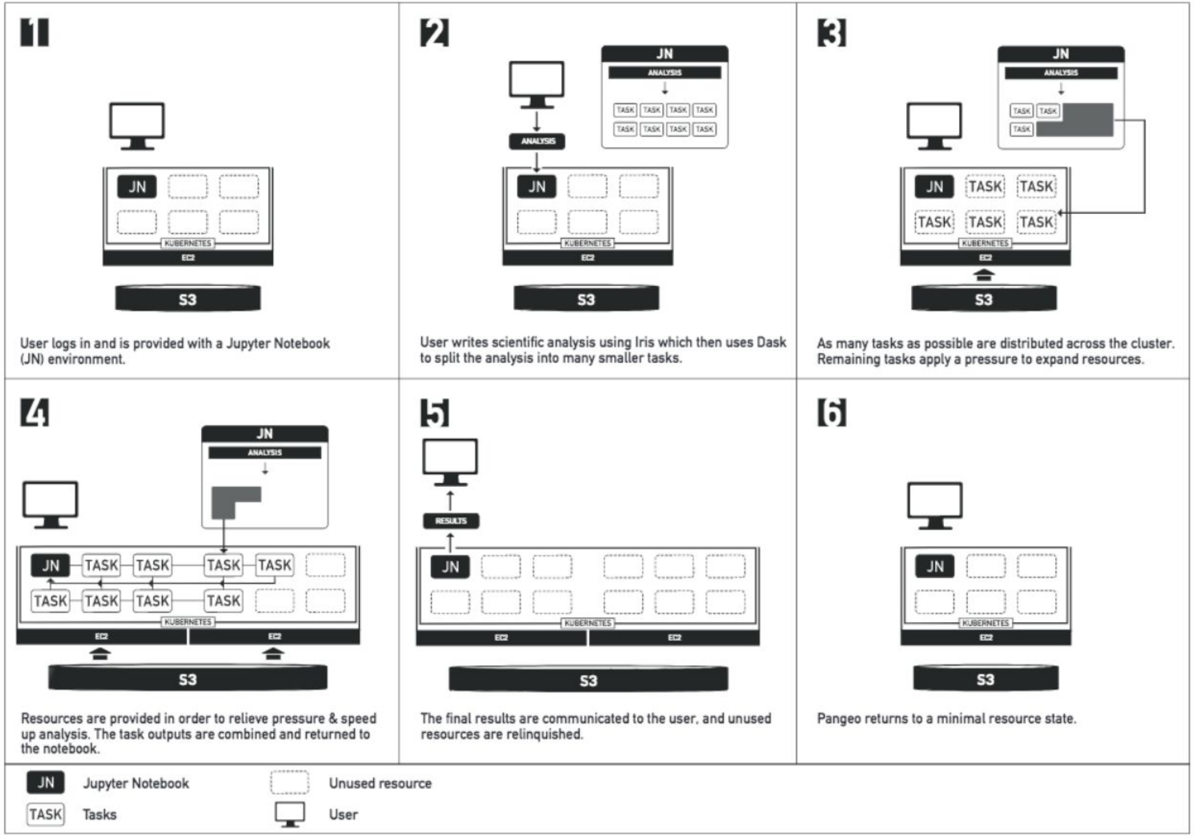
ok, let me calculate that for you...

Weather And Climate Bot

the answer is around here somewhere...

Weather And Climate Bot


typhoon_...
Secure
https://pangeo.informaticslab.co.uk
Apps
GitHub
Other Bookmarks
Run
Code
Dashboard View
Memory: 1201 / 16384 MB
Edit Presentation
Show Presentation
x: 207
y: 142
z: 86
Index: 122, 142, 167
Present
Slides
Themes
Help



1. User logs in and is provided with a Jupyter Notebook (JN) environment.

2. User writes scientific analysis using Iris which then uses Dask to split the analysis into many smaller tasks.

3. As many tasks as possible are distributed across the cluster. Remaining tasks apply a pressure to expand resources.

4. Resources are provided in order to relieve pressure & speed up analysis. The task outputs are combined and returned to the notebook.

5. The final results are communicated to the user, and unused resources are relinquished.

6. Pangeo returns to a minimal resource state.

JN	Jupyter Notebook	⬚	Unused resource
TASK	Tasks	🖥	User

https://bit.ly/2O9qJr3

On the cloud?

- Doesn’t have to be but…

Workloads are volatile by nature

On Prem this gives two options:

- Very big cluster
- quick results
- Inefficient utilisation

vs

- Smaller cluster
- suppress volatility with queues
- higher utilisation

Volatile workloads in the cloud

Fast

↕

Slow

1. Scheduler creates many jobs for an individual user
2. Many schedulers submit to the same orchestrator, smoothing volatility somewhat
3. Orchestrator asks for more cloud resources in response to spikes in demand
4. Many Many users use the cloud smoothing demand a lot
5. The Cloud installs more racks in response to demand

What’s next

- Data discovery - “Hey pangeo, get me data on UK storms last year”
- Science to service - one click APIs and dashboards.
- Environmental data analysis for all - deployments, tools, APIs and tutorials to make environmental data accessible to SMEs
- On demand hardware configured clusters
- Improved data storage and handling

What’s next: our data challenge

- Fast metadata access
- Parallel and elastic over fast and immediately consistent
- Universally addressable
- Reluctant to embrace one size fits all
- Still need to deal with the high volume and velocity data as is currently generated

Zarr

Why pangeo?

Met Office

JADE

Alfred P. Sloan FOUNDATION

NASA

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

NSF EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

Met Office

W
UNIVERSITY of WASHINGTON

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

ANACONDA.

Element 84

Environmental Futures & Big Data

Impact Lab

University of Exeter

Rothamsted Research

Plymouth University

Exeter CityFutures

Plymouth College of Art

European Union
European Regional Development Fund

PML Plymouth Marine Laboratory

Met Office

Innovation is hard… and always has been

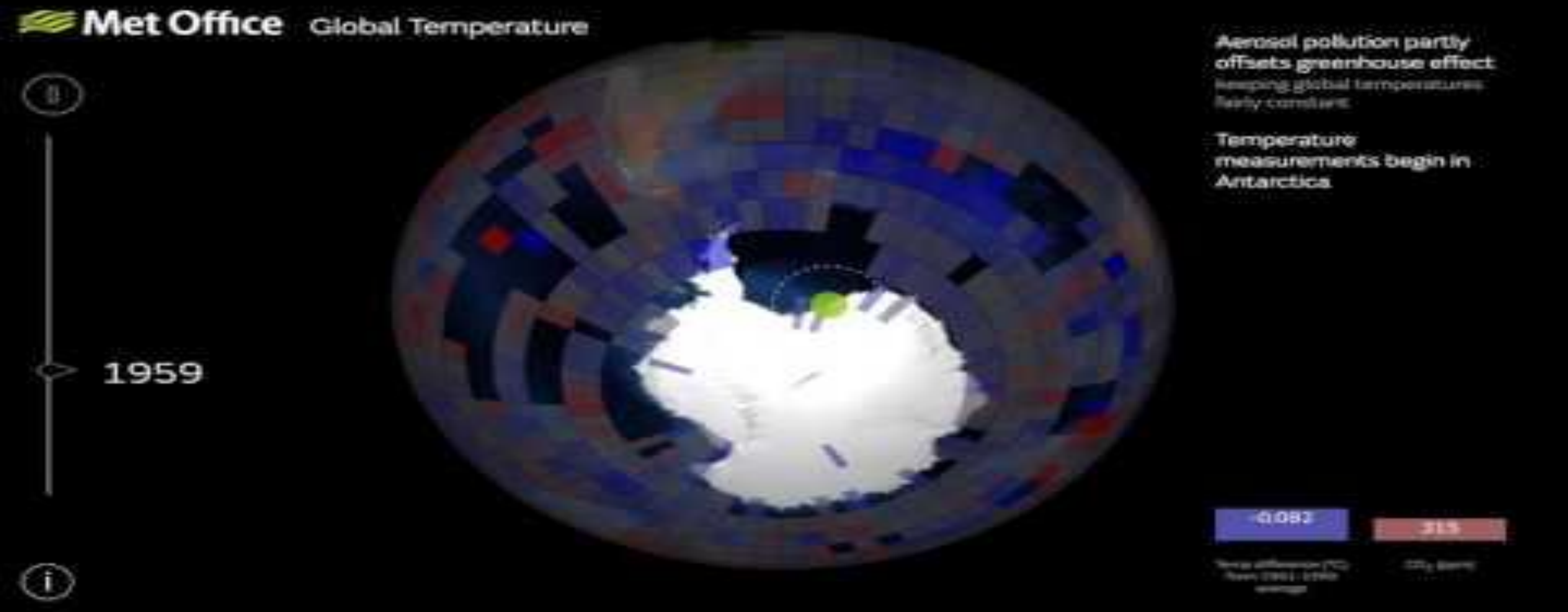
introduction of innovation…is a struggle against stupidity and envy, apathy and evil, secret opposition and open conflict of interests, a horrible period of struggle with man, a martyrdom even if success ensues.

Diesel, 1858-1913

Met Office INFORMATICS LAB

Thank you!

✉ hello@informaticslab.co.uk @informatics_lab informaticslab.co.uk



Met Office Global Temperature

1959

Aerosol pollution partly offsets greenhouse effect
keeping global temperatures fairly constant

Temperature measurements begin in Antarctica

-0.083	315
Temp difference (°C) from 1961-1990 average	CO_2 (ppm)

Met Office INFORMATICS LAB