PML | Plymouth Marine Laboratory

Listen to the ocean

Can EO afford big data
An assessment of the temporal and monetary costs of existing and emerging big data workflows

Oliver Clements & Peter Walker

Introduction

The storage, delivery and analysis of big data is a `hot topic`. Many new technologies are appearing, should we simply adopt these without good reason?

NO

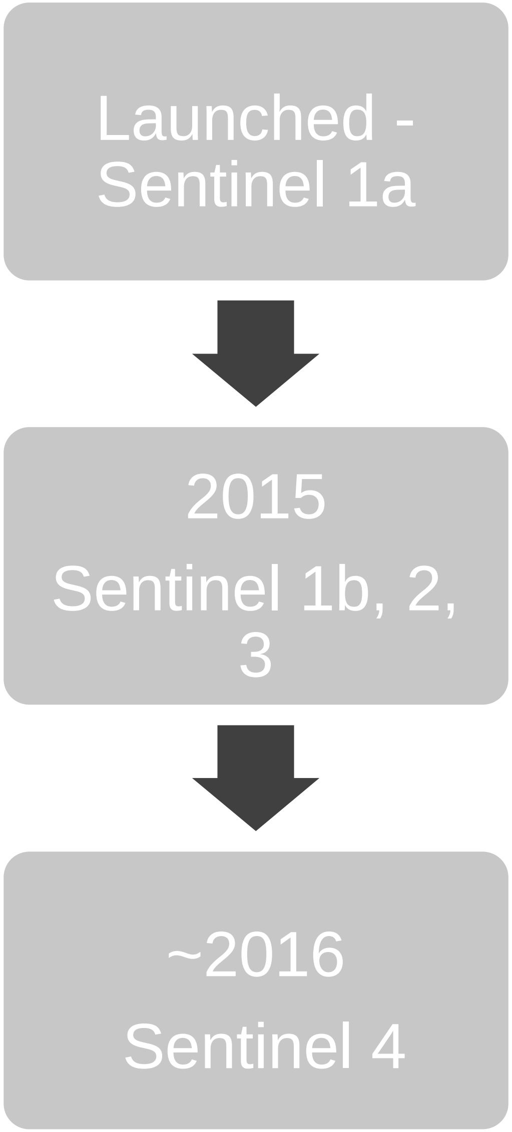
- Data
- Costs
- Engaging new types of users –" Citizen scientists"
- Workflows

Data –It's getting bigger

- Data volume is ever growing
- 40 years ago < 100Mb
- Today > 1Pb
- Remote sensed EO data makes up a large proportion
- More powerful systems are allowing models to create huge datasets

Data – Sentinel programme

- 5 satellites over the coming years
- Just Sentinel 1 will produce ~ 2.5 Tb or data per day
- When the programmes is finished it will be producing ~ 8Tb per day
- Dealing with this will be a major challenge



Launched - Sentinel 1a

2015

Sentinel 1b, 2, 3

~2016

Sentinel 4

Data -challenges

This influx of data will create challenges in several fields

- Storage
  - Storing all the data will be very costly
- Transfer
  - For users to actually get value from the data they will have to be able to discover it and transfer the data over networks
  - To effectively do this a very fast internet connection would be needed
- Analysis
  - To be able to produce answers to scientific questions users will have to process unprecedented volumes of data

COSTS

Costs -Monetary

If we take the three identified necessary elements for hosting big data -storage, transfer & processing

Storage

- Most accurate pricing is a well kept secret (unless you buy)
- Based on our experience for network storage you could pay as much a £5000 for ~ 10Tb of high end disk and controller

Transfer (networking)

- Again industry pricing is a well kept secret
- For a network switch architecture with appropriate number of ports you could pay around £5000 again

Processing

- Costs here can range from £3-5k to as much money as you can spend, E.G. a recent NERC big data infrastructure spent around £80k on processing nodes.

Costs -Temporal

As well as the monetary costs, temporal costs must be considered. Both data transfer and processing take time.

For instance, most users will not be on the same network as the data archive, this results in a severe penalty for data transfer.

Internet speeds are getting faster, but not everywhere

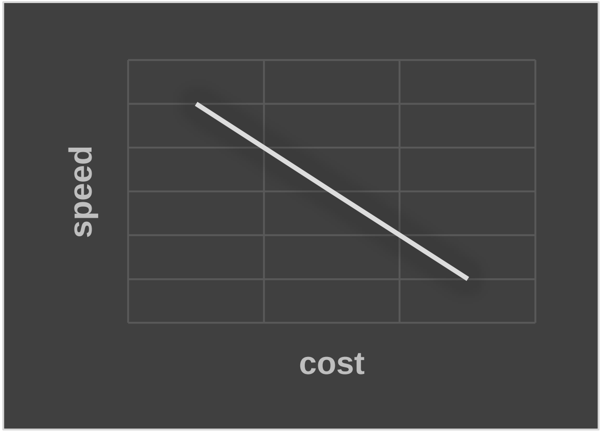
- Europe : ~30mb/s
- Americas : ~10mb/s
- Asia : huge variation 60mb/s – 4mb/s
- Africa : ~3mb/s

Taken from ooklaglobal test data

Costs -Temporal

Once a user has the data, processing times can again be huge.

There can be a trade off here of monetary cost vs temporal cost



Processing speed vs infrastructure cost

The more expensive your processing infrastructure the quicker it will be. But even on a > £100k system you could be taking hours if not days to process over a long time series

This creates an environment where most time is spent getting and processing data that actually doing science with the results!!

Costs – Citizen Science

The costs discussed are a major barrier to citizen science and scientists in the developing world. Both the processing and data transfer times make use of EO big data a challenge for these groups.

Citizen scientists in the developed world :

- Have access to high speed internet but probably no processing infrastructure

Scientists in the Developing world :

- Do not have access to either high speed internet or processing infrastructure

How can we fix this with modern workflows patterns?

Workflows

There are several existing and emerging workflows available to users and data providers.

- Simply download all the data and process locally
- Download a subset of data and process locally
- Utilise predefined processing at the data provider and download result
- Utilise user defined ad-hoc processing at the data provider and download result

Workflows – download and process locally

This can probably be described as the classic workflow. It was popularised when no data transfer standards existed and data were relatively small. This workflow suffers from unnecessary data download if the user only requires a subset for processing.

- Transfer
  - FTP
  - HTTP
  - DVD or Hard Disk (still used today!)
- Processing
  - Custom code
  - Any language

Workflows – subset data and process locally

This is an adaptation on the previous workflow that utilises some, now, well established standards. By only downloading the geo/temporal bounds required you can minimise the data transfer cost. There is still a potentially considerable processing cost. For this test we used the WCS server THREDDS.

- Transfer
  - Web Coverage Service
  - OpenDAP
  - ThreddsNetCDFsubset service
- Processing
  - Custom code
  - Any language

Workflows – predefined processing @ the data

With more advancements in standards a new workflow emerged. By utilising the OGC standard Web Processing Service a user can request a predefined process chain be applied to data and then only download the result. For this test we used PyWPS.

- Transfer
  - Web Processing Service
  - HTTP
- Processing
  - Only predefined processing tools
    - This is the biggest weakness if the data provider does not have the processing tool the user requires

Workflows – user defined processing @ the data

This is the latest workflow. It takes advantage of a new OGC standard called Web Coverage Processing Service. This standard allows user defined processing to be run at the data provider with only the result being downloaded. The Analysis engine used by this system (Rasdaman)is also a highly optimised Array database

- Transfer
  - Web Coverage Processing Service (WCPS)
  - HTTP
- Processing
  - User defined processing using the WCPS query language
    - This is the biggest advancement, it allows the user to make use of the processing power at the data provider and only download the result

Analysis of workflows

Scenario : Calculating an average chlorophyll concentration for a given geospatial area and a given time period.

Data : OC-CCI 8460X4310px daily composites

Temporal bounds :

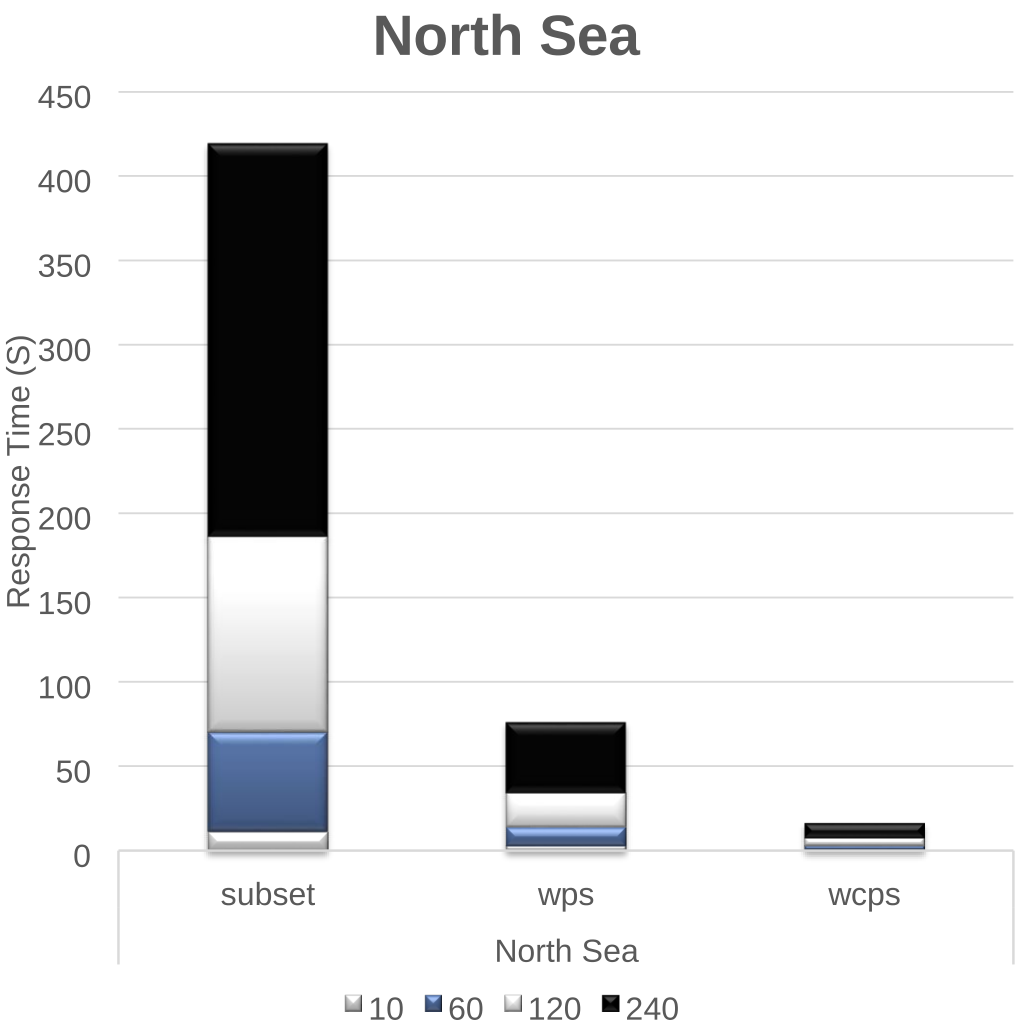
- 10 days
- 60 days
- 120 days
- 240 days

Geospatial area :

- Africa(-60,-65,88,45)
- North Atlantic (-66,-1,-4,68)
- North Sea (-3,52,7,61)

Analysis – North Sea

- Local processing of complete dataset not plotted as the download time alone is ~10hrs for 20Gb @ 5Mbps
- The response time for downloading then processing a subset is considerably faster than downloading the whole dataset
- However both the modern workflows presented a significant improvement
- Primarily because only the result is downloaded
- The final workflow, WCPS, provides yet another speed gain due to the efficient data access and analysis.

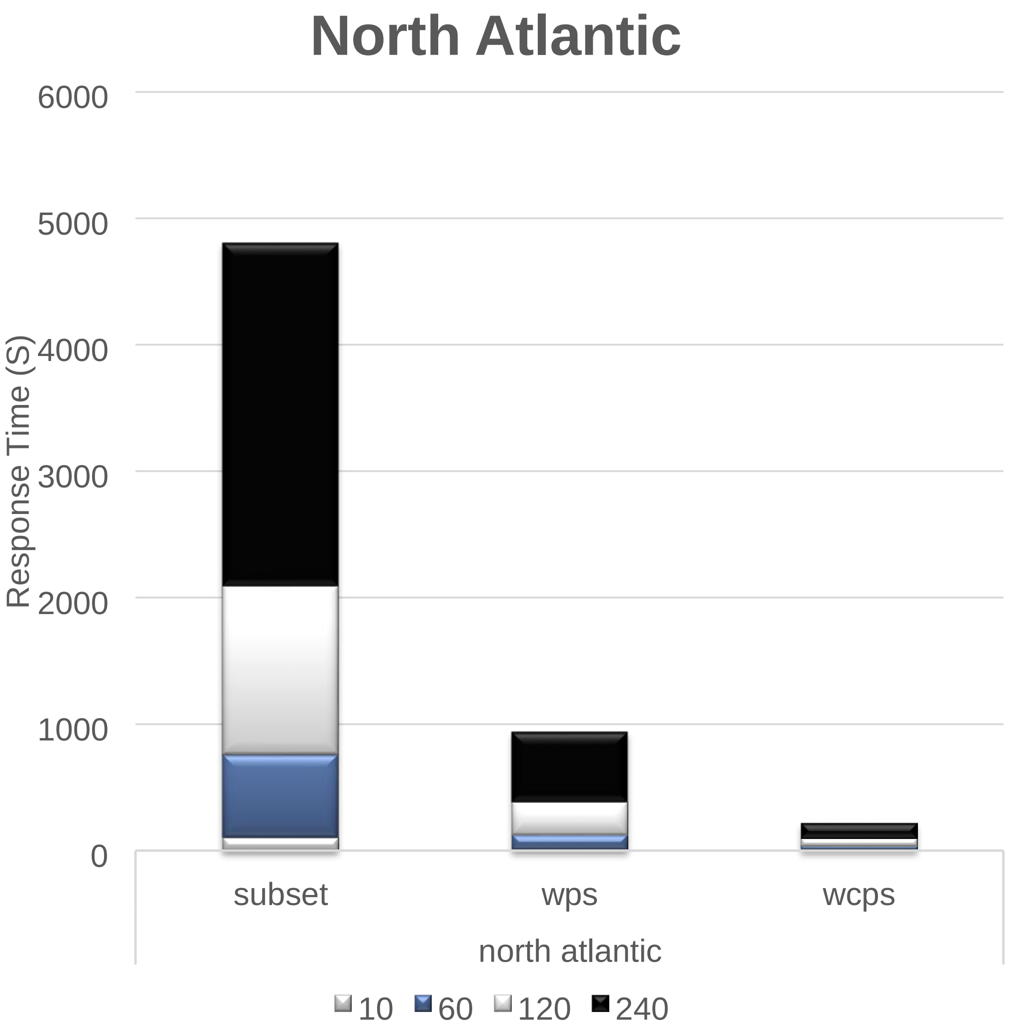


Analysis – North Atlantic

- Again local processing has been removed due to the download time overhead. There would also be a storage cost that might not be able to be met by users in the developing world

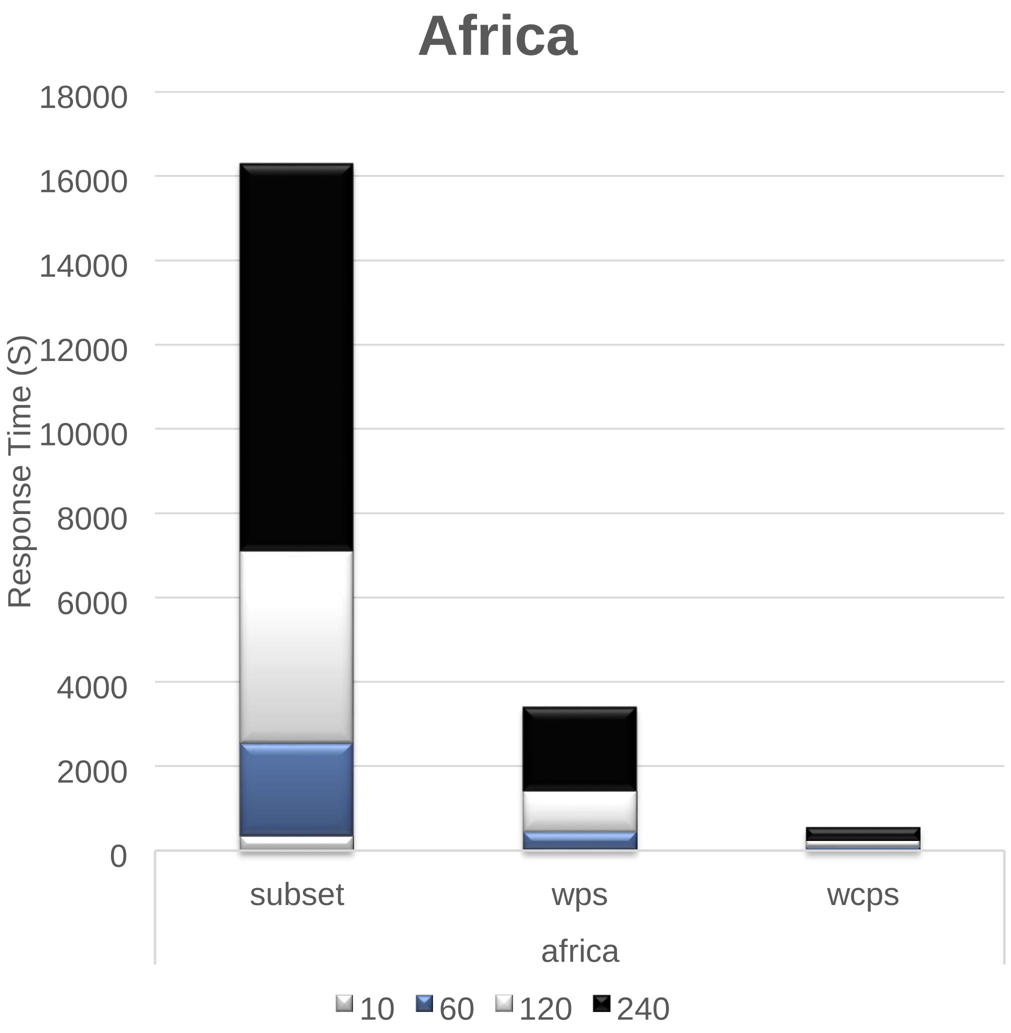
- The Gap between sub-setting and WPS is slightly smaller but still very large

- The processing time difference between WPS and WCPS is still significant



Analysis -Africa

- The Gap between sub-setting and WPS is slightly smaller again.
- The processing time difference between WPS and WCPS is still significant
- All of these tests assume a download speed of 5Mbps. This is to emulate the lower global average.
- To get the sub-setting anywhere near the WCPS result you would need > 100Mbps but the processing would still be slower



Conclusion

- The volume of data being produced and made available to researchers and the public is growing rapidly
- To make the data accessible to as many users as possible we must make systems available that protect the users from the costs associated with data storage and data processing
- To allow analysis to be carried out by citizen scientists and users in the developing world we must minimise data transfer as well as processing
- The existing workflows provide some of these features, but new emerging workflows that utilise standards such as WPS and WCPS provide the greatest benefit to the users.
- Data providers should embrace the new standards

Thanks For Listening

Questions?

Big Data Management and Visualisation -Earth Observation Community Workshop

Satellite Application Catapult -Harwell, Oxfordshire

8th May 2014

Register for free @

http://earthserver.pml.ac.uk/portal/big_data_workshop

Thank you