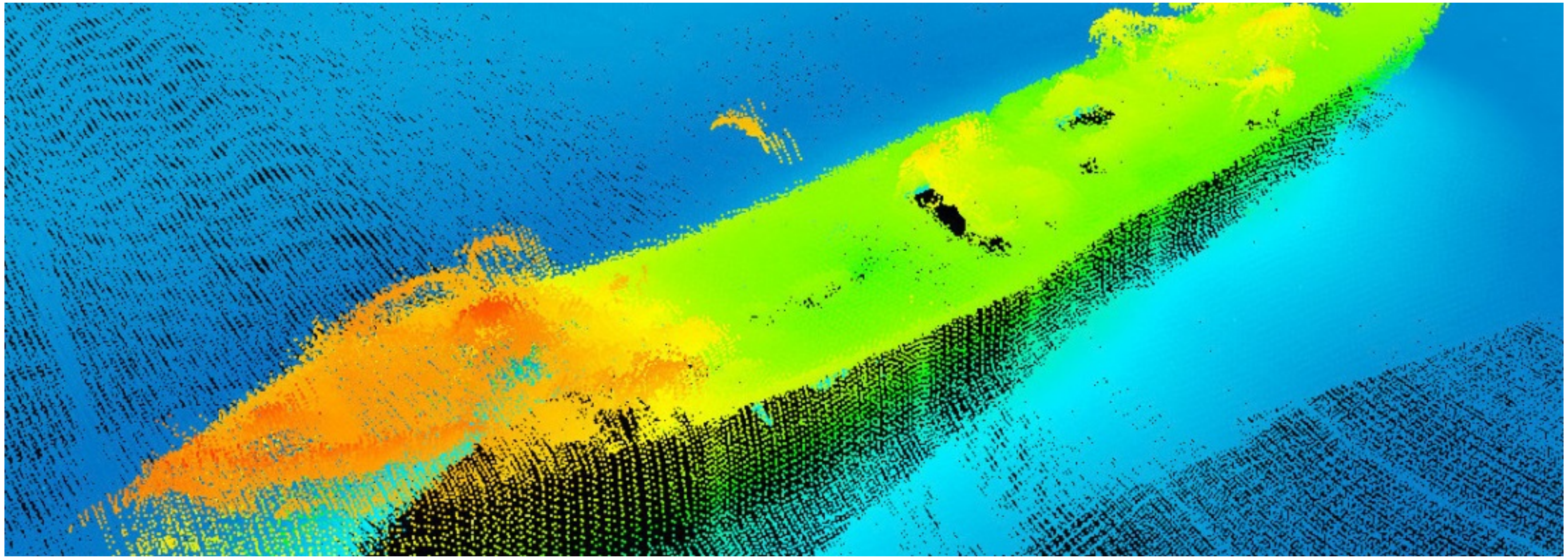
High Performance Computing Approaches for Processing Hydrographic Data

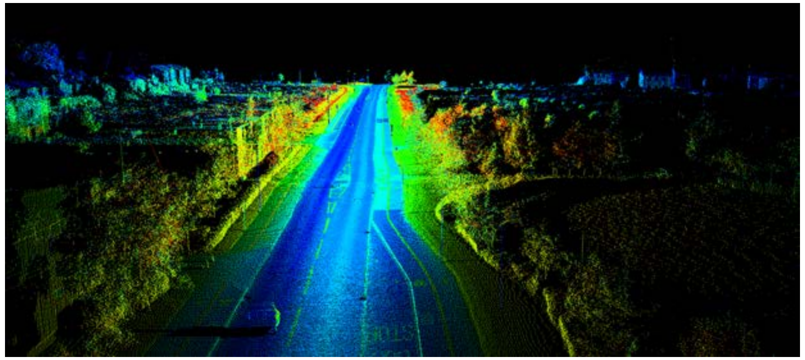


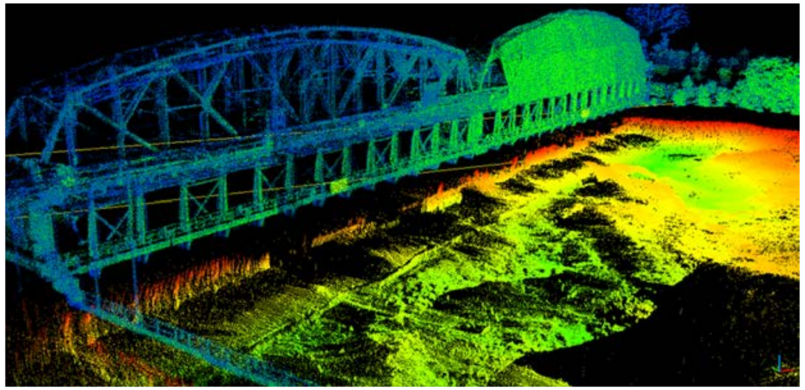
Australian Government
Geoscience Australia

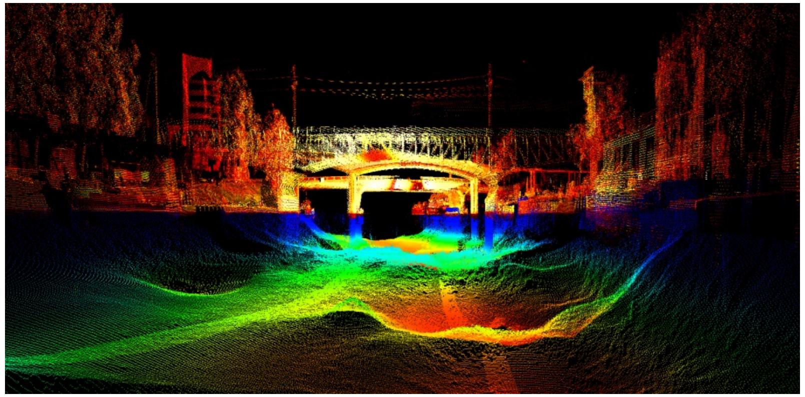
Australian Government
Department of the Environment
Australian Antarctic Division

The need

- Large volumes of information-rich point data are becoming increasingly available
- Greater volumes of data can mean greater detail – but regional-scale mapping requires large amounts of computing power
- Centralisation can be difficult with multiple providers, constant updates, different submission formats etc.
- But transfer speeds and costs can also be a bottleneck
- The ability to process large quantities of information in a distributed manner is needed -> High Performance Computing



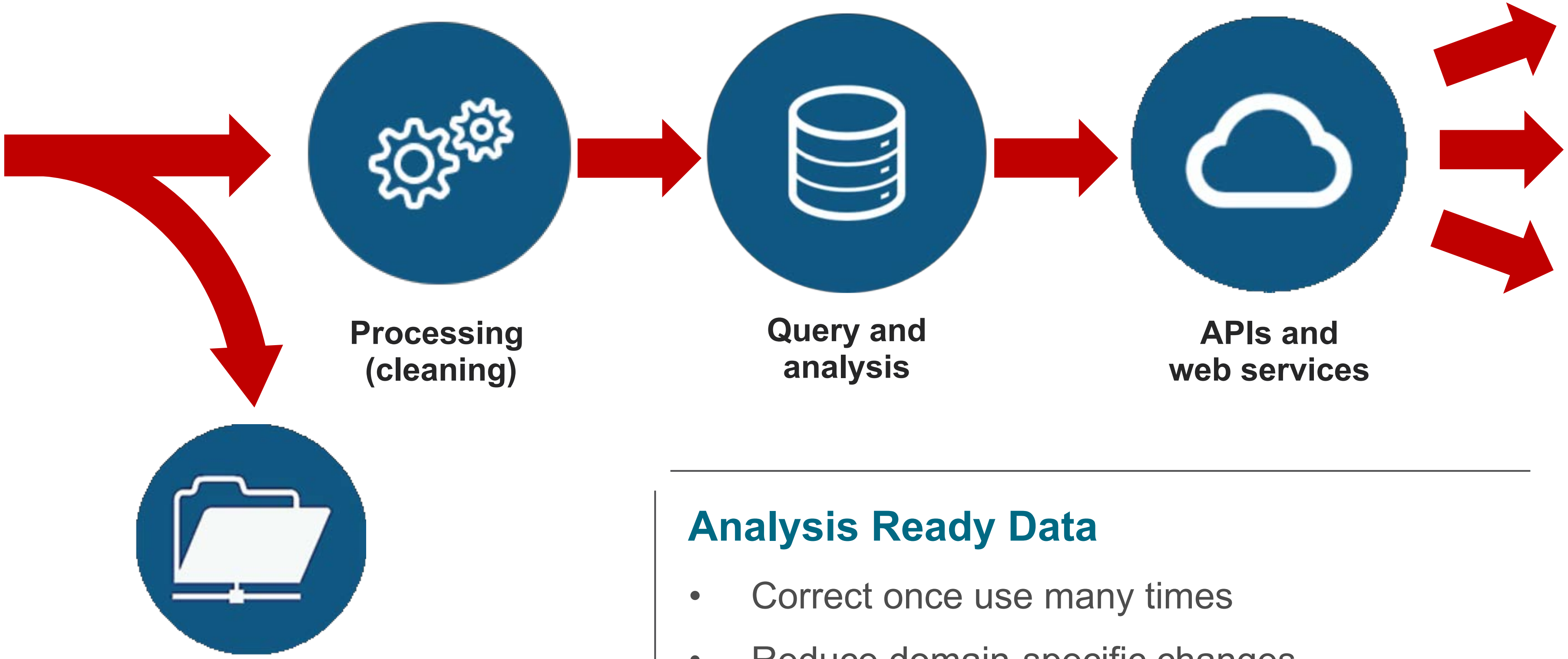




Survey to Service

Marine survey

Airborne survey



Processing (cleaning)

Query and analysis

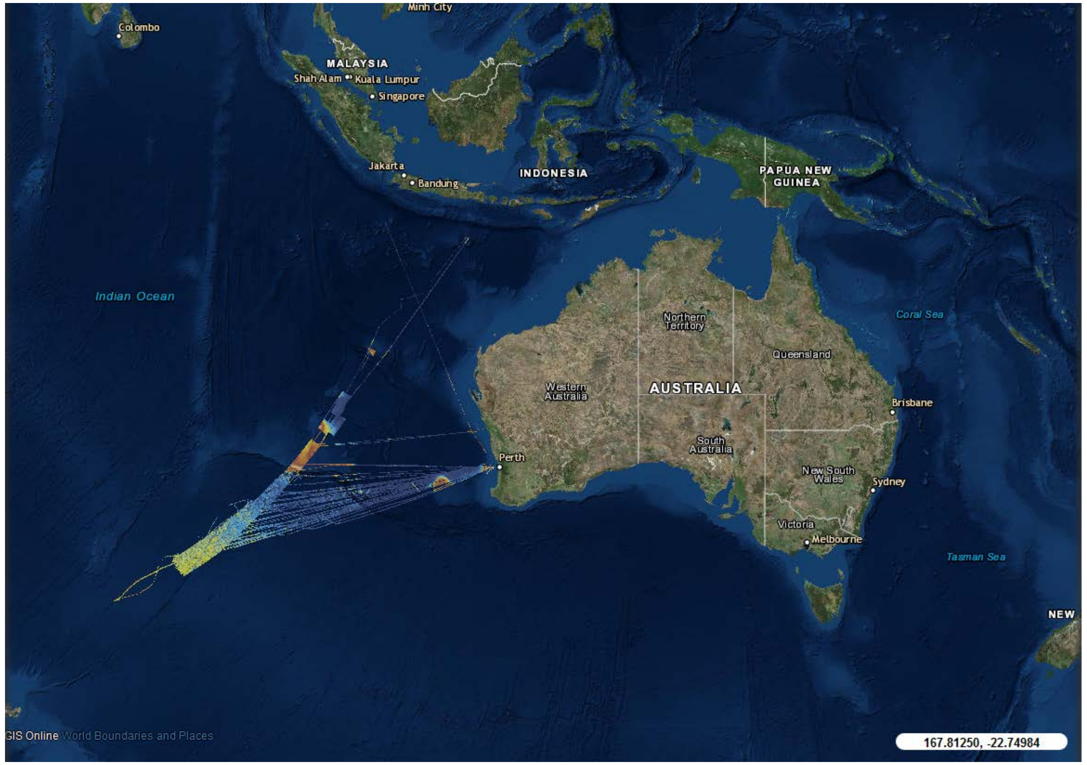
APIs and web services

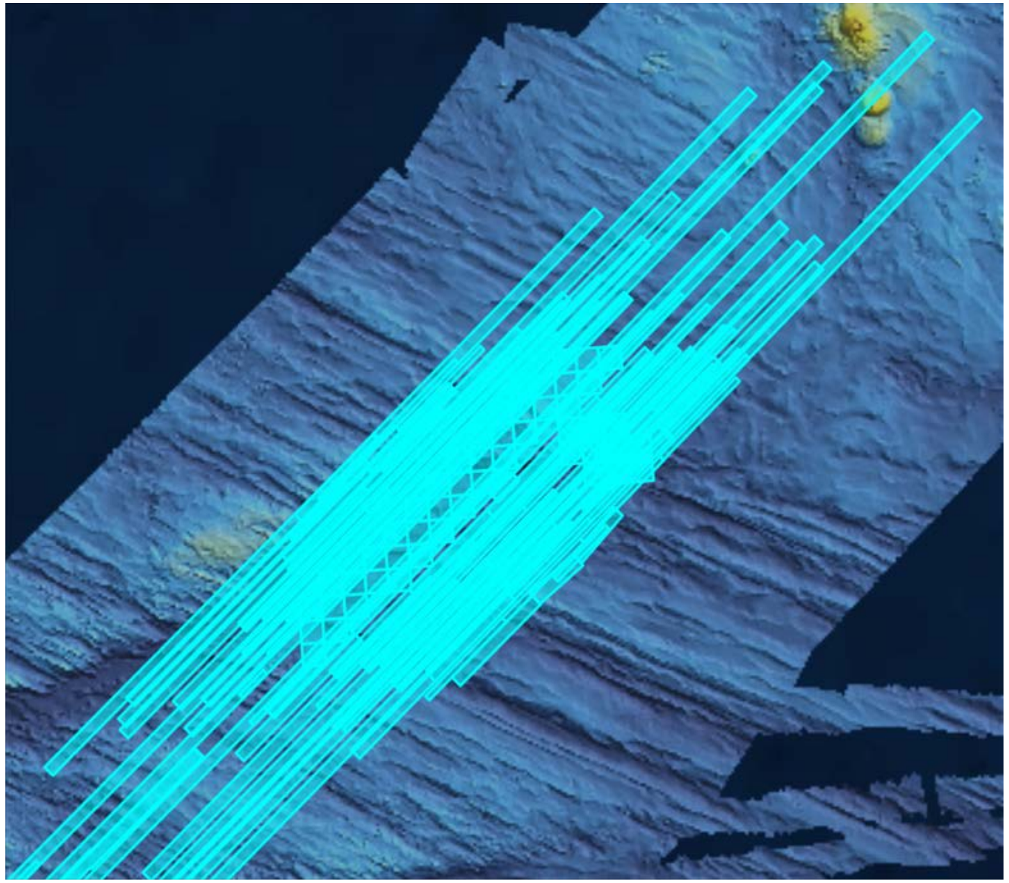
Raw archive

Analysis Ready Data

- Correct once use many times
- Reduce domain-specific changes
- Correct up to the point before products ‘branch’
- Self-describing data

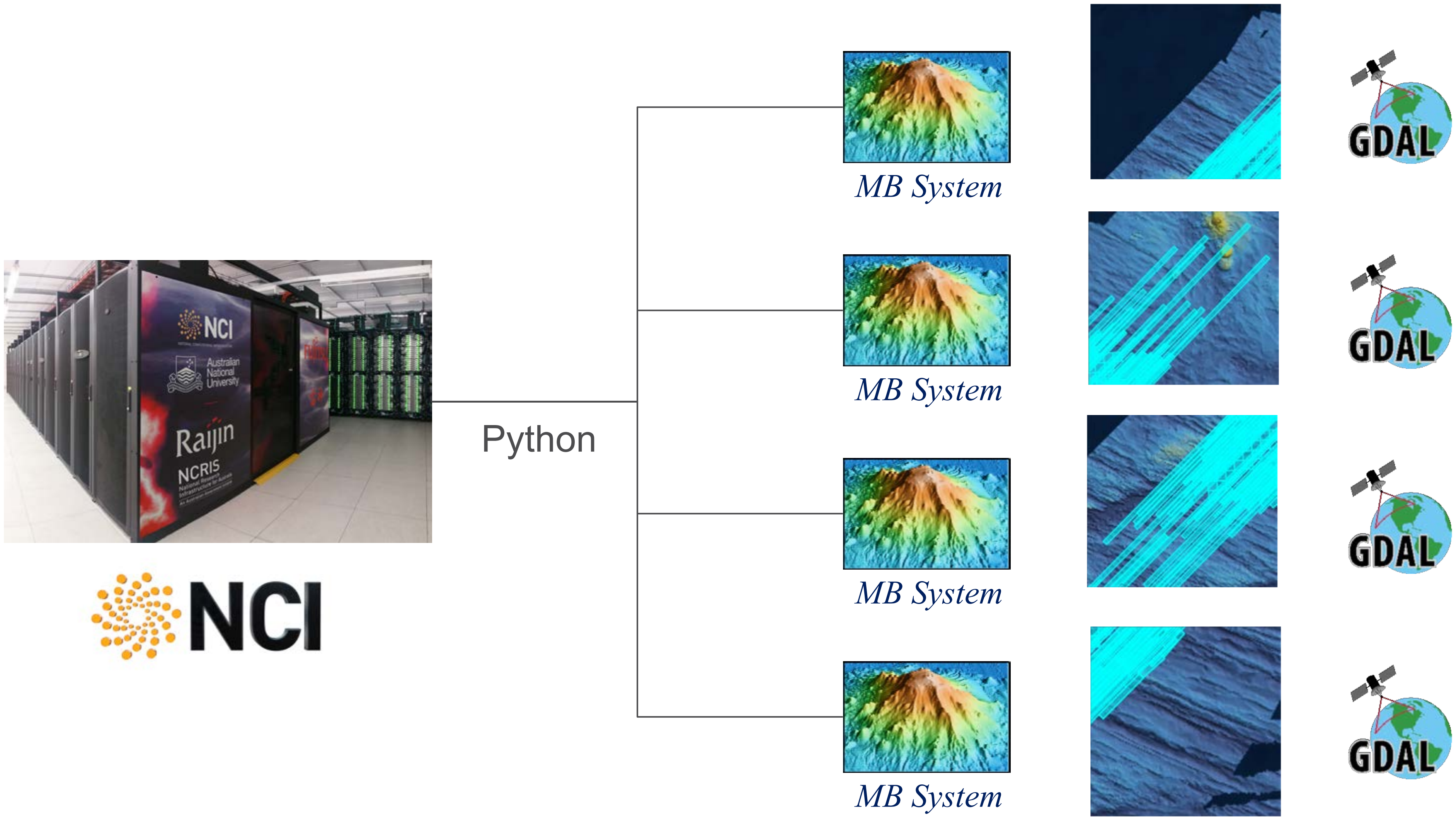
Building footprints for raw data





http://marine.ga.gov.au

Concurrent processing at NCI



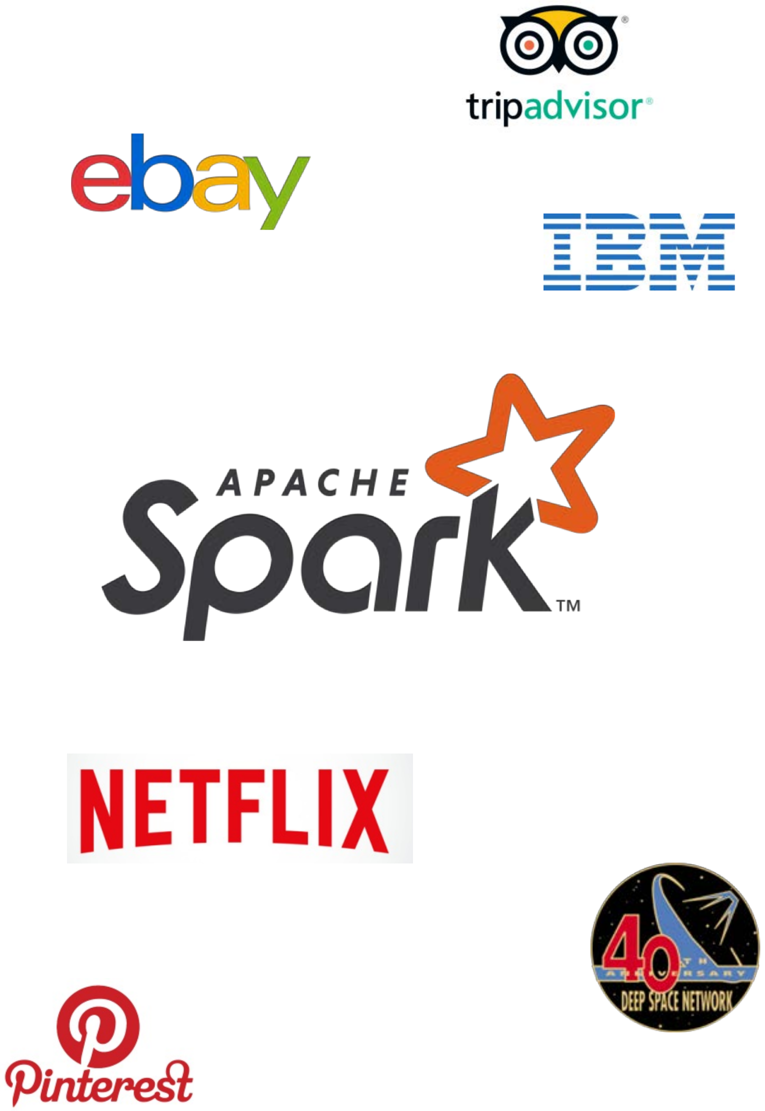
Days to minutes

Apache Spark

- Provides a means of performing scalable computing across multiple (possibly virtual) machines

- Can read data distributed across machines and platforms (e.g. reading directly from S3 buckets, databases, Lustre, HDFS)

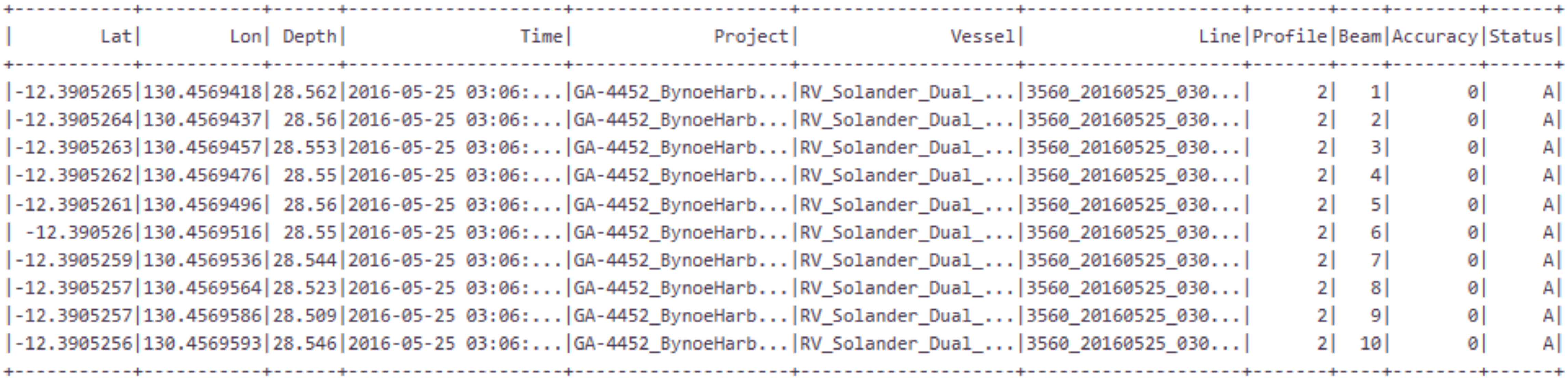
- Can be coded using Python, R, Java or Scala, and can also run SQL (database) commands



Bathymetry processing with Spark

http://bit.ly/2wUwuC0

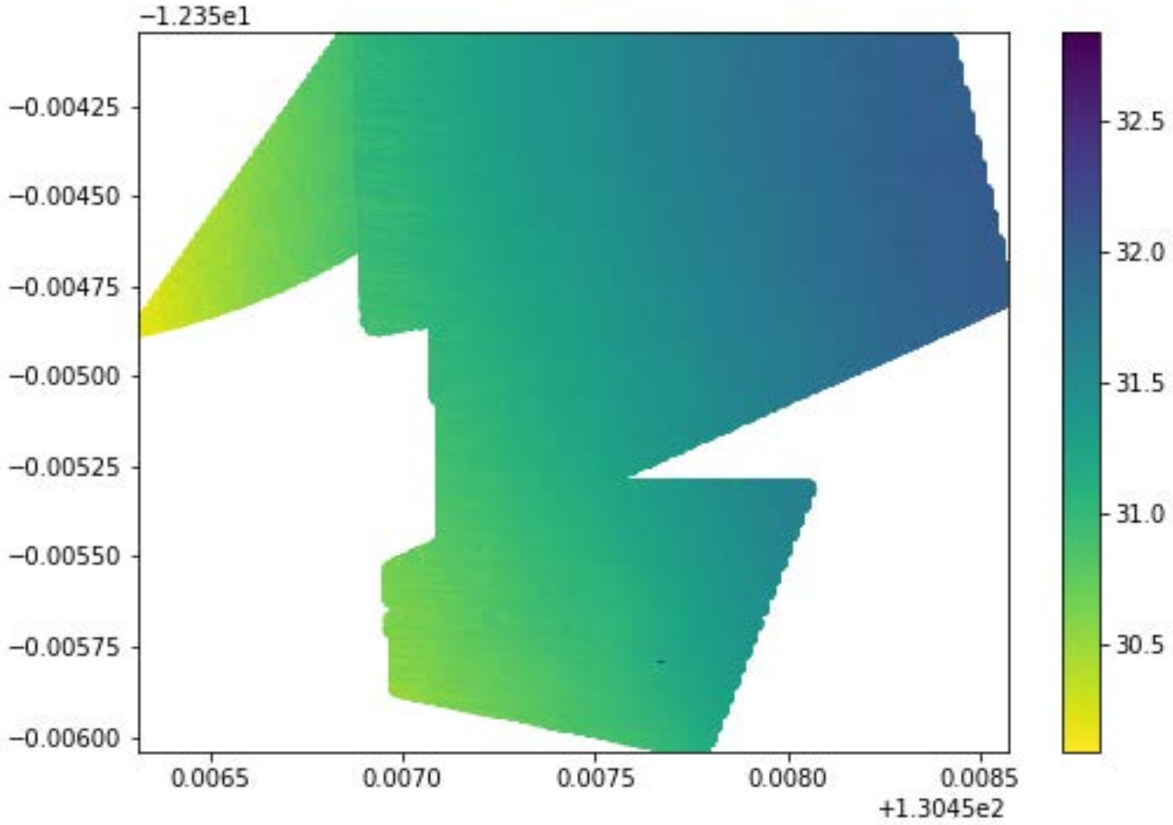
```
val s3 = spark.read.format("csv").load("s3a://test-bathymetry/*")
```



```
+-----------+-----------+------+--------------------+--------------------+--------------------+--------------------+-------+----+--------+------+
|        Lat|        Lon| Depth|                Time|             Project|              Vessel|                Line|Profile|Beam|Accuracy|Status|
+-----------+-----------+------+--------------------+--------------------+--------------------+--------------------+-------+----+--------+------+
|-12.3905265|130.4569418|28.562|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   1|       0|     A|
|-12.3905264|130.4569437| 28.56|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   2|       0|     A|
|-12.3905263|130.4569457|28.553|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   3|       0|     A|
|-12.3905262|130.4569476| 28.55|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   4|       0|     A|
|-12.3905261|130.4569496| 28.56|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   5|       0|     A|
| -12.390526|130.4569516| 28.55|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   6|       0|     A|
|-12.3905259|130.4569536|28.544|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   7|       0|     A|
|-12.3905257|130.4569564|28.523|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   8|       0|     A|
|-12.3905257|130.4569586|28.509|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|   9|       0|     A|
|-12.3905256|130.4569593|28.546|2016-05-25 03:06:...|GA-4452_BynoeHarb...|RV_Solander_Dual_...|3560_20160525_030...|      2|  10|       0|     A|
+-----------+-----------+------+--------------------+--------------------+--------------------+--------------------+-------+----+--------+------+
```

Bathymetry processing with Spark

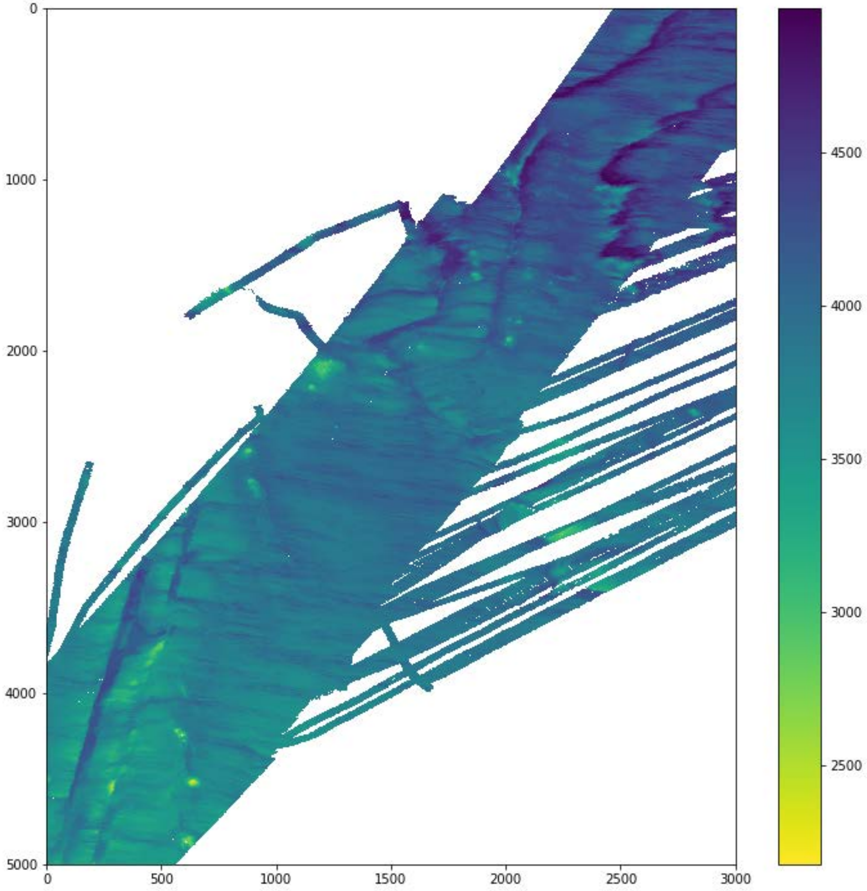
http://bit.ly/2wUwuC0



ESRI GeoTools for Hadoop
(and Spark!)

Bathymetry processing with Spark

http://bit.ly/2wUwuC0



Approximately *45 minutes* for >4.6 billion (cleaned) points (at 150m) using 8 m3.xlarge nodes, approximately AUD$0.48 Using AWS.

(previously > 8 hours)

Moving further ahead

geomesa

GeoTrellis

GeoWave

If you have questions:

Johnathan Kool (AAD)
johnathan.kool@aad.gov.au

Kim Picard (GA)
kim.picard@ga.gov.au