Big Data Analytics WG: Use Case Array Databases

RDA 4th Plenary

2014-sep-22, Amsterdam, The Netherlands

Peter Baumann

Jacobs University | rasdaman GmbH

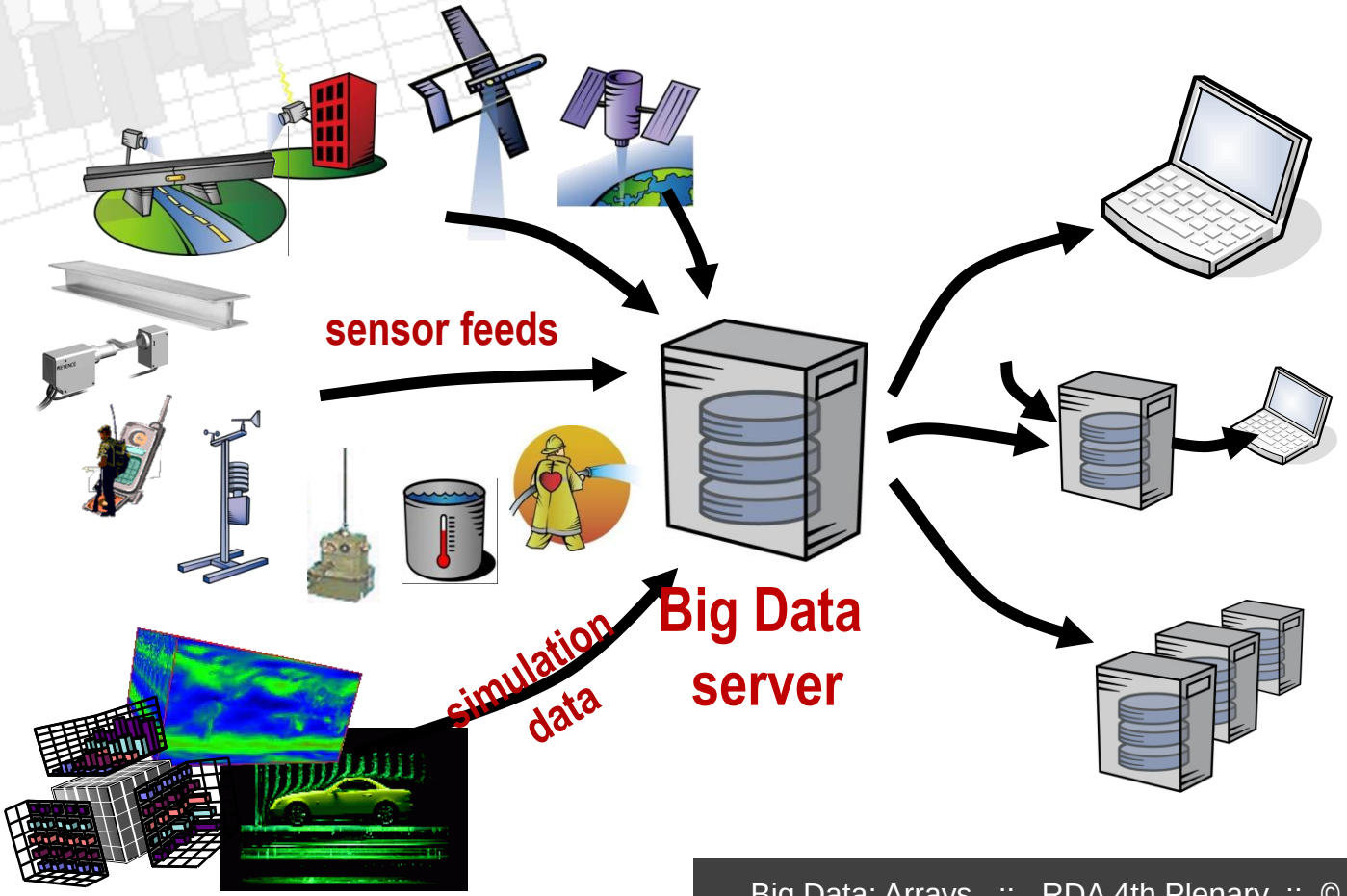
baumann@rasdaman.com

Structural Variety in Big Data

- Stock trading: 1-D sequences (i.e., arrays)
- Social networks: large, homogeneous graphs
- Ontologies: small, heterogeneous graphs
- Climate modelling: 4D/5D arrays
- Satellite imagery: 2D/3D arrays (+irregularity)
- Genome: long string arrays
- Particle physics: sets of events
- Bio taxonomies: hierarchies (such as XML)
- Documents: key/value stores = sets of unique identifiers + whatever
- etc.

Arrays in [Geo] Science & Engineering

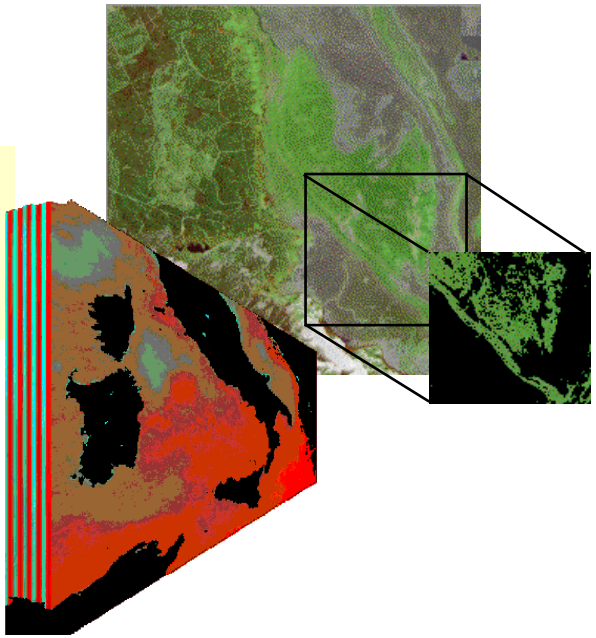
- spatio-temporal sensor, image, simulation, statistics data(cubes)



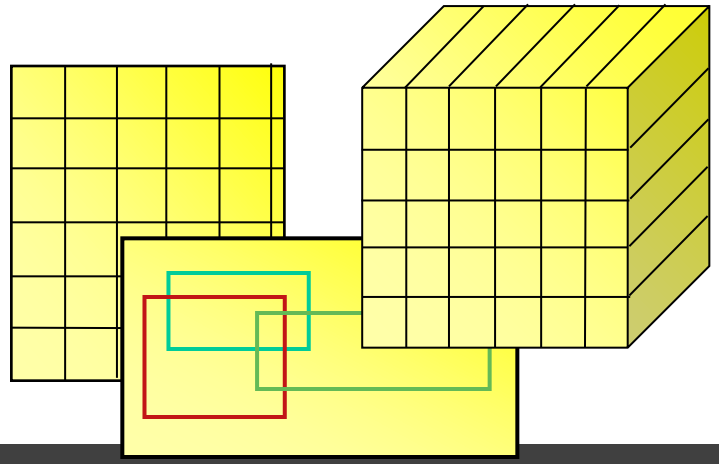
rasdaman: Agile Array Analytics

- „raster data manager“: SQL + n-D raster objects

```
select img.green[x0:x1,y0:y1] > 130
from   LandsatArchive as img
where  avg_cells( img.nir ) < 17
```

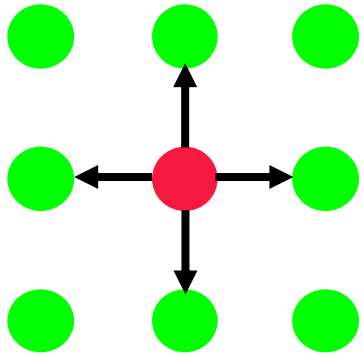


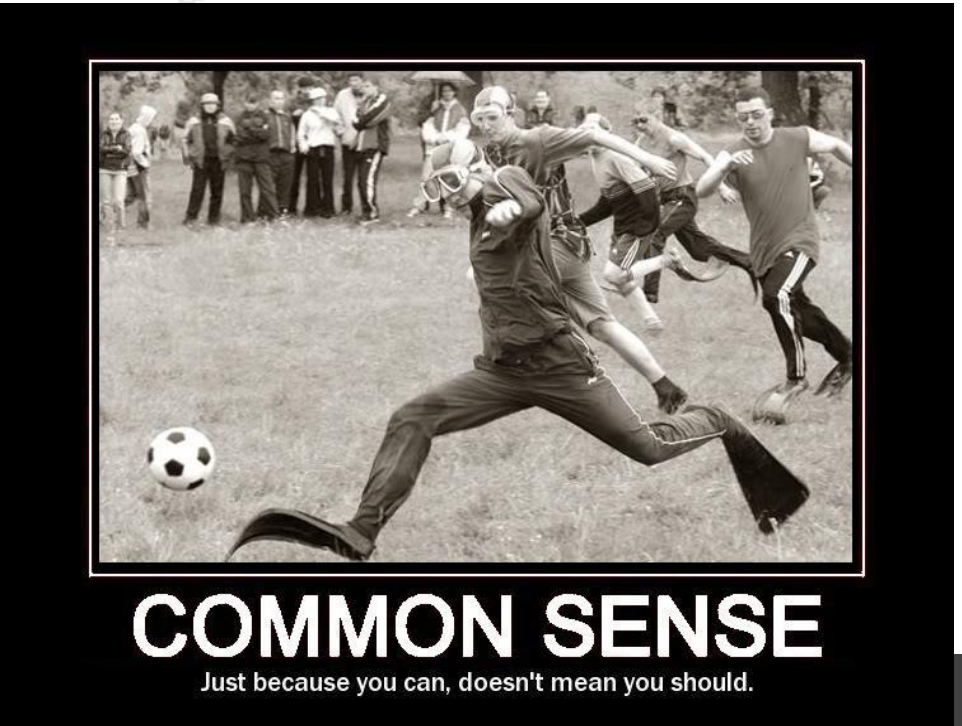
- Scalable parallel “tile streaming” architecture
- In operational use
  - OGC Web Coverage Service Core Reference Implementation

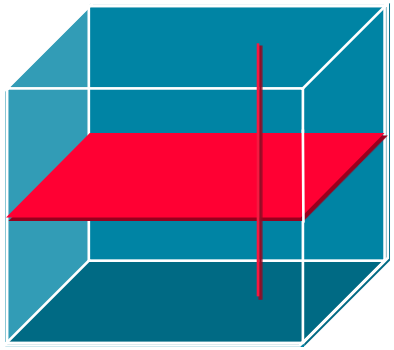


Inset: Hadoop is not the Answer to All

- no builtin knowledge about structured data types
  - "Since it was not originally designed to leverage the structure […] its performance […] is therefore suboptimal" [Daniel Abadi]
  - M. Stonebraker (XLDB 2012): „will hit a scalability wall"

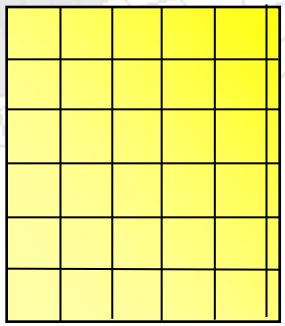




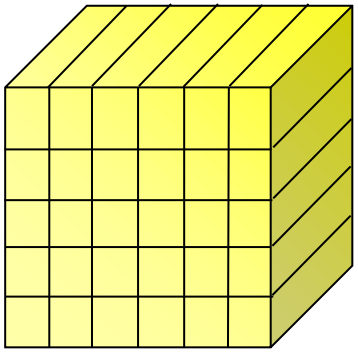


Adaptive Tiling

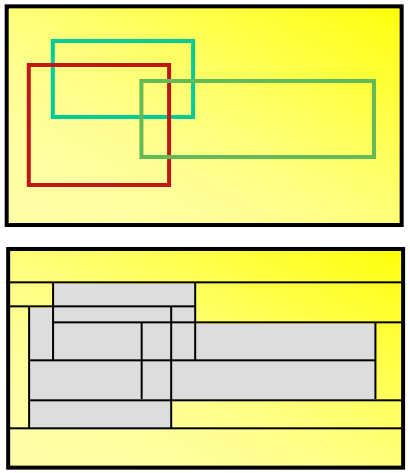
- Sample tiling strategies [Furtado]:
  - regular



directional



area of interest

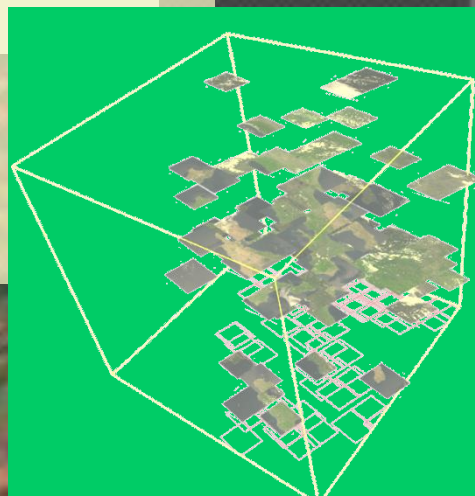


- rasdaman storage layout language

```
insert into MyCollection
  values ...
  tiling area of interest [0:20,0:40], [45:80,80:85]
  tile size 1000000
  index d_index storage array compression zlib
```

Sample Application: Database Visualization

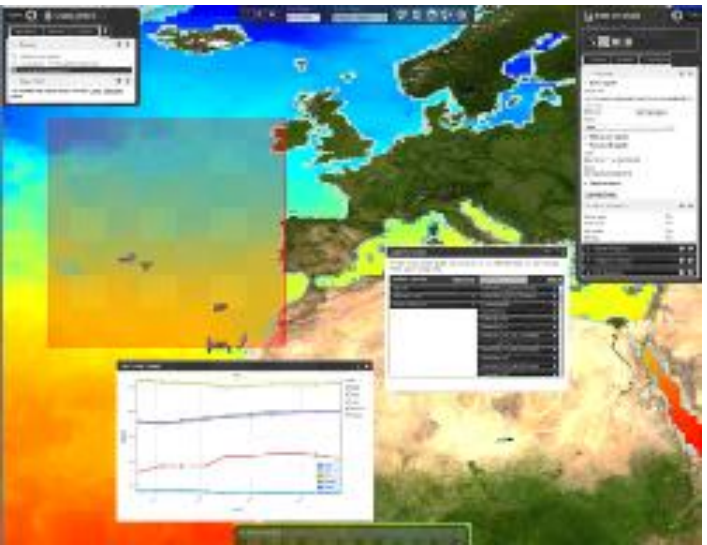
```
select
  encode(
    struct {
      red:   (char) s.image.b7[x0:x1,x0:x1],
      green: (char) s.image.b5[x0:x1,x0:x1],
      blue:  (char) s.image.b0[x0:x1,x0:x1],
      alpha: (char) scale( d.elev, 20 )
    },
    "image/png"
  )
from SatImage as s, DEM as d
```

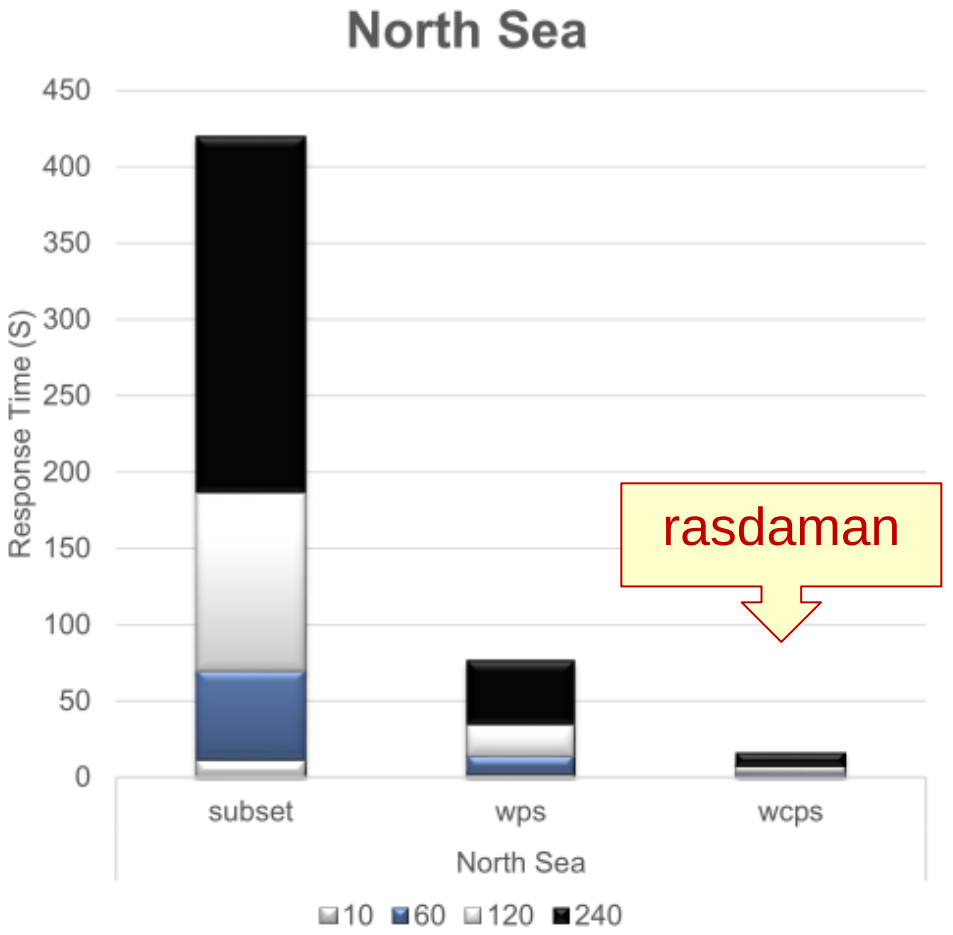


[Jacobs U, Fraunhofer: data courtesy

Use Case: Plymouth Marine Laboratory

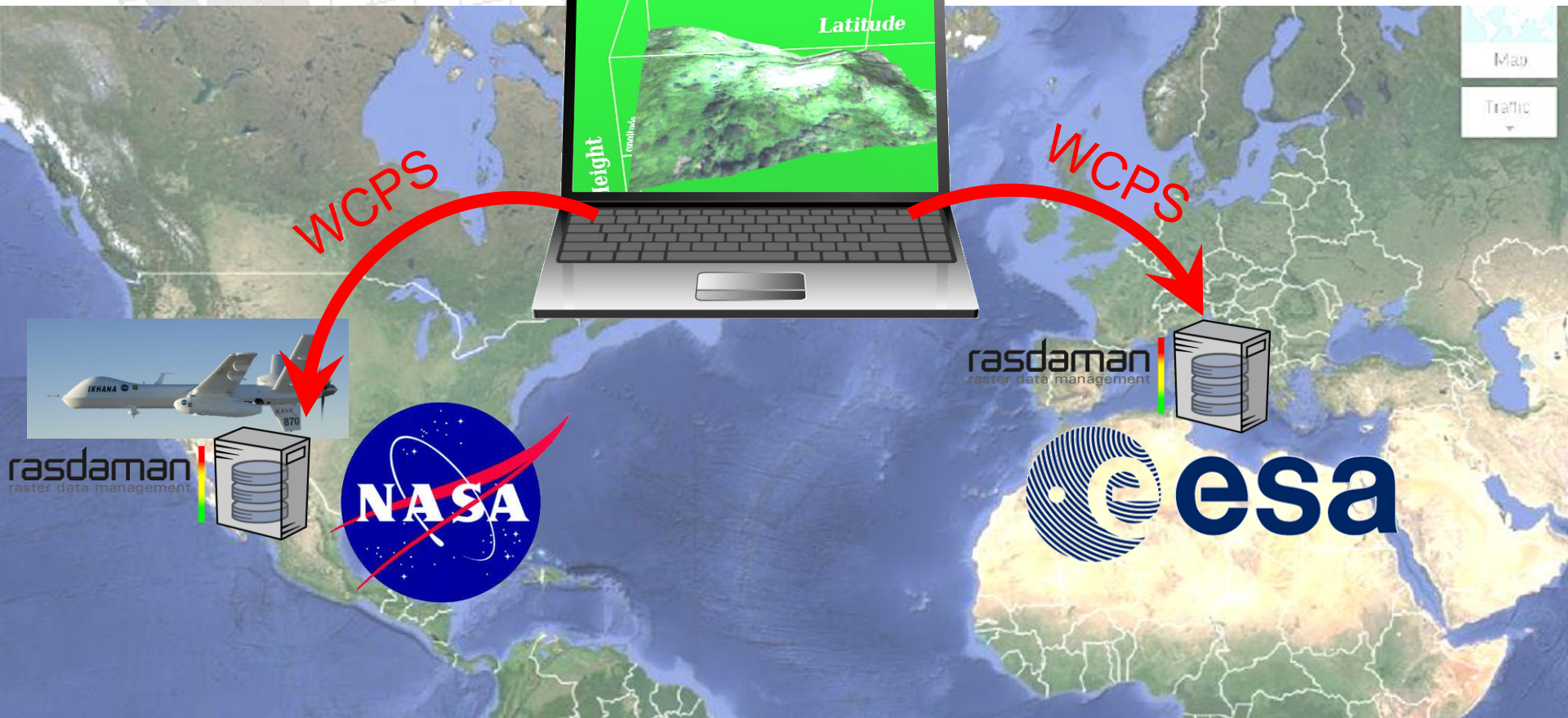
- "Avg chlorophyll concentration for given area & time period, from x/y/t cube"
  - 10, 60,120, 240 days
- Conclusions:
  - „we must minimise data transfer as well as [client] processing"
  - "standards such as WCPS provide the greatest benefit"





Secured Archive Integration

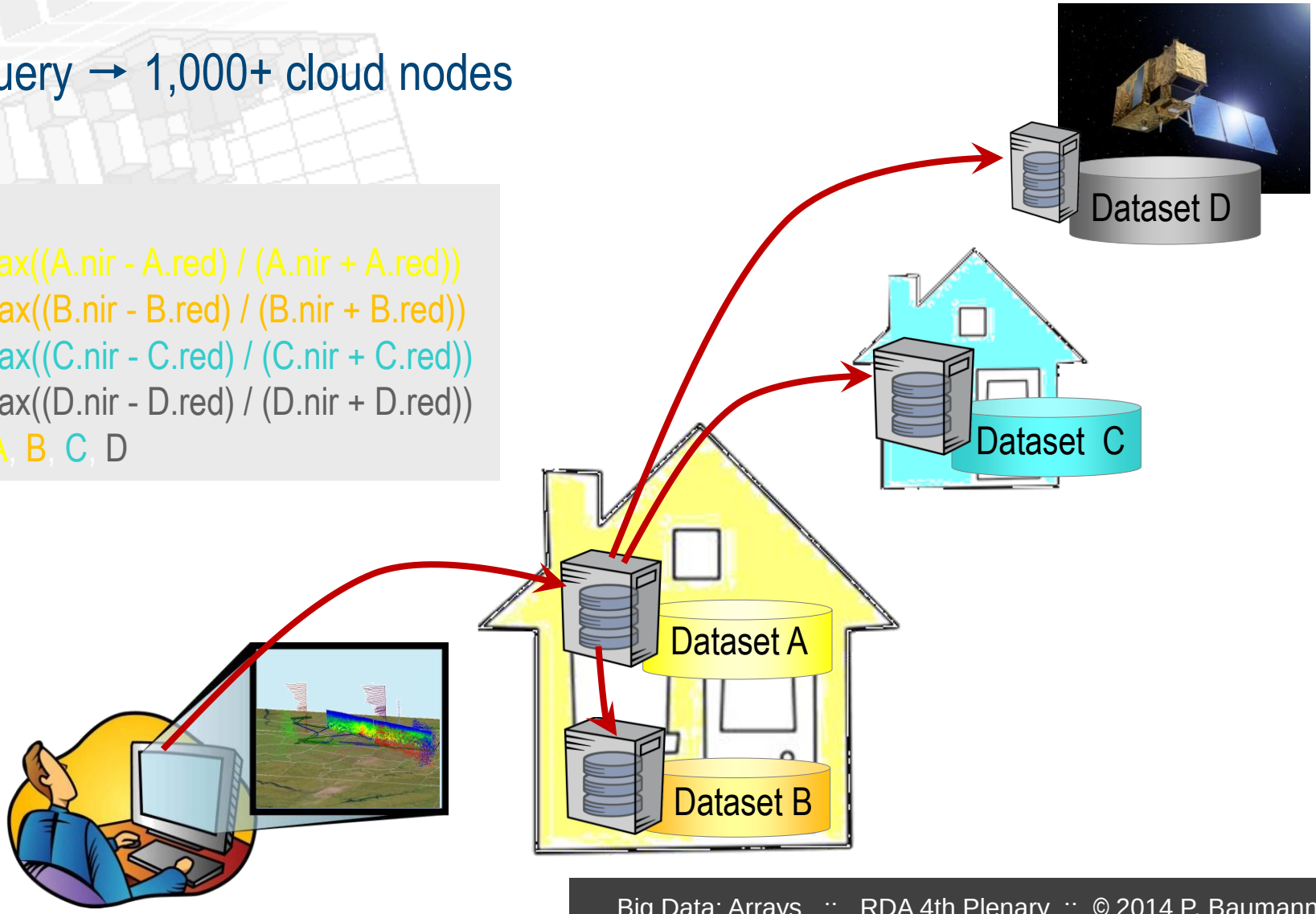
First-ever direct, ad-hoc mix from protected NASA & ESA services in OGC WCS/WCPS Web client (EarthServer + CobWeb)



Parallel / Distributed Query Processing

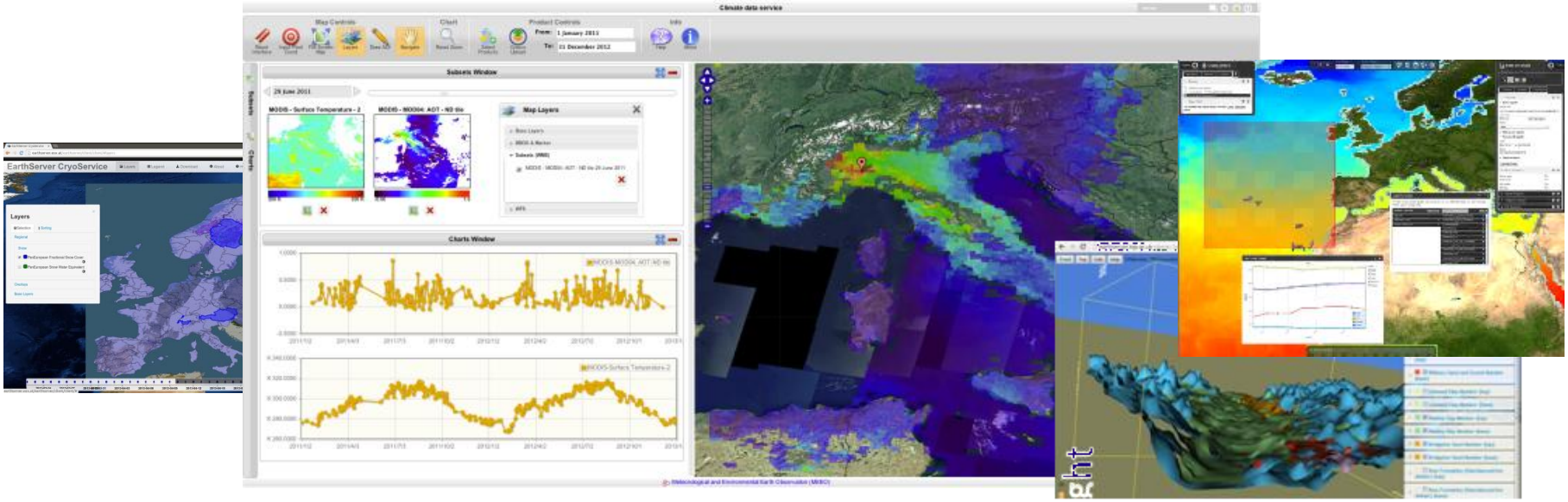
- 1 query → 1,000+ cloud nodes

```
select
      max((A.nir - A.red) / (A.nir + A.red))
    - max((B.nir - B.red) / (B.nir + B.red))
    - max((C.nir - C.red) / (C.nir + C.red))
    - max((D.nir - D.red) / (D.nir + D.red))
from A, B, C, D
```



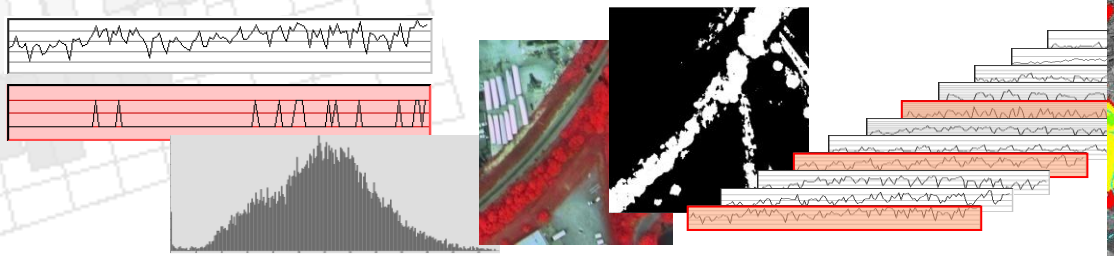
Array Databases: Practice Proven with rasdaman

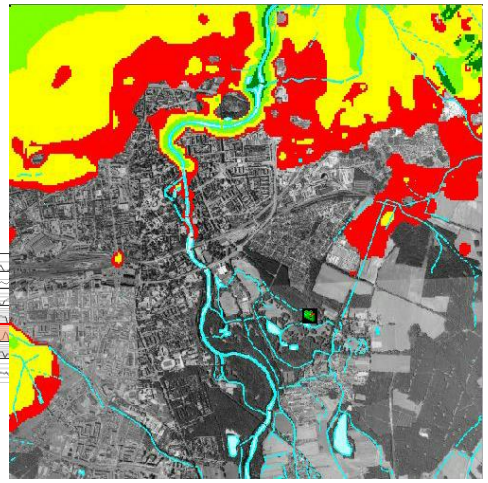
- from simple data access to agile analytics
  - strictly based on open OGC Big Geo Data standards
- 130+ TB databases, 2D, 3D x/y/z & x/y/t, 4D x/y/z/t timeseries
- single query distributed to 1,000+ cloud nodes



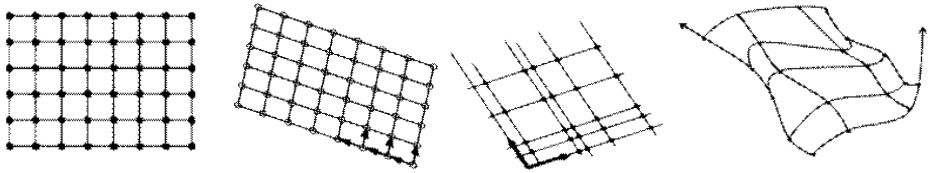
OGC WCPS

- OGC Web Coverage Processing Service (WCPS)
  = high-level geo raster query language; adopted 2008





  - WCPS 2: all grid types:



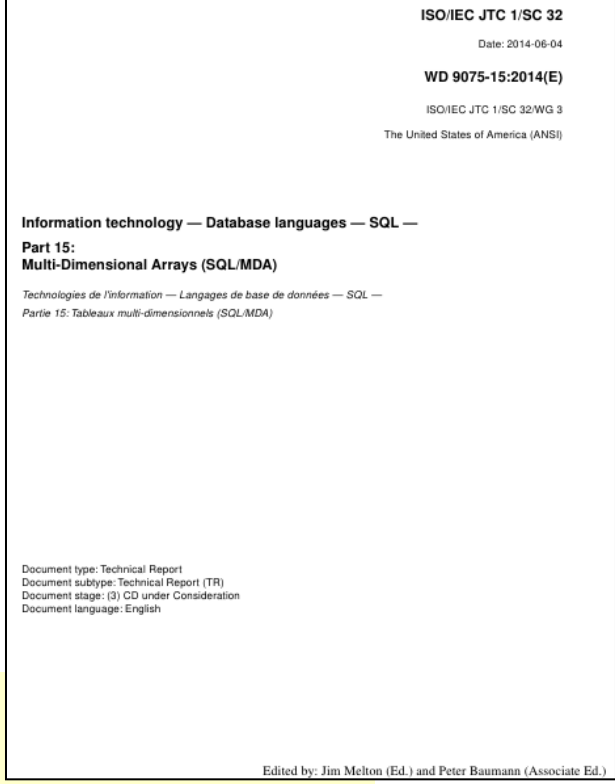
- "From MODIS scenes M1, M2, M3: difference between red & nir, as TIFF"
  - …but only those where nir exceeds 127 somewhere

```
for $c in ( M1, M2, M3 )
where some( $c.nir > 127 )
return encode( $c.red - $c.nir, "image/tiff" )
```

$(tiff_A, tiff_C)$

Recent Progress: ISO Array SQL

- ISO 9075 Part 15: SQL/MDA
  - resolved by ISO SQL WG in June 2014


ISO/IEC JTC 1/SC 32

Date: 2014-06-04

WD 9075-15:2014(E)

ISO/IEC JTC 1/SC 32/WG 3

The United States of America (ANSI)

Information technology — Database languages — SQL —

Part 15:
Multi-Dimensional Arrays (SQL/MDA)

Technologies de l'information — Langages de base de données — SQL —
Partie 15: Tableaux multi-dimensionnels (SQL/MDA)

Document type: Technical Report
Document subtype: Technical Report (TR)
Document stage: (3) CD under Consideration
Document language: English

Edited by: Jim Melton (Ed.) and Peter Baumann (Associate Ed.)

- n-D arrays as attributes

```
create table LandsatScenes(
    id: integer not null, acquired: date,
    scene: row( band1: integer, ..., band7: integer ) array [ 0:4999,0:4999] )
```

- declarative array operations

```
select  id, encode(scene.band1-scene.band2)/(scene.nband1+scene.band2)), „image/tiff“ )
from    LandsatScenes
where acquired between „1990-06-01“ and „1990-06-30“ and
          avg( scene.band3-scene.band4)/(scene.band3+scene.band4)) > 0
```