rasdaman: Big Data Analytics auf multidimensionalen Rasterdaten

Open-Source Park, INTERGEO 2013

Peter Baumann
Jacobs University | rasdaman GmbH
Array DB Research @ Jacobs U

- Large-Scale Scientific Information Systems research group
  - focus: large-scale n-D raster services & beyond
  - www.jacobs-university.de/lis
- Spin-off company: rasdaman GmbH
- Main results:
  - Array DBMS, rasdaman
  - Geo service standards: Chair, OGC raster-relevant working groups, editor of 10+ stds & candidate stds
    - Geo Array QL standard (adopted)
  - Further: Array SQL
- **raster data manager**
  = Array DBMS for massive n-D raster data
- SQL + imaging operators

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

- **Flexibility** ← query language
- **Scalability** ← „tile streaming“ architecture, parallelization
- In operational use

[www.rasdaman.org](http://www.rasdaman.org)
Array Query Operators: rasql

- **selection & section**

```sql
select c[ *:* , 100:200, *:* , 42 ]
from ClimateSimulations as c
```

- **result processing**

```sql
select img * (img.green > 130)
from LandsatArchive as img
```

- **search & aggregation**

```sql
select mri
from MRI as mri, masks as am
where some_cells( mri > 250 and m )
```

- **data format conversion**

```sql
select png( c[ *:* , *:* , 100, 42 ] )
from ClimateSimulations as c
```
Configurable Tiling

- Sample tiling strategies [Furtado]:
  - regular
  - directional
  - area of interest

- rasdaman storage layout language

```sql
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
```
Distributed Query Processing

select max((A.nir - A.red) / (A.nir + A.red))
from A

select max((B.nir - B.red) / (B.nir + B.red))
from B

select max((A.nir - A.red) / (A.nir + A.red))
- max((B.nir - B.red) / (B.nir + B.red))
from A, B
Calling Tools from Database Queries

- UDF = invocation of external code within query
  - Transparently integrated with tile streaming, optimization, parallelization
- Ex: “NDVI from raw Landsat subset, orthorectified with Orfeo Toolbox“

```sql
select
  encode(
    otb.orthoRectifFilter(
      ((img.red-img.nir)/(img.red+img.nir)) [x0:x1,y0:y1],
      outputSpacing, deformationFieldSpacing
    ),
    "png"
  )
from   LandsatRawArchive as img
```
In-Situ Databases

- Traditionally: data imported into database
  - full data control → efficient data organization
- Problem: Large-scale data centers sometimes object to copying
  - Data simultaneously used by other actors (big NO-NO in classical databases!)
- Approach: reference external files, use as tiles; cf [Ailamaki et al 2010]

```sql
insert into MyCollection
  referencing /my/path/*.tif

update MyCollection
  referencing /oops/forgot/*.jpg
```

- Different from storing tiles in files!
- Challenges: efficiency, consistency, caching, ...
3D Database Visualization

- rasdaman Web client toolkit, and...

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

[data courtesy BGS, ESA]
A Brief History of Array DBMSs

first appearance in literature (not first implementation)
EarthServer: Big Earth Data Analytics

- Scalable On-Demand Processing for the Earth Sciences
  - EU FP7-INFRA, 3 years, 5.85 mEUR
- 100+ TB databases for all Earth sciences + planetary science
  - Platform: rasdaman
Ex: Climate Data Service, MEEO
Ex: Plymouth Marine Laboratory
Ex: British Geological Service
Conclusion

- Multi-dimensional Arrays are „Big Data“
  - earth / space / life sciences, business, ...
- rasdaman: Flexibility, scalability, information integration, and more
  - www.rasdaman.org, standards.rasdaman.org