Agile Analytics on Big Data

Geospatial World Forum
Rotterdam, The Netherlands, 2013-may-16

Peter Baumann & the rasdaman Team

Jacobs University | rasdaman GmbH
Bremen, Germany
p.baumann@jacobs-university.de

Supported by EU FP7 eInfrastructure, contract 283610 EarthServer
Collecting Spatio-Temporal Data

sensor feeds
[OGC SWE]

Big Data server

simulation data
Serving Spatio-Temporal Data

sensor feeds
[OGC SWE]

simulation data

Big Data server
A Standards-Based Approach

- Core element in OGC: geographic **feature**
  - = abstraction of some real-world phenomenon

- Special kind of feature: **coverage**
  - = space-time varying multi-dimensional phenomenon
  - raster images...**but there is more!**

- Typically, coverages are **Big Data**
Ex: Spatio-Temporal Image Services

[Diedrich et al, DLR-DFD]

[CC BY]
The rasdaman Raster Analytics Server

- **Array DBMS** for massive n-D raster data
- Extending SQL with array operators:
  
  ```sql
  select img.green[x0:x1,y0:y1] > 130
  from   LandsatArchive as img
  ```

- “tile streaming” architecture
  - extensive optimization, hw/sw parallelization
  - Scaling from laptop to cloud

- open source from www.rasdaman.org
What Can You Do with Raster Queries?

- **ad-hoc navigation, extraction, aggregation, analytics**
- **Time series**
- **Image processing**
- **Summary data**
- **Sensor fusion & pattern mining**

- current value is 8220.0;
- average over all values up to now currently is 7461.7692307692305.
Scaling Big Data

- Data scalability
  - Compression +++ adaptive partitioning +++ transparent tape integration +++ raw file data integration

- Processing scalability
  - "tile streaming“ architecture +++ effective query optimization +++ cluster / cloud parallel query processing +++ external code invocation

- User spectrum scalability
  - Ad-hoc queries +++ graphical query frontends
Scalability: Parallel Query Processing

- Heterogeneous federation / cloud
  - Can optimize for data location, transport volume, node load, ...

\[
\text{select } \max\left(\frac{A.\text{nir} - A.\text{red}}{A.\text{nir} + A.\text{red}}\right) - \max\left(\frac{B.\text{nir} - B.\text{red}}{B.\text{nir} + B.\text{red}}\right) \\
\text{from } A, B
\]

\[
\text{select } \max\left(\frac{B.\text{nir} - B.\text{red}}{B.\text{nir} + B.\text{red}}\right) \\
\text{from } B
\]
3D Database Visualization

- Problem: coupling DB / visualization

- Approach:
  - via X3D:
    - deliver RGBA directly into client GPU

```sql
select
  encode(
    { 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
```

[JacobsU, Fraunhofer 2012]
EarthServer: **Big Earth Data Analytics**

- Scalable On-Demand Processing for the Earth Sciences
  - EU funded, 3 years, 5.85 mEUR
  - Platform: rasdaman (Array Analytics server)
  - Distributed query processing, integrated data/metadata search, 3D clients

- Strictly open standards: OGC WMS+WCS+WCPS; W3C Xquery; X3D

- 6 * 100+ TB databases for all Earth sciences + planetary science
Rasdaman Impact

- Has established the research field of Array Databases
- Registered GEOSS component for Big Data Analytics
- OGC WCS Core, OGC WCPS reference implementation
- Experience heavily drives OGC coverage & ISO Array SQL standardization
Conclusion

- Sensor, image, simulation, & statistics data = a main source of Big Data in Earth Sciences
  - Petrol industry has “more bytes than barrels”

- OGC W*S standards offer coverage interoperability
  - www.ogcnetwork.net/wcs

- Core paradigm: high-level integrated query language as c/s interface
  - data + metadata search in same query
  - Visual clients hide QL
  - optimizations, parallelization on rasdaman server

- **EarthServer** introduces Agile Analytics
  - www.earthserver.eu