SCALABLE EARTH OBSERVATION ANALYTICS WITH SCIDB

Marius Appel
marius.appel@uni-muenster.de
EO DATA ORGANIZATION

LANDSAT 8
EO DATA ORGANIZATION

SENTINEL 2
EO DATA ORGANIZATION

<table>
<thead>
<tr>
<th>Name</th>
<th>Änderung...</th>
<th>Typ</th>
<th>Größe</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>3.663 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>119.141 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>123.402 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>129.554 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.858 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.773 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.792 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.629 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>129.694 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.999 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>3.682 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>2.281 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.905 KB</td>
</tr>
<tr>
<td>S2A_OPER_PRD_MSL1C_PDMAC_20160514T170317_R121_V20150810T084516_201508</td>
<td>14.05.2016</td>
<td>JP2-Datei</td>
<td>32.927 KB</td>
</tr>
</tbody>
</table>

**SENTINEL 2**
EO DATA ORGANIZATION

SENTINEL 2
EO DATA ORGANIZATION

SENTINEL 2
EO DATA ORGANIZATION

SENTINEL 2
EO DATA ORGANIZATION

• EO image deployment is file-based
• GDAL interfaces EO imagery with GIS software
• Difficult to analyze large image collections due to
  – data volume
  – Irregularities
  – lack of time support in GDAL
• Higher-level data organization as an alternative to files?
  – Key requirement: scalability
SCIDB INTRODUCTION

• Array-based data management and analytical system [1]
• Relies on shared nothing architectures
• Open-source version available, extensible by UDFs
• Basic data representation as multidimensional arrays:
  – $n$ dimensions, $m$ attributes with different data types

SCIDB ARCHITECTURE

• arrays are divided into equally sized chunks

• chunks are distributed over many SciDB instances

• Size and shape of chunks are defined by users per array and have strong effects on computation times

• Storage is nearly sparse
• SciDB query language: Array Functional Language (AFL)

• Built in functionality:
  – Load / write arrays from / to files
  – Arithmetic operations
  – subsetting by dimensions, attributes, or values
  – Aggregations
  – Joins
  – Changing array schemas (repartitioning, redimensioning)
  – Linear algebra routines: (GEMM, GESVD, basic statistics)
  – ...
EXTENSIONS FOR EO DATA

• scidb4geo (https://github.com/appelmar/scidb4geo)
  – SciDB plugin adds metadata and simple operations on space-time referenced arrays

• scidb4gdal (https://github.com/appelmar/scidb4gdal)
  – ingest / download to / from GDAL supported files
  – spacetime mosaicing

• R package scidbst (https://github.com/flahn/scidbst)
  – mimics functionality of common packages on SciDB arrays
SCIDB CLIENTS

• Low-level clients: iquery, Shim

• High-level R client (similar for Python)
  – overrides standard methods, e.g. %*%
  – make extensive use of proxy objects
  – lazy evaluation:
    • compute things when result is being read
    • ignore computations for unread parts of the results
SCIDB STREAMING

- Run external programs (e.g., R, python) within SciDB at chunk level parallelism

→ chunk size selection must be adapted to the analysis
STUDY CASE: LAND USE CHANGE MONITORING IN SOUTH WEST ETHIOPIA FROM LANDSAT 7 IMAGERY

- Landsat 7 data from 12 tiles captured between 2003-07-21 and 2014-12-27 → 1975 scenes
- approx. 325,000 km²
- monitor changes starting with 2010-01-01
- using R and Breaks For Additive Season and Trend and its R implementation [1]

EO DATA AS REGULAR ARRAYS
LANDSAT 7 IN SCIDB

Images form a single three-dimensional array with **daily temporal resolution** and

- 49548 x 47713 x 4177 cells in total
- Only 0.5% (54 \cdot 10^9) of the cells contain data \(\rightarrow\) sparse storage
STUDY CASE IMPLEMENTATION

1. Ingestion using GDAL

2. Preprocessing (with built-in SciDB functionality)
   - remove any values \(\leq -9999\) or \(>10000\)
   - compute NDVI vegetation index
   - Reorganize chunks such that one chunk stores complete time series of
     64 x 64 pixels

3. Run R scripts on all chunks using streaming

4. Postprocessing (with built-in SciDB functionality)
   - Reshape one-dimensional result array to form a two-dimensional map

5. Export results using GDAL
STUDY CASE: RESULTS

Year of detected changes
- 2010
- 2011
- 2012
- 2013
- 2014

Landsat Tiles
STUDY CASE SCALABILITY

• 16 SciDB instances

• running change analysis repeatedly with different number of available CPU cores
CONCLUSIONS

• The array model with chunking and sparse storage seems well-suited to represent large EO datasets from many scenes at a higher level than files.

• Analyses scale well with available hardware.

• Little reimplementation needed to scale complex time-series processing through streaming (and no need to care about parallelization / external memory).

• Installation and data ingestion not straightforward and time-consuming.

• Mostly useful for re-analysis but not real-time processing.

• Missing interactive(!) user interfaces (á la Google Earth Engine) to make the technology more accessible to end users?
THANK YOU

• Questions?

• Hands-on with SciDB tomorrow!

• Slides available at GitHub: https://github.com/appelmar/edcforum2017

• Contact marius.appel@uni-muenster.de