Pangeo -TileDB

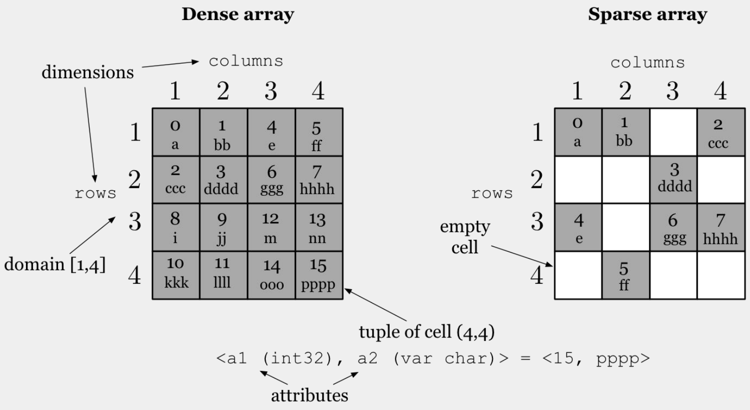
8/22/2019

[tile]DB

About TileDB

- TileDB introduces a novel format and a powerful storage library for storing and accessing massive dense and sparse multi-dimensional array data
- The TileDB storage engine is optimized for the cloud
- TileDB, Inc. was founded in 2017 to further develop and maintain the TileDB project, which was a research project at Intel Labs and MIT
- TileDB has raised $4 million to date in Seed funding by Nexus Venture Partners and Intel Capital

TileDB Features



Dense array

	1	2	3	4
1	0 a	1 bb	4 e	5 ff
2	2 ccc	3 dddd	6 ggg	7 hhhh
3	8 i	9 jj	12 m	13 nn
4	10 kkk	11 llll	14 ooo	15 pppp

Sparse array

	1	2	3	4
1	0 a	1 bb		2 ccc
2			3 dddd	
3	4 e		6 ggg	7 hhhh
4		5 ff		

- Native support for sparse and dense arrays
- Rapid updates (and appends on dimensions)
- Multiple attributes for a single cell
- Parallel reads/writes for network storage, e.g. S3
- Integration with Python, Dask, R, Java, Spark, Presto, Go, PDAL and GDAL
- Well documented
- MIT license

[tile]DB

TileDB for Geospatial Data Processing

- Can model complex raster data, with multiple attributes / pixel
- GDAL integration makes common geospatial ETL operations (re-projection, registration, translation, etc.) easy and efficient.
    - Full support for complex data types (I / Q) in the same array
- Supports efficient updates / appends / labeling.
- Full support for multi-variable datasets, e.g. HDF 4/5 , netCDF and NITF
- Attribute labelling of values in the array
- Time series in the same array (appending datasets)

[tile]DB

Dask/TileDB

- GDAL code modified to support parallel writes to an existing TileDB array
- Dask
  - Parallel computations that scale up to thousands of nodes
  - https://docs.dask.org/en/latest/

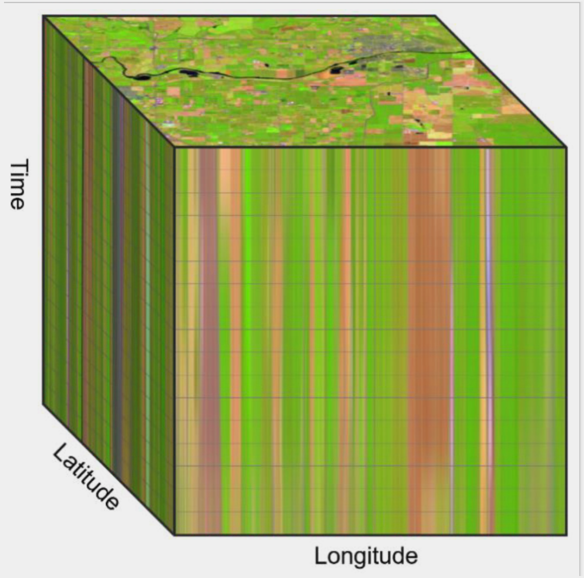
TileDB enables fast parallel reads/writes with Dask

TileDB and the GDAL driver support multiple attributes per pixel value e.g. latitude, longitude and complex value

Applicability to ARDs

Analysis Ready Datasets

- TileDB Differentiators
  - Multi-dimensional support
    - Multiple attributes per pixel
    - Dense arrays
    - Sparse arrays
  - Append
  - Update
  - Parallel processing/tooling around these areas
  - No additional metadata indexing required



Achieving the Full Vision of Earth Observation Data Cubes
https://www.mdpi.com/2306-5729/4/3/94/htm

COGs / TileDB

- COGs can be a data management headache
- COGs are limited by the options available in the TIFF spec
    - COGDumper
    - Libtiff
    - GDAL

TileDB provides an open source SDK in multiple languages, Python, Java, R, Spark in addition to a novel format that matches the needs of an ARD

TileDB / ARDs

Current state of play

- GDAL
  - Multi-dimensional support and translation from HDF / netCDF
  - Parallel appends / updates
  - Support for complex data types
- PDAL
  - Sparse array support with parallel appends / updates suitable near real time collects