Array Data Management and Querying in the Cloud

Hannes Dröse

German Aerospace Center (DLR), Institute of Data Science Jena
hannes.droese@dlr.de

Motivation

- scientific data often stored in dense multi-dimensional arrays
- data typically resides in cloud storage due to PB-scale
- relational model not ideal
  - ➢ inefficient
  - ➢ inconvenient

State-of-the-Art

GeoSpatial Extensions to DBMSs

- e.g. PostGIS, Oracle Spatial
- well suited for vector data
- problematic for >2 dimensions
- require time-consuming ingestion

Array DBMSs

- e.g. rasdaman, SciDB
- only a subset is still active
- no standard data model or query language
- most require ingestion

=> often not used by scientists

Current Approaches

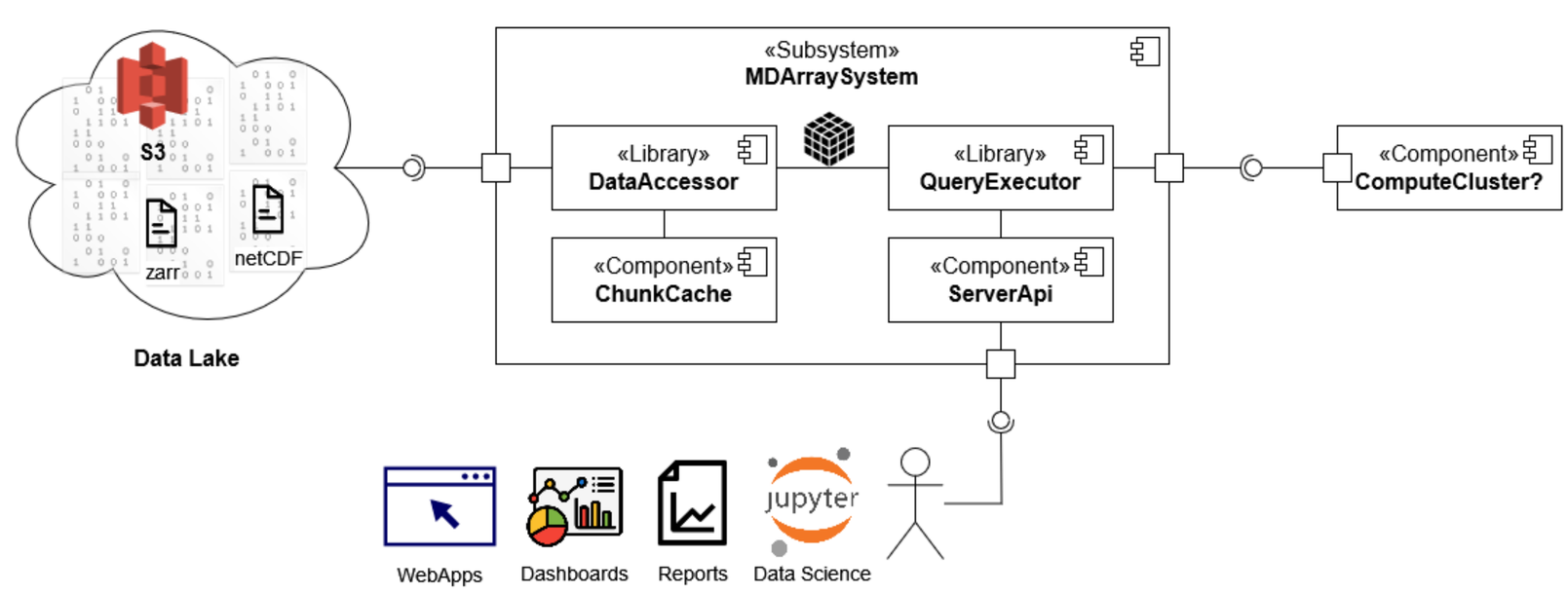
- data stored as files/objects in cloud services
- „cloud-ready file formats“ e.g. zarr
- usage of Data Science libraries for tensor data (TensorStore, Xarray, Dask)
- imperative script execution

Our Approach

- multi-dimensional array data lakehouse

Current State:

- initial server implementation
- based on Python & Xarray for rapid prototyping
- access to netCDF and zarr datasets
- supported operations: slicing, dicing, aggregation



Architecture of Proposed System

Future Plans

- implement data access library with chunk-based caching
- develop actual query language
- query optimization and execution

Research Topics

- query language and optimization
- caching strategies
- dynamic rechunking to match access patterns
- file formats and their implications
- benchmarking

Current Interest

- no standard array benchmark
- previous work only domain-specific fixed-sized benchmarks or micro-benchmarks

Goal:

- state-of-the-art benchmark
- scalable, realistic, choke points

DLR