TECHNISCHE UNIVERSITÄT
ILMENAU

Definition of a Benchmark for Raster Data Processing

by

Proteeti Prova Rawshan

Supervisor: Dr. Marcus Paradies

Research in Computer & Systems Engineering
proteeti-prova.rawshan@tu-ilmenau.de

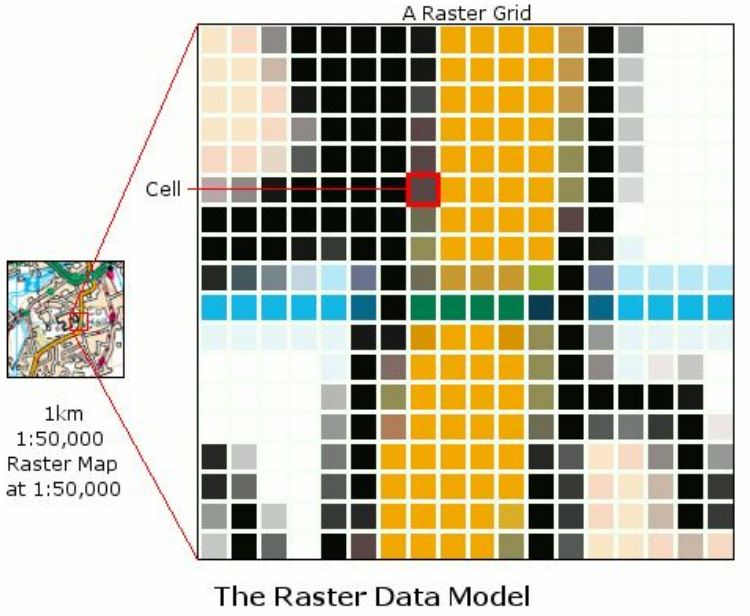
21 March, 2024
Research Project- WS 2023/24
Department of Databases & Information Systems

Outline

- Introduction
- Benchmarking Raster Data
- Methodology
- Implementation
- Evaluation
- Conclusion & Future Directions

Introduction: Raster Data

- Originates from satellite imagery for precise earth surface capture.
- Treats Earth as a continuous surface, storing data like a digital photograph.
- Uses a grid of cells (pixels), each holding a value
- Cells/ pixels represent specific ground areas in rows and columns.
- Values in cells can represent classes (e.g., land use, vegetation) or measurements (e.g., elevation).



The Raster Data Model

Fig. 2024. Spatial Data Models. https://gisinfo.hertfordshire.gov.uk/gisdata/vectorraster.htm. Accessed 2024-03-20.

Introduction: Raster Data Processing

Xarray

Rasterio

GDAL

DuckDB

Raster Data Processing Systems

Complex analysis, transformation & visualization

Diverse functional scopes & characteristics

Introduction: Raster Data Processing

Xarray

Rasterio

GDAL

DuckDB

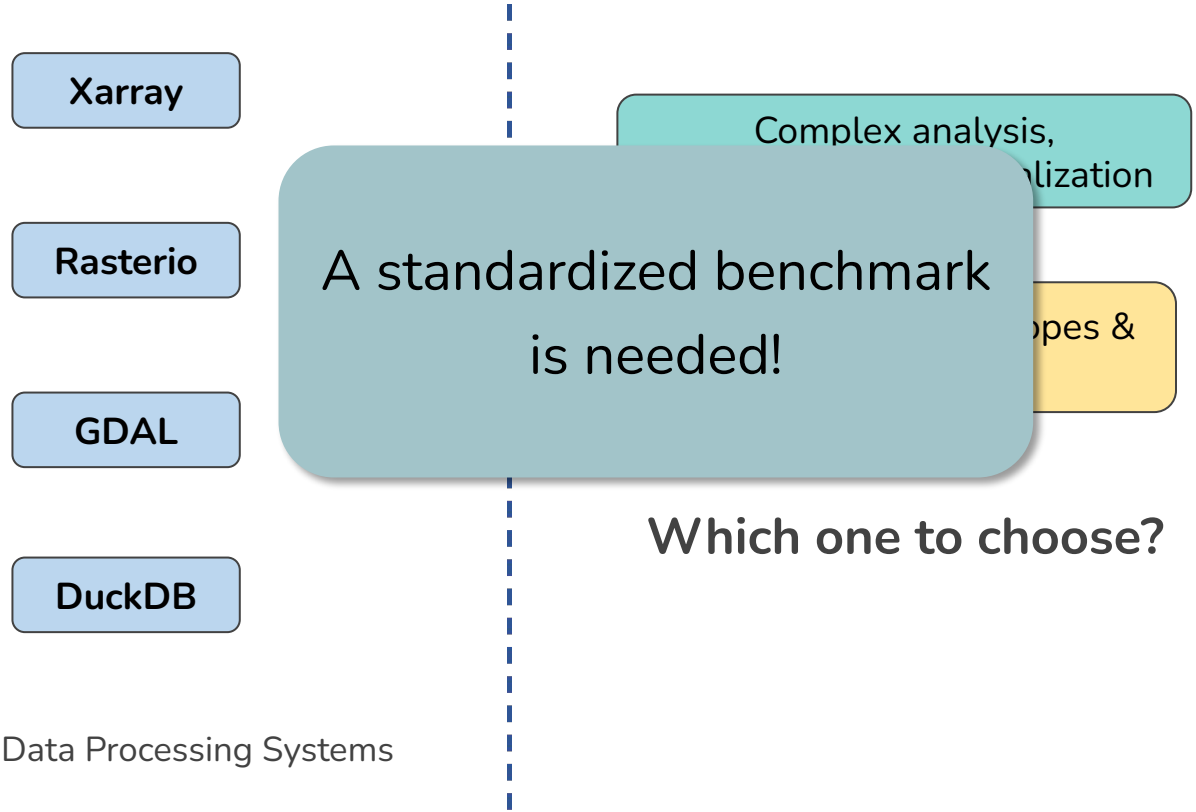
Raster Data Processing Systems

Complex analysis, transformation & visualization

Diverse functional scopes & characteristics

Which one to choose?

Introduction: Raster Data Processing



Raster Data Processing Systems

Benchmarking Raster Data Processing Systems

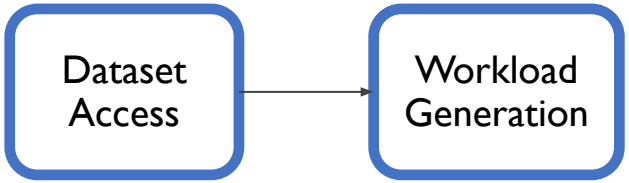
- Existing benchmarks focus on specific characteristics or optimization strategies
- They do not offer a comprehensive evaluation, especially in the **geospatial domain**
- Our solution provides an **end-to-end benchmarking framework**, performing complex aggregations - covering a wider range of use cases
- Fulfills the requirements of handling **high workloads & efficient processing** in read world settings
- Also evaluates **system performance** in large-scale raster data operations

Methodology

Dataset Access

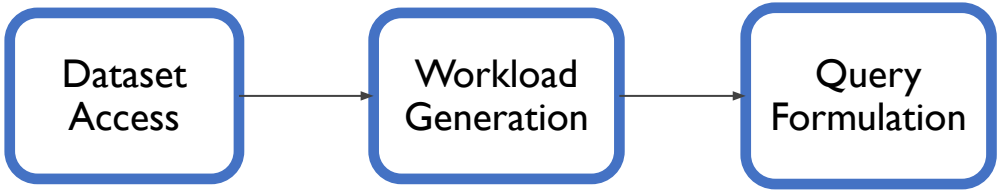
➔ ERA5-Land climate data

Methodology



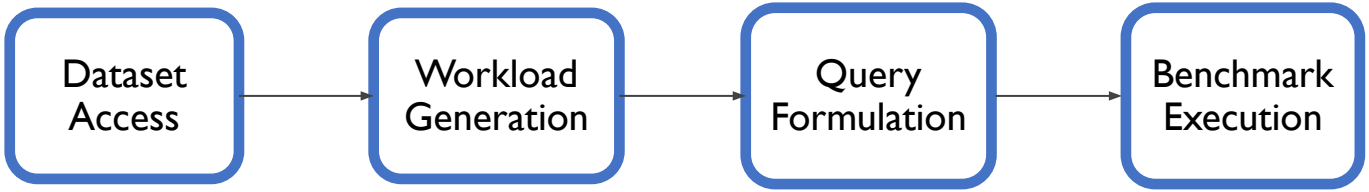
➔ mimics complex climate data analysis tasks

Methodology



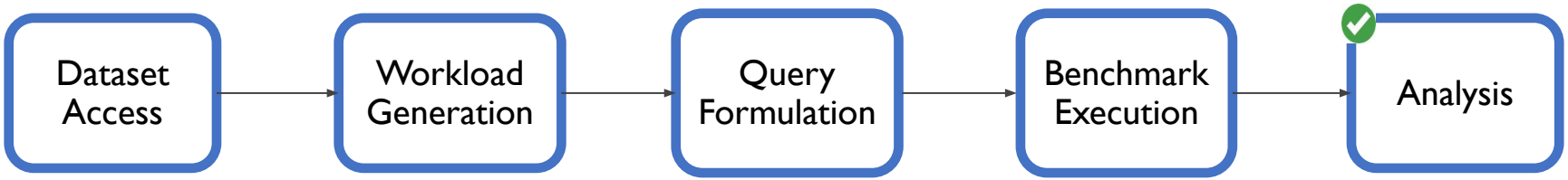
➔ basic to advanced aggregations, common analytics in climate research

Methodology



➔ measures performance under different computational demands research

Methodology



➔ measures performance under different computational demands research

Crafting the benchmark

- ERA5-Land t2m or "t2m" variable, which is the 2-meter temperature data across Europe (2000-2022).

- Example query: **Temporal Range Aggregation**
  - aggregate 't2m' temperature data over specific time frames
  - showcasing how the benchmark handles time-series data efficiently

Crafting the benchmark

Temporal Range Aggregation

Calculates the average or total temperature for a specific period.

query function definition

Benchmark Configuration

Temporal Range Aggregation: 30

query type: weight

Workload Generation

if type == Temporal Range Aggregation:

generate parameters

returns query params & aggregation type

Crafting the benchmark

Temporal Range Aggregation

Calculates the average or total temperature for a specific period.

query function definition

Benchmark Configuration

Temporal Range Aggregation: 30

query type: weight

Workload Generation

generates workload.csv

populated with query params

Crafting the benchmark

Temporal Range Aggregation

Calculates the average or total temperature for a specific period.

query function definition

Benchmark Configuration

Temporal Range Aggregation: 30

query type: weight

Workload Generation

generates workload.csv

populated with query params

Workload Execution

➔ takes time range & operation type (mean/sum) from workload, executes query

➔ logs before/after results of resources used

Crafting the benchmark

Temporal Range Aggregation

Calculates the average or total temperature for a specific period.

query function definition

Benchmark Configuration

Temporal Range Aggregation: 30

query type: weight

Workload Generation

generates workload.csv

populated with query params

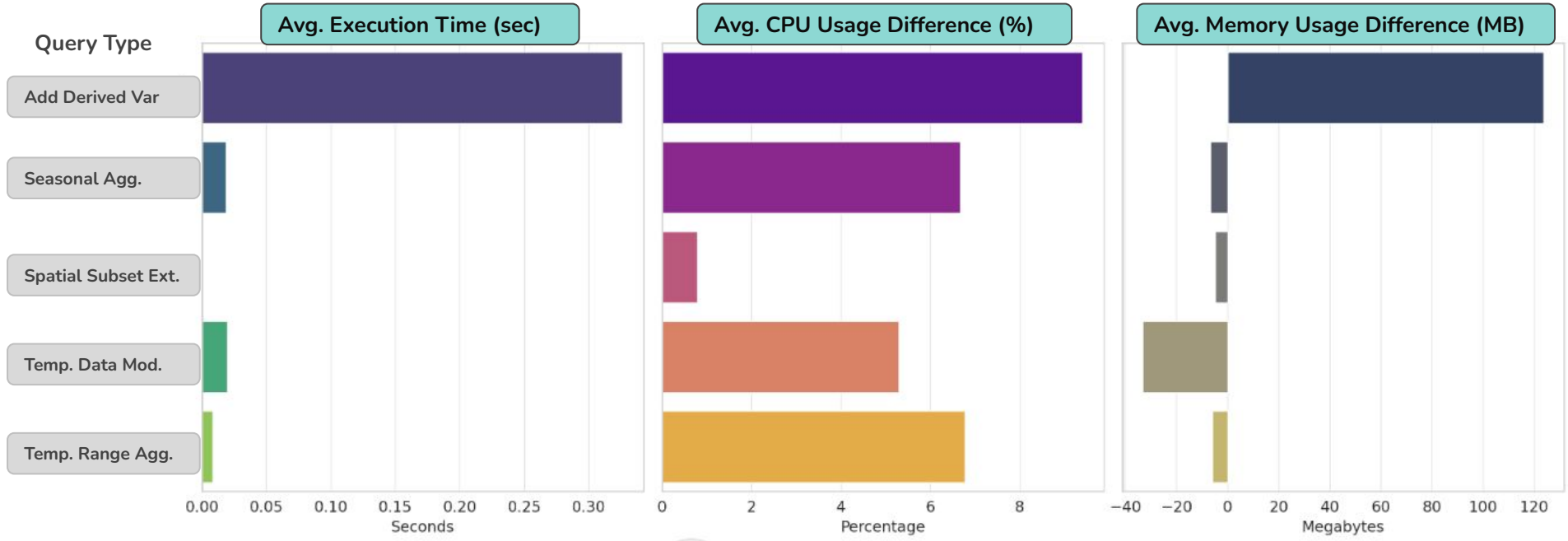
Workload Execution

➔ takes time range & operation type (mean/sum) from workload, executes query

➔ logs before/after results of resources used

benchmark results!

Evaluation:


Query Type
Add Derived Var
Seasonal Agg.
Spatial Subset Ext.
Temp. Data Mod.
Temp. Range Agg.
Avg. Execution Time (sec)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Seconds
Avg. CPU Usage Difference (%)
0
2
4
6
8
Percentage
Avg. Memory Usage Difference (MB)
−40
−20
0
20
40
60
80
100
120
Megabytes

Next Steps

- ➔ Comparison between similar tools & systems (i.e Rasterio, DuckDB)
- ➔ Diversifying query types, extending dataset
- ➔ Testing in different computational environments
- ➔ Include additional volumes from ERA5 & other datasets
- ➔ Compare with other geospatial data analysis tools
- ➔ Create interface for interactive queries & real time data visualization

To Summarize

- ➔ Establishes a comprehensive **benchmark for raster data processing**, focusing on climate data analysis
- ➔ **Core Concept:** proposes a framework which separates the queries & workload - ranging from simple selections to complex aggregations
- ➔ **Benchmark Results** give an insight to different performance metrics, helping to select the most suitable tool based on scenario
- ➔ Addresses **critical gaps in current methodologies**, offering a new perspective on evaluating different systems for complex geospatial data

References

- J. H. Stephan Hoyer, “Xarray n-d labeled arrays and datasets in python,” 2023. Accessed on 2023-10-30.
- G. Szárnyas, J. Waudby, B. A. Steer, D. Szakállas, A. Birler, M. Wu, Y. Zhang, and P. Boncz, “The ldbc social network benchmark: Business intelligence workload,” Proceedings of the VLDB Endowment, vol. 16, no. 4, pp. 877–890, 2022.
- P. Cudre-Mauroux, H. Kimura, K.-T. Lim, J. Rogers, S. Madden, M. Stonebraker, S. B. Zdonik, and P. G. Brown, “Ss-db: A standard science dbms benchmark,” Under submission, vol. 114, 2010.
- R. A. R. Zalipynis, “Array dbms: past, present, and (near) future,” Proceedings of the VLDB Endowment, vol. 14, no. 12, pp. 3186–3189, 2021.
- C. C. C. Service, “Ecmwf reanalysis v5 (era5),” 2023. Accessed on 2024-03-20.
- QGIS 2.18 Documentation. Raster data, 2024. Accessed on 2024-03-20.
- ECMWF. https://cds.climate.copernicus.eu/#!/home, 2024. Accessed on 2024-03-20.