Definition of a Benchmark for Raster Data Processing

Proteeti Prova Rawshan

November 1, 2023

Type of work: Research Project
Student: Proteeti Prova Rawshan
Supervisor: Dr.-Ing. Marcus Paradies
Chair: Prof. Dr.-Ing. habil. Kai-Uwe Sattler
Department: Department of Databases and Information Systems
Semester: Winter semester 2023/2024

Keywords: benchmarking; raster data processing; query formulation; benchmark metrics

Motivation

Benchmarking plays a crucial role in assessing the performance of various raster data processing systems, which are widely used in applications such as environmental modeling and geospatial analysis. These systems, including open-source tools like xarray [1], exhibit diverse functional scopes and performance characteristics, making it challenging to select the most suitable one for specific tasks. The absence of a standardized benchmark for raster data processing inhibits unbiased comparisons among these tools, hindering users' ability to make informed choices. To address this gap, there is a pressing need to develop a comprehensive benchmark encompassing a range of representative queries, from simple data access to complex analytics. This benchmark will provide an objective means to evaluate different systems and assist users in selecting the most appropriate tool for their needs, addressing the challenges highlighted in the LDBC Social Network Benchmark [2], the SSDB benchmark [3], and a tutorial on raster data processing systems [4].

Task Description

The research project aims to develop a benchmark for comparing raster data processing systems, with the primary goal of creating a prototype that can run against at least two systems: xarray and a relational database system, such as DuckDB. The research questions to be addressed include defining a set of representative and objective queries for the benchmark, focusing initially

on the query side, and later considering data aspects. This entails selecting data sets, determining the types of queries (e.g., point queries and aggregations), and evaluating the performance of the systems, either through multiple implementations or by assessing individual systems' query performance and identifying bottlenecks. The project also aims to adhere to established benchmarking guidelines to ensure the benchmark's quality.

Scientific Approach

The research approach for this project involves utilizing various data sources, such as the ERA5 dataset [5] and the xarray documentation, to define and develop a benchmark for raster data processing systems. Additionally, the project will explore existing benchmarking efforts in this field. The methodology for creating the benchmark is inherently creative, making it distinct from typical research methods. The implementation process will include two key aspects: a reference implementation of specific queries in xarray and the development of a workload driver, which will issue queries based on configurable parameters. This workload driver will be tailored for Python-based systems, and the data ingestion process is simplified, with no complex indexing or import required. Regarding metrics, the project will assess how systems perform under the benchmark, considering metrics like query execution time, throughput (queries per second), and query latency. However, the specific metrics will be determined in the course of the project.

Time Schedule

Time period	**Tasks**
Nov 1st	finishing and uploading this exposé
Nov 15th	acquainting with the dataset and exploration
Dec 31st	query formulation
Feb 10th	reference implementation
Feb 20th	implementation of the workload driver
Feb 29th	evaluation
March 1st	finishing and uploading of the final project report.

References

[1] J. H. Stephan Hoyer, "Xarray n-d labeled arrays and datasets in python," 2023. Accessed on 2023-10-30.

[2] 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.

[3] 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.

[4] R. A. R. Zalipynis, "Array dbms: past, present, and (near) future," *Proceedings of the VLDB Endowment*, vol. 14, no. 12, pp. 3186–3189, 2021.

[5] C. C. C. Service, "Ecmwf reanalysis v5 (era5)," 2023. Accessed on 2023-10-30.