A multi-objective optimization algorithm for EO data processing based on Dask library

Arthur Lalayan [1,2] , Hrachya Astsatryan [1], Gregory Giuliani [3,4]
1 Institute for Informatics and Automation Problems of NAS RA, Armenia
2 National Polytechnic University of Armenia
3 Institute for Environmental Sciences, University of Geneva, Switzerland,
4 UNEP/GRID Geneva, Switzerland

Background and motivation

Earth observation (EO) data are widely used for environmental monitoring, urban occupation analysis, or risk detection. The parallel processing of EO data using High-Performance Computing (HPC) or cloud resources improves the processing performance of the growing amounts of geospatial data. Parallel computing frameworks divide the workflow into chunks for further parallel processing on several independent computational worker nodes to speed up the simulation time. The distributed computing libraries scale the workflow using mainly a master-slave architecture, such as the open-source parallel python Dask library. Multiple computational and data platforms and environments can be configured and customized for the Dask, including virtual, physical, cloud-based, and on-premises solutions.

For efficient and optimal processing of EO data, it is necessary to manage and consider several aspects, such as the cost and the utilization of a set of running computational nodes. Therefore, finding a tradeoff between cost and performance is an actual challenge, as it depends on the input data size, the number and the complexity of processing instructions, and other factors. The article aims to propose a Pareto-based collaborative multi-objective optimization algorithm, which is implemented on the suggested scalable EO data processing platform. A multi-objective optimization non-dominated sorting genetic algorithm is applied to the historical dataset of experiments to find the optimal point considering cost and performance.

The evaluation results of the algorithm for several EO data processing workflow is presented, which offers users the optimal amount of resources by considering the number of Dask worker nodes and computational resources per each. The normalized difference vegetation index (NDVI) has been used for the experiments using Dask clusters with a different number of worker nodes. For each case, the processing time is found, and based on it, the price is calculated.

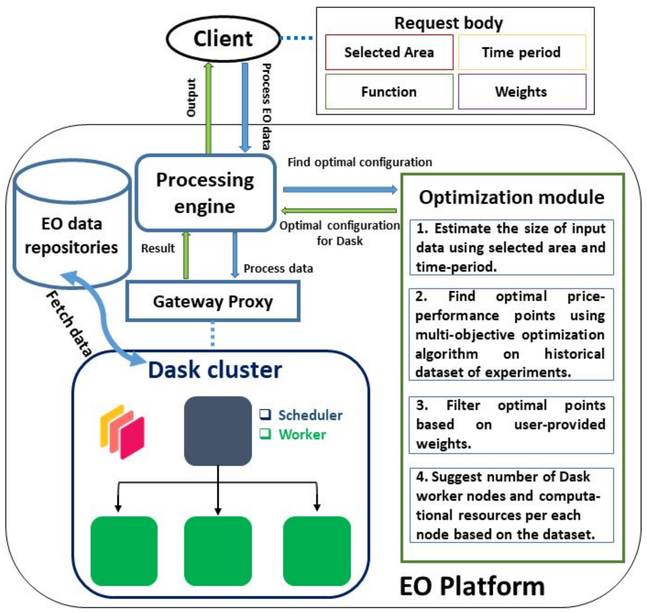
Experimental results

To find the optimal trade-off between EO data processing performance and the cost of computing resources, several experiments are carried out and the results are collected. For instance, an average NDVI has been evaluated for the territory of Armenia using the data from the Sentinel-2 satellite, considering different Dask clusters with varying numbers of worker nodes and amounts of computing resources. Standard instances of global cloud providers were chosen as the specification for the Dask worker node. Three types of workloads light, medium, and heavy are selected (see table).

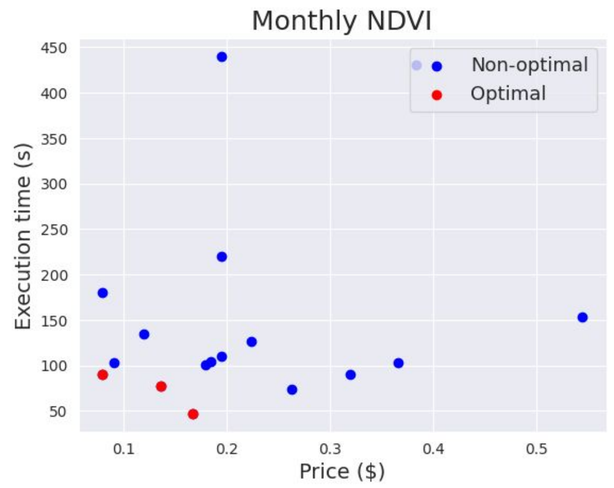
Workload type	Studied period	Input data size
Light	Week	148 GB
Medium	Month	624 GB
Heavy	Season	2.34 TB

The execution time for each type of each simulation is stored as a simulation result, and a price is calculated based on the number of worker nodes, the number of CPUs per node, the execution time of the task, and standard hourly rates for VM instances.

Optimal price-performance points are determined by applying a Pareto-based multi-objective optimization non-dominated sorting genetic algorithm on the historical dataset of experiments. Those points are filtered based on user-provided weights considering price or performance importance. Using this approach, a decision-making service is implemented as a part of the EO data processing platform, which estimates the input data size of the user request and recommends the optimal configuration and computational resources for the Dask cluster. The skeleton of the service is presented in the figure below.



The result of the multi-objective optimization algorithm is presented for the monthly workload in the figure, where the red points correspond to the performance-cost optimal points.



When using one of the optimal points, the performance increase will be approximately 2 times, and resource costs 1.7 times less compared to the average cases.

Conclusion

A multi-objective optimization algorithm is applied to the historical dataset of experiments to find the optimal computing resources for the Dask cluster to process the EO data. The experimental results show the benefits of the suggested algorithm. In the case of calculating the average monthly NDVI, the algorithm can improve the performance by 2 times and reduce the cost of the computational resources by 1.7 times compared with the average cases. The algorithm is implemented as a part of the suggested scalable EO data processing platform.

Acknowledgments

The research was supported by the University of Geneva Leading House and the State Committee of Science of the Republic of Armenia by the projects entitled ”ADC4SD: Armenian Data Cube for Sustainable Development”, ”Self-organized Swarm of UAVs Smart Cloud Platform Equipped with Multi-agent Algorithms and Systems” (Nr. 21AG-1B052) and ”Remote sensing data processing methods using neural networks and deep learning to predict changes in weather phenomena” (Nr. 21SC-BRFFR-1B009).