Multi-Cloud workflows with Pangeo and Dask Gateway

Tom Augspurger[1], Martin Durant[2], Ryan Abernathey[3], and Joe Hamman[4]

[1]taugspurger@continuum.io
[2]Anaconda Inc.
[3]Columbia University
[4]CarbonPlan

November 22, 2022

Abstract

As more analysis-ready datasets are provided on the cloud, we need to consider how researchers access data. To maximize performance and minimize costs, we move the analysis to the data. This notebook demonstrates a Pangeo deployment connected to multiple Dask Gateways to enable analysis, regardless of where the data is stored. Public clouds are partitioned into regions, a geographic location with a cluster of data centers. A dataset like the National Water Model Short-Range Forecast is provided in a single region of some cloud provider (e.g. AWS's us-east-1). To analyze that dataset efficiently, we do the analysis in the same region as the dataset. That's especially true for very large datasets. Making local "dark replicas" of the datasets is slow and expensive. In this notebook we demonstrate a few open source tools to compute "close" to cloud data. We use Intake as a data catalog, to discover the datasets we have available and load them as an xarray Dataset. With xarray, we're able to write the necessary transformations, filtering, and reductions that compose our analysis. To process the large amounts of data in parallel, we use Dask. Behind the scenes, we've configured this Pangeo deployment with multiple Dask Gateways, which provide a secure, multi-tenant server for managing Dask clusters. Each Gateway is provisioned with the necessary permissions to access the data. By placing compute (the Dask workers) in the same region as the dataset, we achieve the highest performance: these worker machines are physically close to the machines storing the data and have the highest bandwidth. We minimize cost by avoiding egress costs: fees charged to the data provider when data leaves a cloud region.

Multi-Cloud workflows with Pangeo and Dask Gateway

Tom Augspurger, Martin Durant, Ryan Abernathey, Joe Hamman

As more analysis-ready datasets are provided on the cloud, we need to consider how researchers access data. To maximize performance and minimize costs, we move the analysis to the data. This notebook demonstrates a Pangeo deployment connected to multiple Dask Gateways to enable analysis, regardless of where the data is stored. Public clouds are partitioned into regions, a geographic location with a cluster of data centers. A dataset like the National Water Model Short-Range Forecast is provided in a single region of some cloud provider (e.g. AWS's us-east-1). To analyze that dataset efficiently, we do the analysis in the same region as the dataset. That's especially true for very large datasets. Making local "dark replicas" of the datasets is slow and expensive. In this notebook we demonstrate a few open source tools to compute "close" to cloud data. We use Intake as a data catalog, to discover the datasets we have available and load them as an xarray Dataset. With xarray, we're able to write the necessary transformations, filtering, and reductions that compose our analysis. To process the large amounts of data in parallel, we use Dask. Behind the scenes, we've configured this Pangeo deployment with multiple Dask Gateways, which provide a secure, multi-tenant server for managing Dask clusters. Each Gateway is provisioned with the necessary permissions to access the data. By placing compute (the Dask workers) in the same region as the dataset, we achieve the highest performance: these worker machines are physically close to the machines storing the data and have the highest bandwidth. We minimize cost by avoiding egress costs: fees charged to the data provider when data leaves a cloud region.