Cloud-Native Repositories for Big Scientific Data

Ryan Abernathey[1], Tom Augspurger[1], Anderson Banihirwe[1], Charles C Blackmon-Luca[1], Timothy J Crone[1], Chelle L Gentemann[1], Joseph J Hamman[1], Naomi Henderson[1], Chiara Lepore[1], Theo A Mccaie[1], Niall H Robinson[1], and Richard P Signell[1]

[1]Affiliation not available

January 19, 2021

Abstract

Scientific data has traditionally been distributed via downloads from data server to local computer. This way of working suffers from limitations as scientific datasets grow towards the petabyte scale. A "cloud-native data repository," as defined in this paper, offers several advantages over traditional data repositories—performance, reliability, cost-effectiveness, collaboration, reproducibility, creativity, downstream impacts, and access & inclusion. These objectives motivate a set of best practices for cloud-native data repositories: analysis-ready data, cloud-optimized (ARCO) formats, and loose coupling with data-proximate computing. The Pangeo Project has developed a prototype implementation of these principles by using open-source scientific Python tools. By providing an ARCO data catalog together with on-demand, scalable distributed computing, Pangeo enables users to process big data at rates exceeding 10 GB/s. Several challenges must be resolved in order to realize cloud computing's full potential for scientific research, such as organizing funding, training users, and enforcing data privacy requirements.

Hosted file

`Cloud_Native_Repositories_for_Big_Scientific_Data.pdf` available at https://authorea.com/users/372628/articles/490577-cloud-native-repositories-for-big-scientific-data