Developing Data Pipelines for Research and Analysis

Seth Goodman[1]* Jacob Hall[1] Cheyenne Hwang[1]

[1]AidData, Global Research Institute, William & Mary

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: containers, geospatial, data, replication, scalable

1 Introduction

The growing demand for data and computational analysis in both research and educational initiatives stretches beyond disciplines such as computer or data science. As the efforts of faculty and students within institutions such as William & Mary become increasingly data-driven, traditional methods of utilizing critical resources such as high performance computing (HPC) clusters and disseminating the resulting work can present practical barriers to innovation. This proposal aims to explore solutions to these issues through the development of a research use case which leverages open source software to build computational environments known as containers that can be deployed to run on computers ranging from individual laptops to HPC clusters or cloud-based clusters with minimal modification.

The use case for the proposed work will focus on evaluating the impact of Chinese financed mining projects on vegetation levels in surrounding areas. The analysis will incorporate an automated data pipeline to acquire satellite imagery on vegetation levels, extract information around mining sites identified by AidData's Global Chinese Development Finance Dataset, and assess trends. The pipeline and analysis are intended to provide an illustrative and adaptable template for a wide range of potential computational tasks incorporating large scale data such as satellite imagery that could be built and deployed using containers. The ability to develop research concepts on one machine, scale up for analysis on a cluster, then distribute to others for replication and review - without having to modify, rebuild, or troubleshoot code - is essential to accelerating the integration of data and computational analysis across fields.

Reference how this work served as the prototype for AidData's own scaled up data processing efforts using Kubernetes (https://github.com/aiddata/geo-datasets). Add more details on how it facilitated this, the amount of data we have processed, and how it has made it easier / supports GeoQuery (add in stats about GeoQuery usage, etc that this will help with).

Acknowledgements Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

*smgoodman@wm.edu