High-performance geospatial analysis

Max Aragón

January 2023

1 Introduction

High Performance Computing (HPC) in a cluster for satellite imagery big data is a method of leveraging powerful computer systems and networks to efficiently process and analyze large amounts of satellite imagery data. This approach utilizes parallel processing and distributed computing to divide the workload among multiple machines, allowing for faster and more efficient data processing. By using a cluster of interconnected computers, HPC can handle the complex and computationally intensive tasks associated with processing large amounts of satellite imagery data, such as image classification, feature extraction, and data fusion. This can enable faster and more accurate decision-making, improved understanding of global phenomena, and more efficient use of resources.

In this report, I summarise the results obtained from parallelising the computation of a series of algorithms using Dask in the SEVIR dataset (Veillette et al, 2020).

1.1 Dask vs Apache Sedona

Dask and Apache Sedona are both open-source big data processing frameworks, but they have different design goals and are more suited to different tasks. Dask is designed to be a flexible and powerful parallel computing library for analytics, enabling users to harness the full power of their CPU and memory resources. Dask can handle large datasets that don't fit into memory, and can also be used for distributed computing across a cluster. Dask is often used for data processing, machine learning and scientific computing workloads. On the other hand, Apache Sedona, is a spatio-temporal big data management system that is built on top of Apache Spark. It provides specific spatial and temporal functionalities for analyzing large amounts of geospatial data. Sedona is mainly used for geospatial data processing tasks like spatial join, range query, and nearest neighbor search. It can handle big data but it's not built to handle large datasets that don't fit into memory. When it comes to big data satellite imagery analysis, Dask is more appropriate than Apache Sedona as Earth observation imagery is often very large in size.

2 Methods

2.1 Data

SEVIR can be downloaded from the Registry of Open Data on AWS. SEVIR contains two main files, a catalog and the data. The catalog is a csv file including metadata of every single event, such as image size, projection, coordinates, time in UTC, percentage of missing values, etc. The data folder contains subdirectories with each image type organised by year in multiple HDF5 files. For this particular proof of concept, a single file of 3GB containing 36 events and a total of 1764 images was processed with and without parallelisation.

2.2 Algorithms

2.2.1 Downsampling

Data downsampling is a technique used to reduce the size of large satellite imagery datasets. This is done by reducing the resolution of the images, which results in fewer pixels and less data to process and analyze. This process can be performed by resizing the images, either by reducing the number of pixels or by aggregating them. Downsampling can be useful for reducing computational requirements and storage space, as well as for making images more manageable for analysis. However, it also results in loss of information, so it should be used with caution and in accordance with the specific analysis and application needs. Additionally, it is important to note that downsampling should be done after all necessary preprocessing steps have been completed, such as atmospheric and geometric correction, to avoid losing important information.

2.2.2 k-means clustering

K-means clustering is a popular technique for grouping similar data points together based on their features. In the context of satellite imagery, this technique can be used to classify different regions of an image based on the visible and infrared channels of the data. The Geostationary Operational Environmental Satellite (GOES) system provides visible and infrared channels that can be used to identify different features on the earth's surface. By applying k-means clustering to the visible and infrared channels of GOES data, it is possible to classify regions of the image based on their spectral characteristics. For example, different land cover types, such as vegetation, urban areas, and water bodies, can be identified by their unique spectral signatures in the visible and infrared channels. Additionally, K-means clustering can also be used to identify cloud patterns and atmospheric conditions.

2.2.3 Frame animation

Data frame animation is a method of compressing satellite imagery data by creating a single animation from a series of images. This is done by concatenating the images into a single file format, such as a GIF or video, which can be played back in a sequence to create an animation. By compressing multiple images into a single animation, the amount of data needed to store and transmit the images is significantly reduced. This technique can be used as a pre-processing method for optical flow, which is a technique for measuring the motion of objects in an image. By having a smaller data size, the computation time for optical flow will be faster and more efficient. Data frame animation can also be used to create visualizations of time-series data, such as changes in vegetation or land use over time. It should be noted that the animation process can also result in losing some information, so it is important to consider the specific application needs before compressing the data.

3 Results

The SEVIR dataset, consisting of 1TB of satellite imagery, was processed using a combination of three image processing algorithms: downsampling, k-means clustering, and frame animation. To evaluate the performance benefits of utilizing Dask for parallel computation, two scripts were implemented. The results revealed that, when applied to a single 3GB image, the original script took 166 seconds to process, while the script utilizing Dask completed the task in 135 seconds. This highlights the scalability of the Dask framework, as it has the potential to effectively process the entire SEVIR dataset in a timely manner.

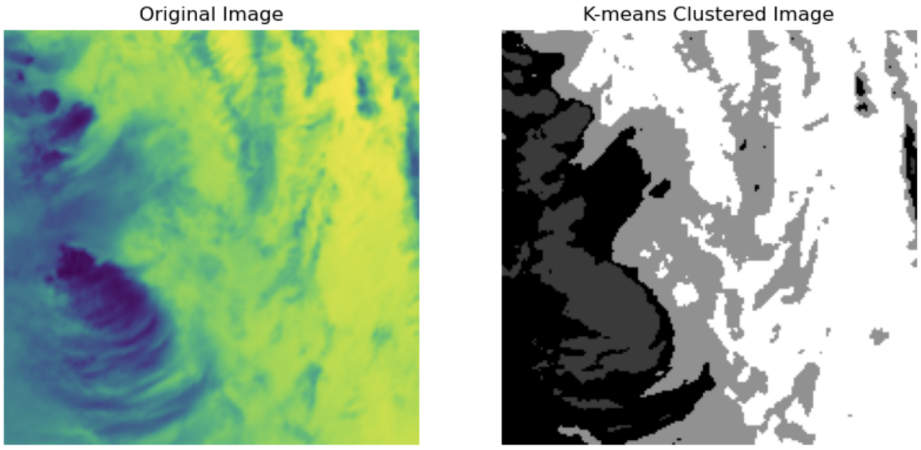

Figure 1: Example of an infrared clustered image into 4 k's

4 Conclusions

In summary, Dask and Apache Sedona are both powerful big data processing frameworks, but they are designed for different types of workloads. Dask is well suited for handling large datasets that don't fit into memory and is often used for data processing, machine learning, and scientific computing. On the other hand, Apache Sedona is designed for analyzing large amounts of geospatial data and is mainly used for geospatial data processing tasks. For big data satellite imagery analysis, Dask's ability to handle large datasets and perform image processing makes it a more appropriate choice than Apache Sedona.

5 References

Veillette, M., Samsi, S., & Mattioli, C. (2020). Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. Advances in Neural Information Processing Systems, 33, 22009-22019.