CloudButton

Serverless Big Data Analytics with Lithops

Josep Sampé (URV), Pedro Garcia Lopez (URV)

7th International Workshop on Serverless Computing (WoSC7) 2021

https://www.serverlesscomputing.org/wosc7/demos/d4

http://cloudbutton.eu

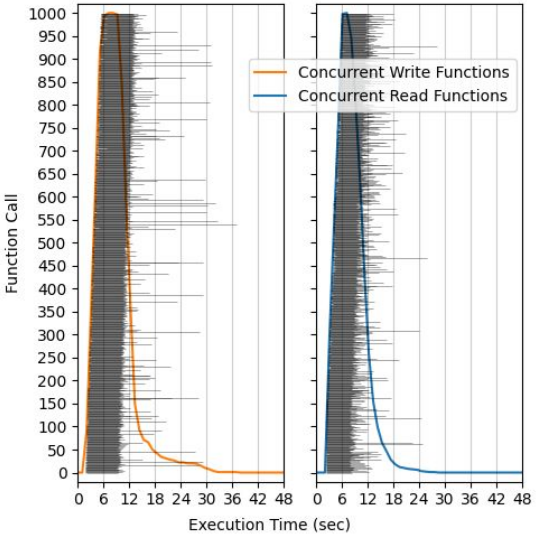
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825184.

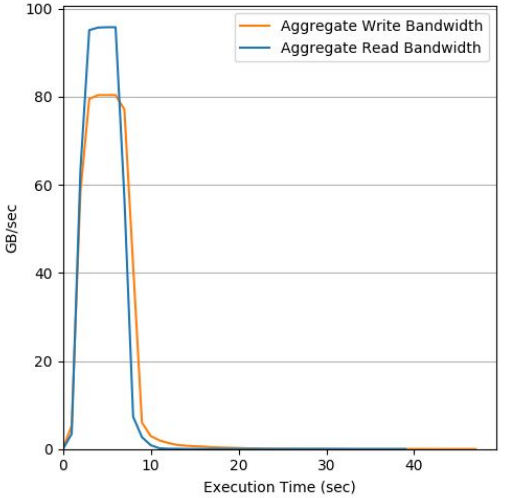
Why Serverless ?

- Good **for stateless embarrassingly parallel Jobs**
- Auto-scaling and parallelism: **1000 parallel vCPUs** and **> 1000 GFLOPS**
- **100 GB/s** aggregate bandwidth between Serverless Funcionts and Cloud Object Storage

Amazon **Lambda**

Amazon S3





https://github.com/lithops-cloud/applications/tree/master/benchmarks

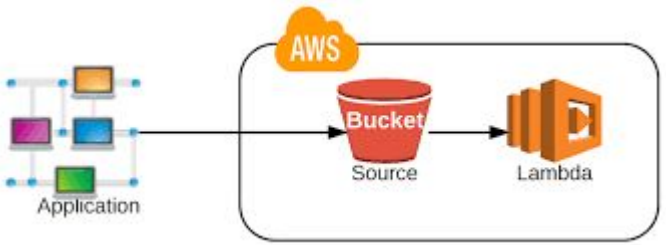
Lithops

Lithops can be defined as a multi-cloud python computing framework that allows multithreaded, local applications to be transparently scaled to massive Cloud resources.

https://github.com/lithops-cloud/lithops

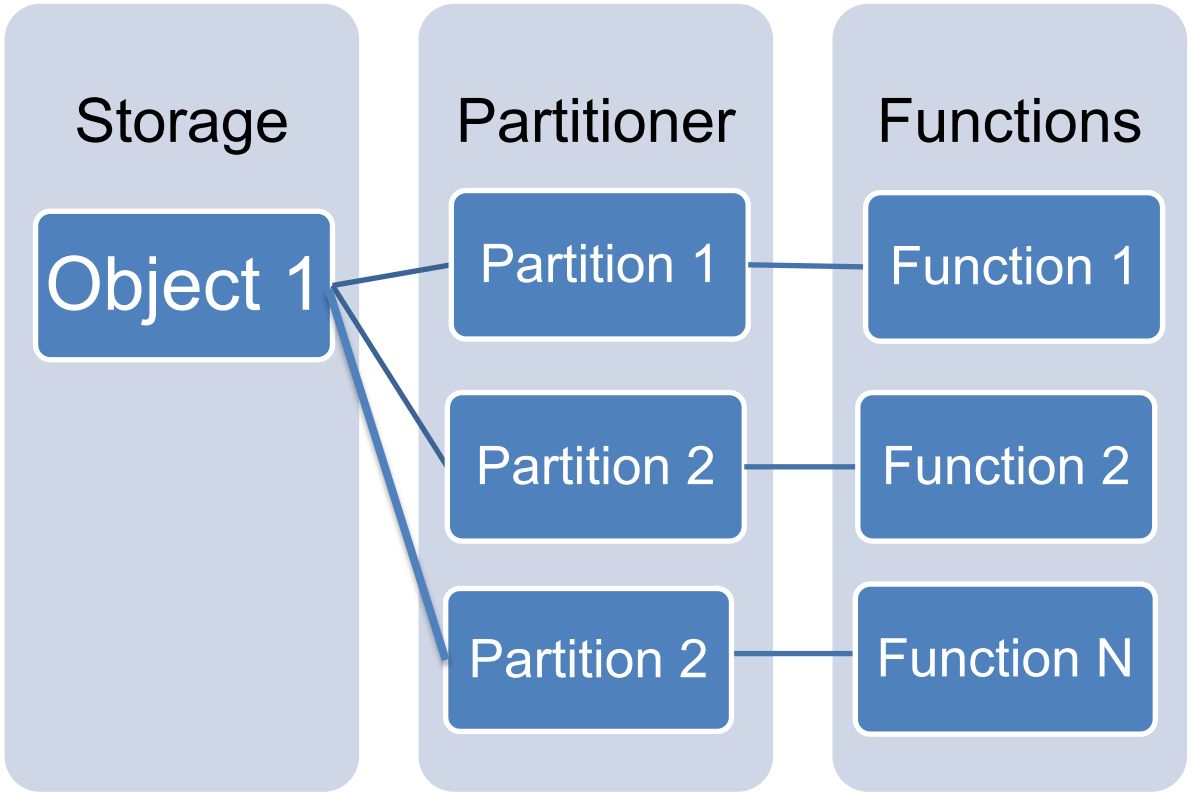
Lithops is data-driven

- Map functions can operates over objects in a bucket
- Lithops partitioner permits on the fly data partitioning to Serverless functions from Object Storage




```python
import lithops

# Bucket with prefix
data_location = 'cos://lithops-sample-data/test/' # Change-me


def my_map_function(obj):
    print('Bucket: {}'.format(obj.bucket))
    print('Key: {}'.format(obj.key))
    print('Partition num: {}'.format(obj.part))
    counter = {}

    data = obj.data_stream.read()

    for line in data.splitlines():
        for word in line.decode('utf-8').split():
            if word not in counter:
                counter[word] = 1
            else:
                counter[word] += 1

    return counter


if __name__ == "__main__":
    fexec = lithops.FunctionExecutor(log_level='DEBUG')
    fexec.map(my_map_function, data_location)
    print(fexec.get_result())
```

Lithops is data-driven



- Preprocessing
  - LIDAR
  - imlZ
- On the fly
  - .txt
  - .csv
  - .gzip

Lithops Geospatial data analytics

Data from Sentinel 2:

- The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery.
- The mission provides a global coverage of the Earth's land surface every 5 days
- Data is available from June 2015
- Data is stored in a public AWS S3 Bucket

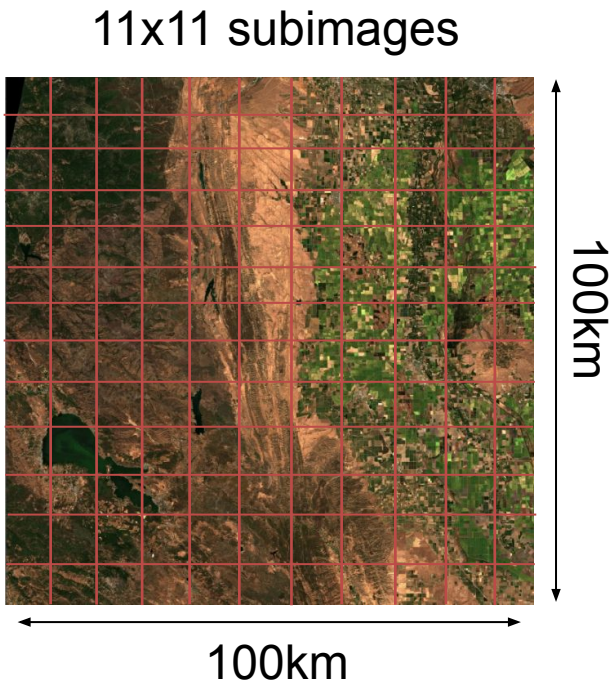
https://registry.opendata.aws/sentinel-2/

Lithops Geospatial data analytics

Cloud optimized GeoTIFF

- Original Sentinel 2 data must be processed by one *product* per function
- Cloud optimized GeoTIFFs is a format that allows to split a *product* in 11x11 subimages
- With the Cloud optimized GeoTIFFs we can spawn a maximum of 121 parallel functions per *product* □ *Speedup processing*

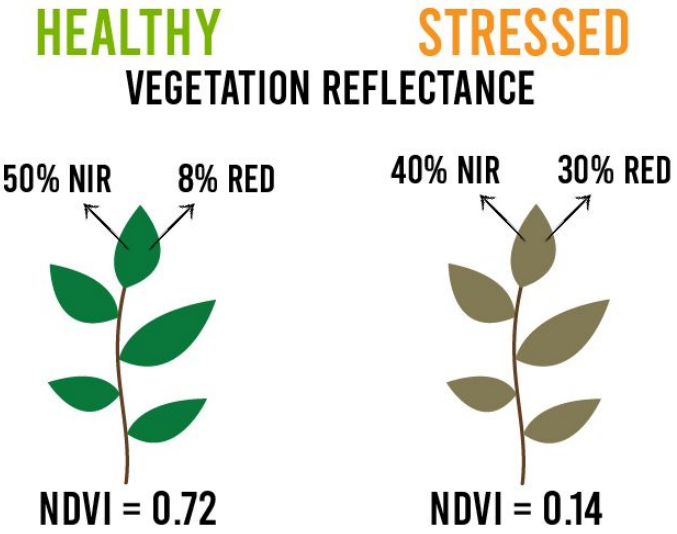
https://registry.opendata.aws/sentinel-2-l2a-cogs/



Lithops Geospatial data analytics

Example: **Calculate the NDVI difference between 2 dates**

- NDVI (Normalized Difference Vegetation Index) is a simple graphical indicator that can be used to analyze remote sensing measurements, assessing whether or not the target being observed contains live green vegetation.
- To calculate the NDVI we need both band04 (RED) and band 08 (NIR)

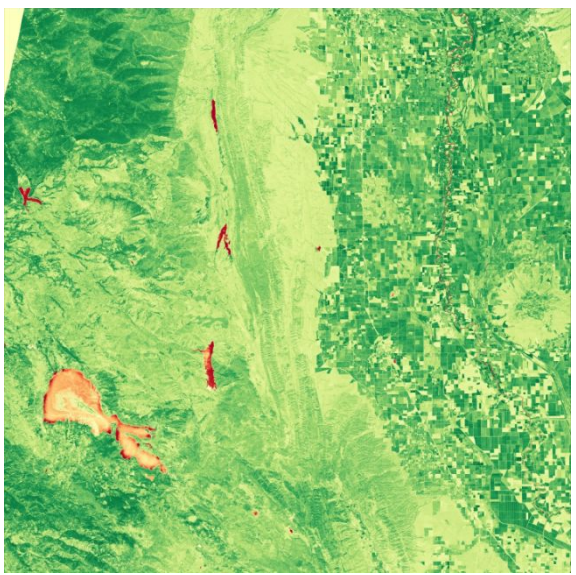


Lithops Geospatial data analytics

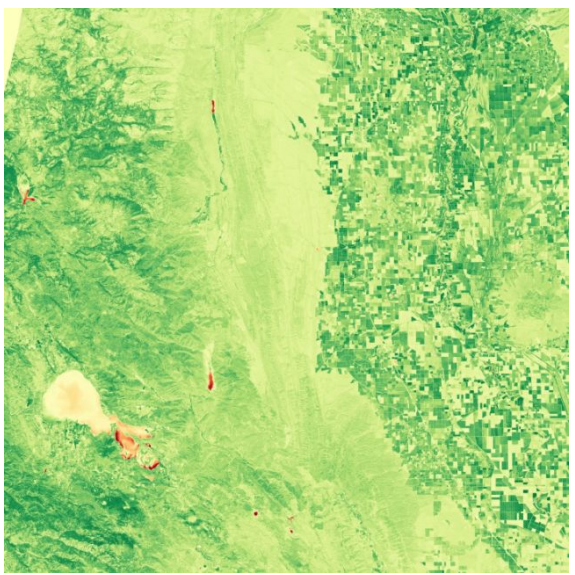
Example: **Calculate the NDVI difference between 2 dates**

- ~25 seconds to compare the NDVI from 2 different dates with Lithops

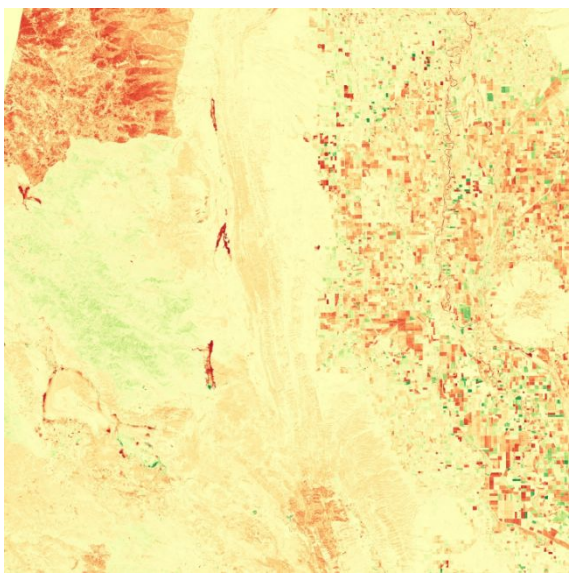
NDVI 16/ Sept /2019



NDVI 16/ Sept /2021



NDVI DIFF



THANK YOU!

Josep Sampé (URV), Pedro Garcia Lopez (URV)

7th International Workshop on Serverless Computing (WoSC7) 2021

https://www.serverlesscomputing.org/wosc7/demos/d4

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825184.