CS 205 Project - M4

Parallelizing Crop Phenology reports via NDVI datasets

Team 18:

Elaine Swanson, Alex Johnson, Saketh Mynampati, Alyssa Ross

Sequential Baseline:

- **Data Collection**
  - Obtained NDVI data from VegScape.
  - Goal – is to also obtain CropLand data from CropScape
- **Data Iteration and Extraction**
  - Process files sequentially from a specified directory
  - Extract month and day from each filename.
  - Create a `Geotiff` object to handle file data.
  - Retrieve raster band data as vectors.
- **Data Analysis**
  - Initialize an `Image` object with Geotiff data.
  - Calculate maximum, mean, and min NDVI value for each month.
  - Update a map with the NDVI values per month.
- **Performance Monitoring**
  - Utilize PAPI for real-time performance metrics.
  - Monitor and recorded time for file operations and data processing.
- **Results**
  - Report maximum NDVI values by month.
  - Showcase performance metrics for efficiency analysis.

```
month: 11, day:12, year:2023
month: 11, day:06, year:2023
month: 11, day:19, year:2023
month: 11, day:13, year:2023
month: 11, day:26, year:2023
month: 11, day:20, year:2023
month: 12, day:03, year:2023
month: 11, day:27, year:2023
month: 12, day:10, year:2023
month: 12, day:04, year:2023
month: 12, day:17, year:2023

NDVI trends:
Year:2023
Month:1, max ndvi: 255, mean:207.127, min:121
Month:2, max ndvi: 255, mean:208.461, min:121
Month:3, max ndvi: 255, mean:201.201, min:122
Month:4, max ndvi: 255, mean:205.602, min:124
Month:5, max ndvi: 255, mean:221.31, min:119
Month:6, max ndvi: 255, mean:225.877, min:125
Month:7, max ndvi: 255, mean:224.481, min:109
Month:8, max ndvi: 255, mean:220.454, min:90
Month:9, max ndvi: 255, mean:219.492, min:123
Month:10, max ndvi: 255, mean:220.171, min:125
Month:11, max ndvi: 255, mean:218.958, min:124
Month:12, max ndvi: 255, mean:218.298, min:123
Year:2024
Month:1, max ndvi: 255, mean:197.146, min:121
Month:2, max ndvi: 255, mean:211, min:123
Month:3, max ndvi: 255, mean:209.834, min:123
Month:4, max ndvi: 255, mean:216.984, min:123
Month:5, max ndvi: 0, mean:-nan, min:0
Month:6, max ndvi: 0, mean:-nan, min:0
Month:7, max ndvi: 0, mean:-nan, min:0
Month:8, max ndvi: 0, mean:-nan, min:0
Month:9, max ndvi: 0, mean:-nan, min:0
Month:10, max ndvi: 0, mean:-nan, min:0
Month:11, max ndvi: 0, mean:-nan, min:0
Month:12, max ndvi: 0, mean:-nan, min:0

Reading files time:
twall:                          3.09669
Total cycles:                138600925
Total instructions:          193897438
Instructions per cycle (IPC): 1.39896
Float performance:           3.22926e-10

Data reduction time:
twall:                          4.22383
Total cycles:                272299283
Total instructions:          562258036
Instructions per cycle (IPC): 2.06485
Float performance:           2.36752e-10
(geospatial) [ajohnson@hal serial]$
```

Sequential Profile

- **Data Collection**
  - Some queries arbitrarily take longer to process, ranging from 60 ms to 300 ms, and occasionally, queries fail and must be retried.
- **Data Iteration and Extraction**
  - Process files sequentially from a specified directory
  - Extract month and day from each filename.
  - Create a `Geotiff` object to handle file data.
  - Retrieve raster band data as vectors.
- **Data Analysis**
  - Initialize an `Image` object with Geotiff data.
  - Calculate maximum, mean, and min NDVI value for each month.
  - Update a map with the NDVI values per month.
- **Performance Monitoring**
  - Utilize PAPI for real-time performance metrics.
  - Monitor and recorded time for file operations and data processing.
- **Results**
  - Report maximum NDVI values by month.
  - Showcase performance metrics for efficiency analysis.

```
month: 10, day:16, year:2023
month: 10, day:29, year:2023
month: 10, day:23, year:2023
month: 11, day:05, year:2023
month: 10, day:30, year:2023
month: 11, day:12, year:2023
month: 11, day:06, year:2023
month: 11, day:19, year:2023
month: 11, day:13, year:2023
month: 11, day:26, year:2023
month: 11, day:20, year:2023
month: 12, day:03, year:2023
month: 11, day:27, year:2023
month: 12, day:10, year:2023
month: 12, day:04, year:2023
month: 12, day:17, year:2023

NDVI trends:
Year:2023
Month:1, max ndvi: 255, mean:207.127, min:121
Month:2, max ndvi: 255, mean:208.461, min:121
Month:3, max ndvi: 255, mean:201.201, min:122
Month:4, max ndvi: 255, mean:205.602, min:124
Month:5, max ndvi: 255, mean:221.31, min:119
Month:6, max ndvi: 255, mean:225.877, min:125
Month:7, max ndvi: 255, mean:224.481, min:109
Month:8, max ndvi: 255, mean:220.454, min:90
Month:9, max ndvi: 255, mean:219.492, min:123
Month:10, max ndvi: 255, mean:220.171, min:125
Month:11, max ndvi: 255, mean:218.958, min:124
Month:12, max ndvi: 255, mean:218.298, min:123
Year:2024
Month:1, max ndvi: 255, mean:197.146, min:121
Month:2, max ndvi: 255, mean:211, min:123
Month:3, max ndvi: 255, mean:209.834, min:123
Month:4, max ndvi: 255, mean:216.984, min:123

GDAL Geotiff initialization:
twall:                          3.04398
Total cycles:                138521193
Total instructions:          193897710
Instructions per cycle (IPC): 1.39977

Data reduction:
twall:                          4.16184
Total cycles:                273153556
Total instructions:          590023850
Instructions per cycle (IPC): 2.16004

Totals:
Execution time:7.20582s, GDAL init(42.2434%), Reduction(57.7566%)
Traffic: (Byte/s):           5.65494e+08
Instructions per byte:       0.730783
(geospatial) [ajohnson@hal serial]$
```

Proposed Parallelization:

- **Data Collection**
  - Parallelize data downloads and interleave with computation to hide latency.
  - Some queries arbitrarily take longer to process or fail.
- **Data Iteration and Extraction**
  - Parallelize read/write operations dynamically across 36 available cores,
  - Cache cropland data to data read from NDVI.
- **Data Analysis**
  - Communication overhead involves communicating relevant crop data between cores processing the same cropland and communication for reductions.
  - Uniformly sized tiles avoid large load imbalances (i.e. large counties)
  - Reductions take place across batches of land and time - minimize with a log-reduction for all NDVI summary metrics.
  - Overheads can be minimized by dynamically chunking.

Parallelization Continued: **Last Time vs. Now**

County Level Analysis