Maximizing Science in Big Data

Astronomy

Hayden Smotherman
University of Washington
Department of Astronomy
Table of Contents:

I. Timeseries: An Overview of Scalable Science

II. Light curves: A Scientific Use Case

III. Images: GPU and HPC Computing
Table of Contents:

I. Timeseries: An Overview of Scalable Science

II. Light curves: A Scientific Use Case

III. Images: GPU and HPC Computing
The Five Vs

**Volume:** The magnitude of the data being processed

**Velocity:** The rate of data that must be processed

**Variety:** The complexity of the data: structured and unstructured

**Veracity:** The fidelity of the data and the associated noise

**Value:** The scientific value or information contained within data as well as the cost.
AXS: Astronomy eXtensions for Spark

- AXS enables big data science by implementing common operations using familiar python and astropy commands.

- AXS extends Spark with data partitioning scheme, sort-merge join optimization, and provides efficient parquet file access to light curves (e.g. 2.9 billion ZTF objects)

Zečević et al 2019
Cesium Scaling Test: Volume and Velocity

Selecting the right hardware configuration can reduce runtime by an order of magnitude.
Selecting the right hardware can reduce costs, but selecting the wrong hardware can cost more without saving time.
Table of Contents:

I. Timeseries: An Overview of Scalable Science

II. Light curves: A Scientific Use Case

III. Images: GPU and HPC Computing
AXS: A Scientific Application

Boyajian's star
Science is about iteration and filtering

AXS helps create an intuitive approach for embedding light curve analyses within UDFs.

Iteratively process the data to filter and correct errors. Repeatedly pruning the data helps to manage the memory footprint.

Final computational resources are modest (450 core hours to process 1.4B light curves) because data are down selected for complex analyses.

There is a growing need for software engineering in astronomy.

Boone et al 2022
Table of Contents:

I. Timeseries: An Overview of Scalable Science

II. Light curves: A Scientific Use Case

III. Images: GPU and HPC Computing
Sifting Through the Static: Moving Object Detection in Difference Images – Smotherman et al. (accepted AJ)
The Kuiper Belt

- A disk of objects beyond the orbit of Neptune extending from about 30 au to 45 au.

- The Kuiper Belt contains objects that have been only lightly-perturbed since shortly after the formation of the Solar System.

Image Credit: https://commons.wikimedia.org/wiki/File:Kuiper_belt_plot_objects_of_outer_solar_system.png
Current Methods: Digital Tracking

- Classical searches use “shift and stack”
  - If examining stacks by eye, limited to small grid of velocity/angle shifts
- Update of this method is “digital” or “synthetic” tracking
  - Algorithmically search for objects along many trajectories and then add the likelihood to find detections
  - Computationally expensive

Credit: Zhai et al. (2014)
The KBMOD algorithm

- Add masks to images, or use difference imaging
- Convolve images with PSF to create Maximum Likelihood Images
- Sum the likelihoods of trajectory locations of the individual images
- KBMOD can search >10^10 trajectories in about a minute on 10-15 4Kx4K images using a 1080 GPU
CNN Filter of False Positives
Recovery as a Function of Known V Mag

Known Object Recovery
0.846 Completeness

- Faintest single-image 10σ depth
- All Objects
- Recovered Objects

Number of Objects

V

0 2 4 6 8 10 12 14 16
20.5 21.0 21.5 22.0 22.5 23.0 23.5 24.0
Comparison to Known Data

Recovered Results vs Predicted Results

- Initial Y Position (arcsec)
  - Recovered Position
  - Predicted Position

- Y Velocity (arcsec/hr)
  - Recovered Velocity
  - Predicted Velocity

- Initial X Position (arcsec)

- X Velocity (arcsec/hr)

- Position Residual (arcsec)
  - Median Residual

- Speed Residual (arcsec/hr)
Inclination Distribution

Brown Inclination Distribution KS Test

- Expected Uniform Distribution
- Observed Distribution

Number of Objects

Inclination
Comparison to Known Magnitude Distribution
Table of Contents:

I. Timeseries: An Overview of Scalable Science

II. Light curves: A Scientific Use Case

III. Images: GPU and HPC Computing
Summary and Challenges

● It’s not just the volume of data; it’s the complexity of the science
  ○ We should focus on how the physics of scientific questions might simplify the computational models.
  ○ Data quality can place important yet hard-to-characterize constraints on large-scale analysis.

● Scaling science is more than just more hardware
  ○ Tools don’t always exist that are robust enough to scale for a typical science analysis.
  ○ Many of the scalable analysis frameworks do not scale easily with memory constraints.
Questions?