University of Washington Astronomy
Survey Science Group

Astronomy In The Cloud: Using MapReduce For Image Coaddition

K. Wiley[1], A. Connolly[1], J. Gardner[2], S. Krughoff[1], M. Balazinska[3], B. Howe[3], Y. Kwon[3], Y. Bu[3]

1. University of Washington, Department of Astronomy 2. University of Washington, Department of Physics 3. University of Washington, Department of Computer Science

1 Introduction

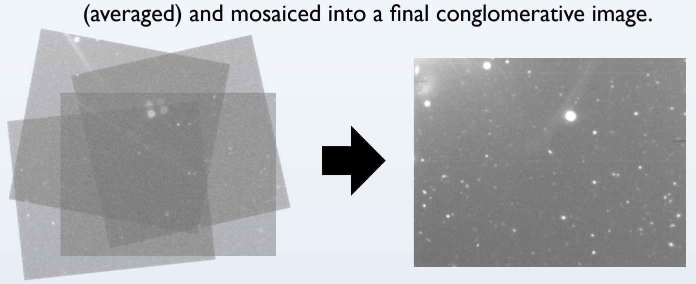
In the coming decade, astronomical surveys of the sky will generate tens of terabytes of images and detect hundreds of millions of sources every night. The study of these sources will involve computational challenges such as anomaly detection, classification, and moving object tracking. Since such studies require the highest quality data, methods such as image coaddition, *i.e.*, registration, stacking, and mosaicing, will be critical to scientific investigation.

With a requirement that these images be analyzed on a nightly basis to identify moving sources (asteroids) or transient objects (supernovae), these datastreams present many computational challenges. Given the quantity of data involved, the computational load of these problems can only be addressed by distributing the workload over a large number of nodes. However, the high data throughput demanded by these applications may present scalability challenges for certain storage architectures.

One scalable data-processing method that has emerged in recent years is *MapReduce* and its popular open-source implementation called *Hadoop*. In the Hadoop framework, the data is partitioned among storage attached directly to worker nodes, and the processing workload is scheduled in parallel on the nodes that contain the required input data. A further motivation for using Hadoop is that it allows us to exploit cloud computing resources, *i.e.*, platforms where Hadoop is offered as a service.

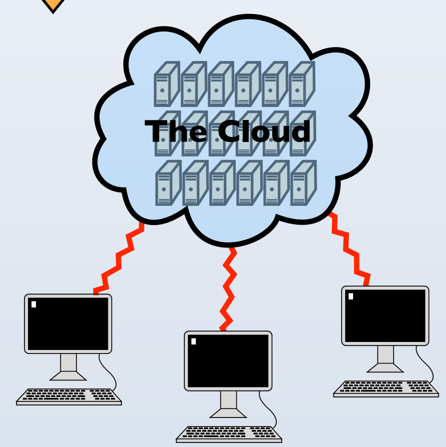
2

Our research focuses on distributed *image coaddition*, wherein multiple partially overlapping images are background-subtracted, registered (aligned), PSF-matched, and finally stacked (averaged) and mosaiced into a final conglomerative image.

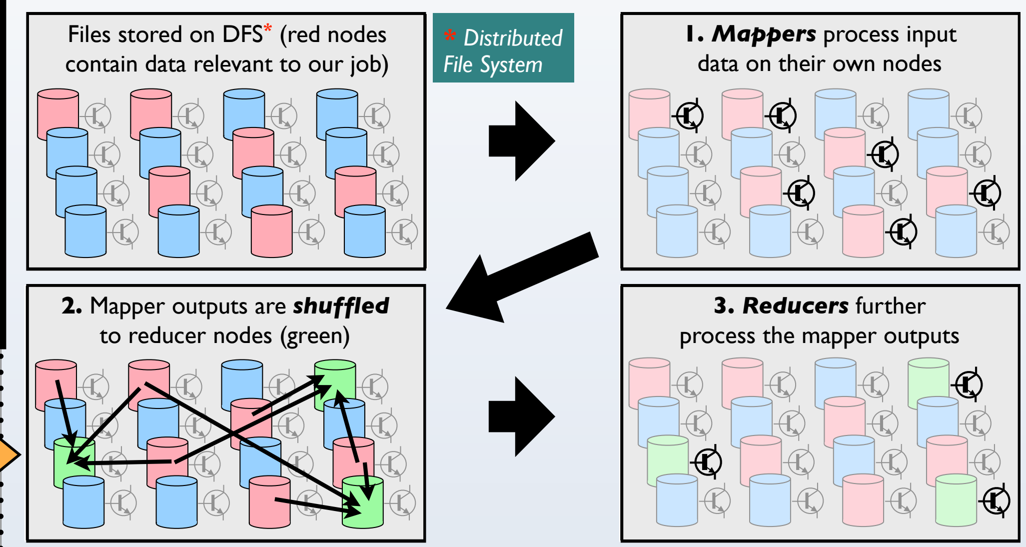


3

Computing *clouds* offer massive clusters as an on-demand service. Users create their own programs and then submit and run them remotely. This arrangement empowers users to use the cloud in their own way while simultaneously mitigating geographic inconveniences.

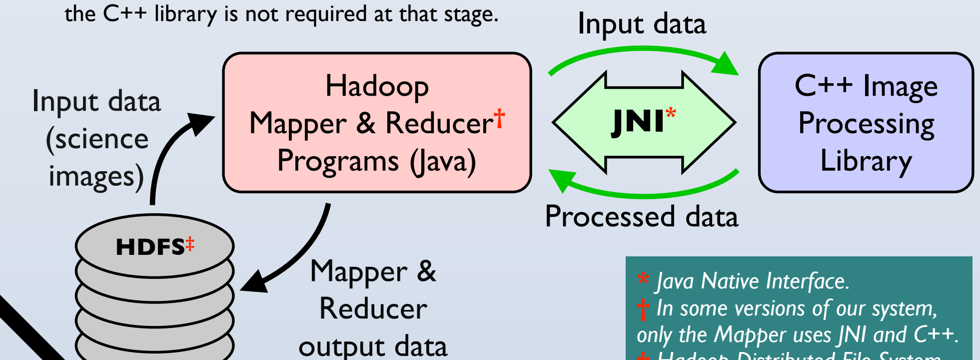


4 MapReduce



5

We wish to use a large in-house C++ image-processing library, but Hadoop is programmed in Java. JNI* coordinates the Java to C++ communication. The Java Mappers do very little work, handing most functionality to C++. However, the Reducer performs its operations in Java as the C++ library is not required at that stage.

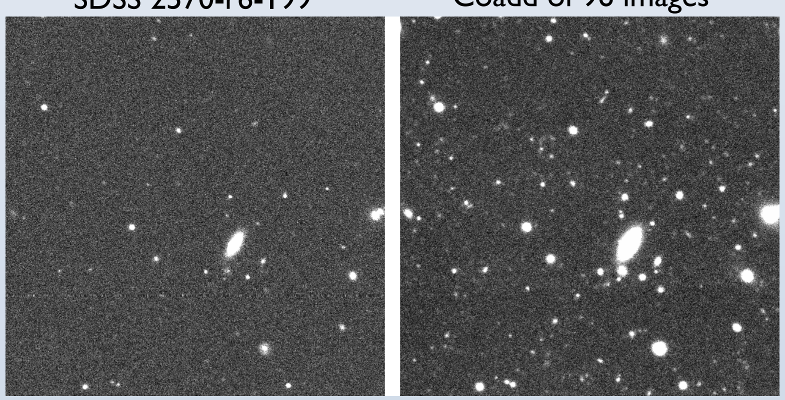


* Java Native Interface.
† In some versions of our system, only the Mapper uses JNI and C++.
‡ Hadoop Distributed File System.

8

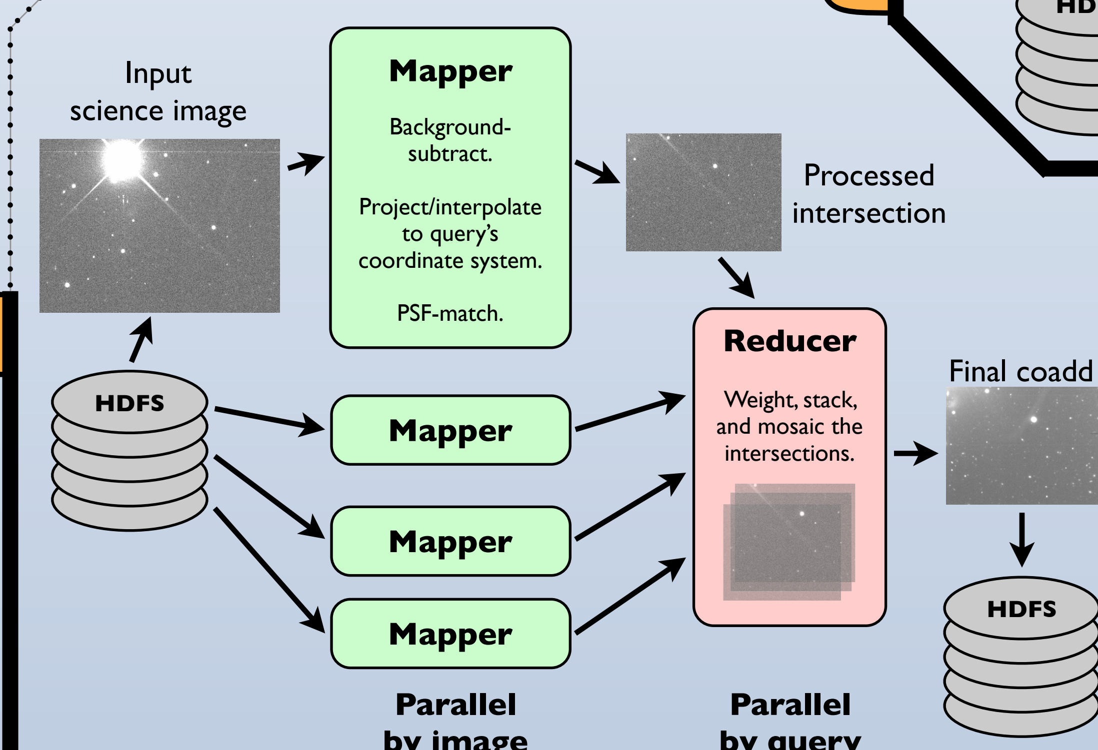
These images show a single input image and an example of a coadd generated using our system. The input dataset is the *Sloan Digital Sky Survey, Stripe 82*. We observe many additional faint sources in the coadd.

SDSS 2570-r6-199 Coadd of 96 images*



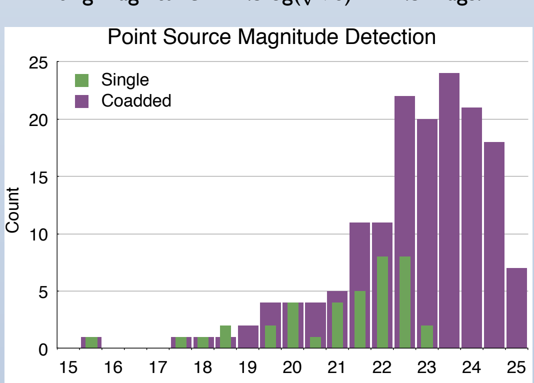
* *Coverage is not necessarily 96 at any given pixel*

7 Image Coaddition in Hadoop



9

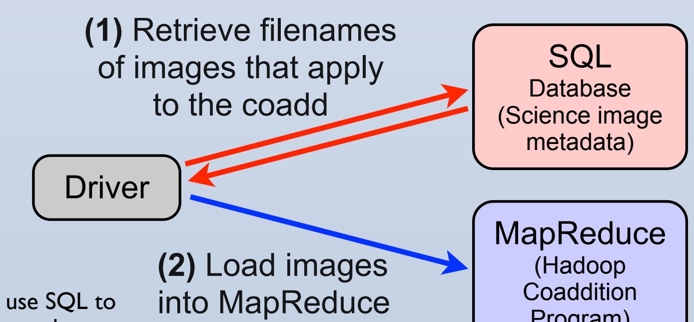
Given a coadd of 96 images, the theoretical improved limiting magnitude = $-2.5\log(\sqrt{96}) \approx -2.5$ mags.



This histogram of point source detections demonstrates that we have achieved an improvement of ~2 mags, slightly less than the theoretical ~2.5 mag improvement, which is expected considering the footnote in ⑧.

6

Hadoop operates on the entire input dataset, but image coaddition only requires a tiny subset of the total images, only those which overlap the query bounds. We must therefore trim the input dataset.



We first use SQL to find the relevant image filenames in a meta-database, then pass only those images to Hadoop.

This work is funded by the NSF Cluster Exploratory (CluE) grant (IIS-0844580) and NASA grant 08-AISR08-0081. The CluE cluster is funded through the CluE grant and maintained by IBM and Google. We thank them for their continued support. We further wish to thank both the LSST group in the astronomy department and the database research group in the computer science department at the University of Washington.