Big Instruments, Large Communities: Data Management in the Decade of the 2020s

Theme: State of the Profession Consideration

Point of Contact: William D. Gropp (National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois)

Co-authors: Margaret W. G. Johnson, Daniel S. Katz, and Donald Petravick (National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois)

Summary

As the scale of the astronomical instruments and the reliance of larger communities on observational data grow, the data ecosystem for astronomy needs to evolve to assure that working scientists have access to the data they need. Given a trend for key science to require data from multiple missions, we foresee the need for a facility that enables multi-instrument pixel-level science, independent from any one observing program. We also comment on what is needed to provide an adaptive, agile, and cost-effective computing environment that is responsive to the community. We stress the adaptation of professional methods and cultivation of the requisite skills within the field to produce processes to understand where computing is going, and to understand the total costs for these ambitious large-scale datasets, rather than asking the committee to pre-maturely endorse a specific technical strategy.

Key Issues and Overview of the Impact on the Field

This white paper focuses on two principal issues: (1) being well prepared for the era in which pixel-level processing and data management involving multiple instruments, such as LSST and WFIRST, is needed to satisfy science goals; and (2) enabling scientists to concentrate on the science, which requires professionalizing the data management in a manner that is sensitive to science needs.

The volume of data produced by astronomy will be dominated by decade-long general-purpose surveys providing data to broad communities. The multi-decade extent of these projects requires a sustainable approach to data management infrastructures and the ability to adapt as technology changes. The availability of high-performance, high-throughput, and cloud-native software approaches provides a variety of potential options for implementing data management systems.

The increase in data size and scale calls for increasing specialization and organizational factoring, which is best done by adapting current professional roles to large astronomical facilities. Changes such as this need to be carefully introduced to the community.

Strategic Goals

Because of the multi-instrument nature of future investigations, we foresee an essential separation of roles where (1) each instrument has teams who transform the pixel-level data from its very raw form, as produced by the instrument, into calibrated quantities that are usable for making scientific conclusions; and (2) a facility, independent of any one observing instrument, supports pixel-level investigations including combining data from multiple sources. This facility needs to be operational in the 2020s. The separation of 1) producing excellent science-ready datasets from an observing program and 2) extracting scientific knowledge from those datasets enables a rich science program. A key aspect of this modified ecosystem is building and maintaining active collaborative relationships between the instrumentation projects, the facility, and the astronomical community.

Provide for Scientific Needs Independent of Any Observing Project: The following examples illustrate scientific needs that would be best fulfilled by an independent facility.

(1) *Fusion of science-ready datasets from different instruments, generating yet more pixels and value-added products.* The facility would allow access to data from, for example, LSST, Euclid, WFIRST, the CMB-S4 legacy survey, and other relevant instruments. An astrophysicist would be able to access pixel-level data from every instrument measuring the sky. Astronomers and astrophysicists could do their analysis on computers at the facility or just retrieve a subset of the data. For Multi Messenger Astrophysics (MMA), the facility would support joint analysis of MMA instrumentation to find phenomena that are just beyond the grasp of any one instrument.

(2) *Detailed integration of the output of simulations to observation at the pixel level.* In its most advanced form, simulations produce pixel data that is processed to pixels, including adding artifacts due to atmospheric conditions, optical effects, and instrumental response of the detectors, and then processed in the same manner as the analogous physical observations. This compensates for biases introduced in the observed data, and leads to a more precise comparison of how well a physical theory fits physical reality.

(3) *Pixel-level studies by individual investigators.* Support of investigations that may be of very specialized scientific interest and beyond the scope of standard data products produced by the instrumental project themselves. These products have been called “value added products” in survey astronomy.

(4) *Refined calibration of instrument data.* When measurements are made of the same phenomena by different instruments, fused data provides a viewpoint on the calibrations of each instrument that is different than the viewpoint when the view is constrained to data from the instrument itself.

Integrate Specialized Skills Needed for Large Scale Science: As mentioned at the 2019 NSF Large Facilities Workshop[1], increased specialization and professional skills, recommended by NAPA[2] and now required by NSF. We see this as including not only project management professionals, but also governance skills and processes that engage the community; systems and software architects using well-known methodologies such as The Open Group Architecture Framework[3] (TOGAF); software engineering teams building foundational infrastructure using well-documented and

[1] https://www.largefacilitiesworkshop.com/wp-content/uploads/2019/04/ProjectMgtCoreComp.pdf
[2] https://www.napawash.org/uploads/Academy_Studies/NSF_Phase_2_Comprehensive_Report.pdf
[3] https://www.opengroup.org/togaf

structured approaches (e.g., Agile); financial management adapted for large facilities from standards such as Technology Business Management[4] (TBM) framework; and integration and operations methodologies such as DevOps and ITILv4[5]. The ensemble of these skills enables an optimal overall implementation of a data management system. A key scaling and reuse challenge is to integrate these specialized skills into effective interactions with the astronomy community for which this is new, and to provide sound feedback to funding agencies. NSF has examples in the from of TrustedCI[6] and the Cyberinfrastructure Center for Excellence[7] (CICoE) pilot project.

Separate Software, Provisioning, Management, Security, and Support within the Facility: Joint pixel-level processing of massive datasets implies co-location of pixel data, which in turn implies all of the responsibilities of a large facility, without prejudice of whether the facility is realized on commercial clouds, on-premises, or some hybrid; or whether the facility is centrally located or distributed. The implementation of an optimal strategy will likely vary over the decade due to many factors, including cost, available technology, acceptability to the user community. A facility composed of the requisite professional strengths will provide for an effective execution in a changing technological and market environment. Such an ensemble of skills can effectively guide the facility over multiple decades.

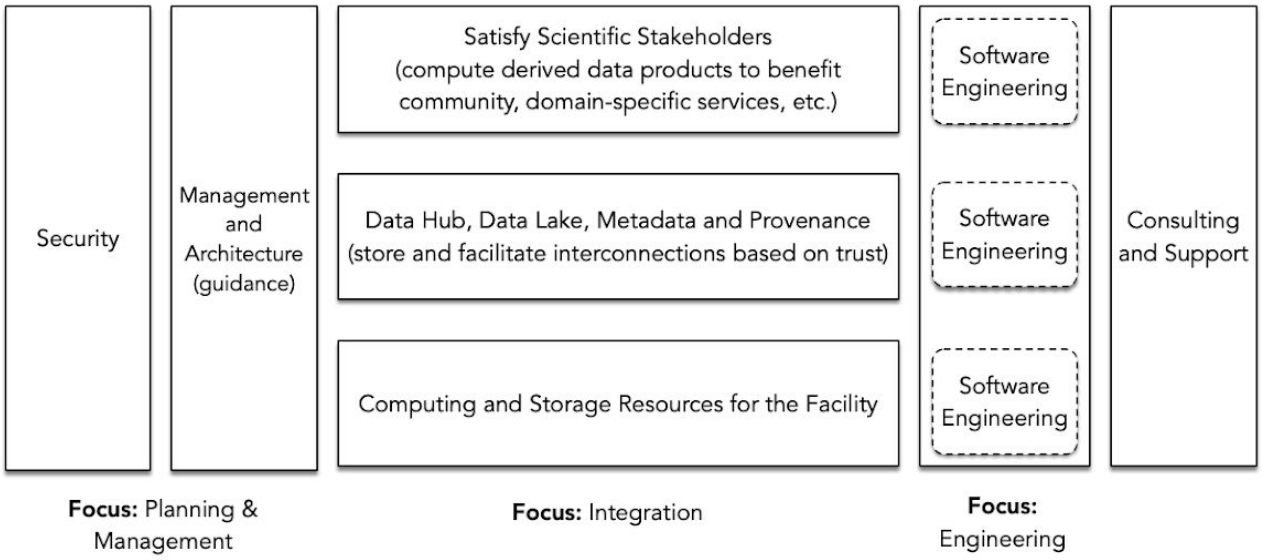

Figure 1: Guide to classifying all elements of a comprehensive pixel-level processing facility.

[4] https://www.cio.gov/priorities/tbm/
[5] https://www.axelos.com/best-practice-solutions/itil
[6] https://trustedci.org/
[7] http://cicoe-pilot.org/

Figure 1 illustrates a functional breakdown of a pixel-level facility, without speaking to specific details of the implementation of each box or a particular organizational structure. Each rectangle represents an area where a best practice for the pixel facility needs to be implemented while maintaining a coherent facility, and illustrates the need for specialized skills commensurate with the investment in the facility.

At a strategic level, managing a facility is about managing risk. Examples of risk areas that would be actively managed include: (1) lock-in due to the data gravity of the large number of pixels; and (2) reliance on multiple software provider communities used to implement the facility and their viability, including cloud-native, high-throughput, and high-performance. It is interesting to note that all these communities provide software that is substantially agnostic to where it is instantiated, either on-premises or on a variety of commercial clouds. These different software communities have different strengths, and we see no reason to make a meaningful decadal decision.

Properly Position the Facility in the Larger Scientific Ecosystem: Specialized skills are needed to make sensible, defensible decisions about data management infrastructure. While algorithms and calibrations are specific to astronomy, pixel-level processing and data management methods are generic to a wider range of disciplines. Technical topics of interest that are generic to many fields include data fusion, virtual data, machine learning, resource management, and other topics related to large data. Skills common to all fields also include the managerial, technical and operational skills mentioned above that are needed, regardless of how systems are physically provisioned. Additional benefits, such as sustaining datasets beyond the lifetime of the projects that produced them can be foreseen from such an approach, as scientific, technical and managerial acumen needed for custodial care of these datasets would persist.

Organizations, Partnerships and Current Status

Of necessity, our plan requires instantiation of a pixel-processing facility, which we see as an instrument-agnostic, astronomy-aware, new facility.

An important aspect of any facility is its intimate relationship with the community it serves, and we see the governance relationship between the community and the facility as key to its success. We can see the governance model having a strong science advisory board, and scientific input on resource allocation committees for limited tangible resources and subsidies for use of commercial resources.

With respect to the implementation of the facility itself, for staff and expertise we are aware of supporting agency initiatives, such as the NSF Harnessing the Data Revolution[8] (HDR) activities, one of which calls for developing institutes for Data-Intensive Research in Science and Engineering[9] (DIRSE) and would necessarily incorporate some of the specialities we have mentioned above. An ideal solution for pixel-level processing of large survey datasets would provide an analogous framework for pixel-level processing of astronomical datasets and perhaps datasets from other domains to achieve an economy of scale. This effort would necessarily be multi-agency and can in practice be informed by the progress in the Tri-Agency Group (TAG), which is considering multi-agency needs for pixel processing for LSST, Euclid, and WFIRST[10]. As mentioned above, experience from existing NSF groups such as TrustedCI and the CICoE pilot project, and experience from provisioning projects serving multiple domains, such as XSEDE[11], can be brought to bear to identify professional skills and science community engagement, which would result in a maximally effective strategy, given the constraints.

Schedule

We foresee that development of the processes needed for managing and operating large-scale persistent support for pixel-level facilities for astronomy and similar sciences would begin organizing immediately and build on the expertise we have cited above along with data management experts from current and past experiments.

Critical dates are based on when science-ready data products for astronomy are available from LSST, Euclid, WFIRST, CMB-S4, and any relevant international projects. The current schedule of Euclid and LSST would see first light in the early 2020s, followed by the production of substantial science-ready datasets. A practical approach is to initially instantiate the facility proximate to LSST, considering the data volume, and allow the facility to gather the skills mentioned in the strategy and allow for its independent evolution while taking on WFIRST data, CMB-S4 data, and other datasets of interest to the community. Ideally, the facility would be free to make additional arrangements with other domains to achieve maximum leverage of its processes, its business arrangements, and any provisioning obligations.

---

[8] https://www.nsf.gov/cise/harnessingdata/sciencedrivers/index.jsp
[9] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505631&WT.mc_id=USNSF_41&WT.mc_ev=click
[10] https://www.nsf.gov/attachments/190967/public/Rhodes_AAAC_9_2017_v3.pdf
[11] https://www.xsede.org/

Cost and Business Considerations

Within decadal timescales, technology and technology markets change. A facility would need to evaluate any potential solutions based on total cost of ownership, evaluate long-term risks, and make the best choice for science based on facts about the risks and costs. A facility broadly supporting the community would consider its own resources plus other kinds of computing available to investigators, which can consist of commercially provided cycles, as in cloud computing; systems that the investigators own and control, which have useful lifetimes approaching a decade; viable computing physical facilities that are already sunken costs and are useful for over a decade; and computing awarded on academic merit provided by agencies and other sources. Given these software, services, and provisioning considerations, a facility will need to focus on how best to support the science while evaluating the costs and risks of potential solutions as they emerge and evolve.

Example topics that would be evaluated in this context are:

(1) *Software.* The community's tendency is towards open-source software. Most importantly, it allows investigators to understand what code does because they can read it and contribute to it as needed. Additionally, it provides a stability stemming from a diverse community of contributors. This stability is relied on to mitigate a number of risks for the operational phase of projects.

(2) *Provisioning.* Cloud provisioning has emerged as a necessary consideration for the types of facilities we envision. While there is a good deal of community interest in clouds, we expect that a full evaluation would consider risks such as "(1) getting locked in to a single vendor, (2) unknown future storage cost, (3) potentially uncapped costs for terminating a vendor, (4) security restrictions, and (5) trust in the network access technologies."[12] An evaluation, if conducted now, would plausibly reach conclusions that commercial cloud provisioning makes sense for certain cases, perhaps supporting retail-scale science due to irregular, urgent, or occasional demand load or unwieldy access management due to number of individual users.

Additional risk evaluations arise due to the prestige value of astronomy science. Because of the scientific prestige of astronomy, projects may benefit from a

[12] https://www.nap.edu/read/24938/chapter/7#177

subsidy underwritten by a commercial cloud provider in exchange for hosting the data. However, the fundamental economics for large astronomy datasets differs from other use cases we have seen cited in genomics and earth science. A fundamental distinction between astronomy and these domains is that genomics and earth science data are able to answer questions of social, economic, and business importance,[13] and hence are able to be monetized, with commercial cloud costs mitigated or even waived because of the market potential of the datasets[14]. While we see no such monetization potential for the astronomical data produced in the 2020s that would obviously sustain gifts or substantial discounts from commercial providers, we also recognize that cloud vendors can use science data holdings to bring on users who want to compute on that data, and can increase the vendor/cloud's science prestige and reputation to encourage usage in other science disciplines.

It is important that costs and risks be evaluated fairly and consider both benefits to the community and the total cost of ownership over time, and consider the required attributes such as fitness for purpose. Unlike a high-performance computer, which is more like an upfront five-year fixed investment, the provisioning must be adaptable, so we also envision the pixel-processing facility to be “evergreen,” meaning that provisioning contracts provide for flexibility at least on the scale of a year, and that any physical investments are spread out over time, and indeed may be leased.

Conclusion

We see separation of large-scale pixel processing as part of the natural progression of the supporting data management systems as the instrumentation systems in astronomy become more expensive and serve larger and larger communities. This structuring of data management enhances science by enabling pixel-level processing of multiple instruments. The fundamental decisions are not only to create the facility, but also to ensure that it is well-governed by community input and participation, and to provide ongoing and agile support based on a staff with the requisite professional skills.

---

[13] https://www.researchgate.net/publication/233531004_The_emerging_science_of_environmental_applications

[14] https://www.noaa.gov/big-data-project