2011 Report of the Review Panel for the Status of the DES-Brazil IT Effort 16 December 2011

Reviewers: Enrique Gaztanaga, Bhuv Jain, Steve Kent, Rich Kron (Chair), Huan Lin, Don Petravick, Brian Yanny

The above-named panel conducted a one-day review of the Brazil IT effort on 7 October, 2011, partly as a follow-on to the similar review of October 2010. A number of documents had been circulated to the panel, including one called "Answers to the 2010 Review Panel" that adequately addressed the various points raised in the first review. The presentations were mostly from the IT Team itself and are posted to docDB in connection with the DES Collaboration meeting at U. Penn.

This Report is organized as follows. An Introduction provides general comments about the status of the development effort and some advice on future directions. Then, comments on each of the five parts of the Charge are listed, in no special order. Paragraphs that contain Recommendations are headed by R:. Not all recommendations are for the Brazil team; some are intended to be considered by some part of the DES project.

A number of themes recur in the more detailed discussion given in the main body of this Report; some of the more important ones are summarized here.

1) Very significant progress is evident across all of the major tasks with respect to our review of a year ago. The team is working stably and effectively.

2) The portfolio of tasks is large, yet the resources to undertake the work are limited. Therefore, priorities should be set so that a smaller number of tasks are addressed, at least at first. These top-priority tasks should be accompanied by a clear set of requirements.

3) The process of establishing what those priorities should be requires coordination with the DES Project. On the one hand, DES should declare what it needs (and define the requirements), but on the other hand Brazil needs to declare what it can reasonably deliver.

4) Similarly, the interactions between the Brazil team and the Science Working Groups should be enhanced in terms of establishing priorities and requirements. That members of the Brazil team are also active members of the Science Working Groups (SWG's) is a positive step.

Introduction

The Brazil Portal concept includes a place where large-collaboration astrophysics analysis can be directed: a central place for checking in codes; compiling, linking, and running those codes; making plots and sharing them, all in a repeatable, archived fashion; and the ability to run things in parallel in larger volume. The vision of the Brazil portal is commendable and should go forward.

The portal is trying to do a very large number of things at once: process Precam data, be a Quick Reduce quality assurance system, analyze simulated data, generate new star catalog simulations, run quality assurance on simulated data, run quality assurance on real data, contrast and compare a set of algorithms (photo-z, star/galaxy separation, weak lensing, cluster finders), save source code and running environments, be a general running platform (supporting C, Fortran, python, ...), serve as an interactive database, foster collaboration within a large international group, and so on. Facilities need to be developed and implemented in phases according to agreed-upon priorities, a notion that was implicit in the presentations and the schedule that were provided to the panel.

The Portal infrastructure has come a long way since the last review. New personnel have been hired into needed positions with excellent skills and appropriate experience, and we were pleased to meet them as presenters.

The team has a clear understanding of the need for different environments - development and testing (including a place for pipeline developers to test) versus production. However, some of the "plumbing" issues have been solved by others, and it seems like some re-invention of the wheel is occurring. For example, the "tawala" tool is much like Fermilab's UPS or LSST's EUPS; and NOAO has NHPPS, which looks like it implements similar functionality as the Orchestration module. We encourage the Brazil team to initiate discussions with appropriate groups at Fermilab, NCSA, and LSST before embarking on developing new infrastructure technology. Alternatively, external consultants could be asked for advice.

More work is needed to formalize a set of requirements that can properly inform the priorities and schedule. For example, which codes in particular are appropriate for running inside the Portal, and what is the prioritization? This is not wholly a Brazil responsibility, since the SWG's need to be involved in the process, and also NCSA and FNAL where there are connections. The acceptance tests for the data and the data processing provide a good example of where the SWG's need to step in: defining these tests is the responsibility of the Collaboration, not Brazil. In effect, the DES project as a whole needs to understand what it needs beyond DES DM, and of that, what is appropriate for the Brazil Portal.

A principal concern emerging from the review in October 2010 was the definition of goals, schedule and deliverables. We recognize considerable effort on this front, but there are a couple issues that remain. First, as already mentioned the nominal scope of the Brazil effort is large and ambitious: current plans cover not only all of the Science Working Groups, but extend to PreCam reductions, QR, a tertiary archive, and much more. This scope would be unrealistic without a clear sense of the most urgent current needs, where additional facilities can be added later as resources permit. (This kind of flexibility is in fact an important feature that is enabled by the Portal design.) Second, the goals, requirements, and scope serve local (Brazil) interests as well, so there should be some way to resolve any tension. In this respect, it is worth noting that to properly define a schedule of tasks with priorities, at least three elements are involved: the DES Collaboration interests, local interests, and general infrastructure otherwise needed for the Portal.

The scope documents are very detailed and are useful for defining functionality for the Brazil team. However, they are not a good substitute for laying out the high-level goals and requirements from the DES point of view.

There are a number of facilities that are definitely needed by DES, which are well adapted to the Portal design, and for which some development has already been accomplished. It may make sense to focus on these places (complete the design, development, documentation, and accessibility of the tools) in terms of priorities and resources. The following facilities are illustrative of ones that can define a realistic set of top-level goals for the Portal. Regardless of the specific ones chosen for the top level, it would be best to start with a small list and then progressively add more capabilities.

1) test the DES DM pipeline outputs with respect to the Science Requirements and the DES DM Requirements and Technical Specifications, using the Data Challenges. This is currently being done by the individual SWG's. The Portal systematizes this work and facilitates the sharing of results.

2) similar to 1), except using real data from DECam.

3) run additional analysis codes, beyond the DES DM codes, designed for additional validation tests.

4) provide a platform for validation of DECam early data - making files and tools available to the Collaboration such that feedback to the commissioning team and to the survey planning team is useful and timely.

Timeliness is important: now is the time to be recruiting new users. To bring in more scientists to become users of the Portal, besides advertising it, one must demonstrate that using it is to their advantage. It will take many sessions with a direct interaction with one of the experienced users/developers to successfully build up a user community. Moreover, good system reliability and up-time are important. These issues could require significant resources.

The reviewers appreciated hearing the reports from users who have succeeded in installing and running modules. In the cases where there have been successes in installing code written externally, some members of the Brazil group were needed to learn the details of the code and get involved in adapting it. The process of making new users productive needs to be streamlined. Current practice does not seem sustainable for the operations phase, which should involve minimal overheads for both the SWG's and for Brazil. Clear goals for the operations phase are needed.

A specific plan for using the Portal by the LSS WG illustrates some additional needs and issues. Various tests need to be run, not just validation of codes specific to LSS science, but also tests of data quality and the performance of the DES DM pipelines. As set up now, the Portal includes checks of specific Science Requirements. It is important to enable members of the LSS WG to contribute additional checks of this kind (e.g. with a different interpretation of a Science Requirement), which will enable cross-comparison of the results. This is an excellent example of the utility of the Portal, but there are some practical aspects that need to be accommodated somehow. Devising tests is development, and developers work in non-static environments. There is naturally a period where it makes more sense to undertake development outside of the Portal. At some later time it becomes more appropriate to make the effort to port the code (and any required other files that the test may need to use) into the Portal. At that point, the SWG code can be incorporated into the toolkit available to everyone on the Portal so that others can similarly run checks and cross-compare the results in a systematic way. Changes in the DES DM pipeline code should be easily checkable by running regression tests - this might place performance (speed) constraints on the system.

At several points we have mentioned the need for connections between the SWG's and the Brazil team, and more generally information flow between the DES project and Brazil, but the actual mechanisms have yet to be worked out. One possibility if for a single point-of-contact to be named on the Brazil end to interface to the Science Working Groups, and similarly perhaps the DES Project could propose one person as a point-of-contact on the Project end.

The current model for interaction between the SWG's and the Brazil team involves the transfer of expertise (know-how) from DES SWG's to Brazil Portal personnel. This is a potentially delicate concept because some expertise could be considered to be proprietary. Proceeding where it is clear there is a mutual benefit and mutual consent seems needed for this to work.

Charge Element #1

Based on the written material and the presentations, please comment on the basic capabilities of the Portal in terms of enabling science, specifically the plan and schedule for implementing the capabilities, and the priorities given to them.

R: Up until now, after a research paper is published, it has been nearly impossible to reproduce results in any systematic fashion. One of the important tasks that can and should be undertaken by the Portal is recording and tracking the provenance of files, such that, for example, the underlying data upon which a published result is based can be recovered later. Accomplishing this single task would be a significant contribution of the Portal to DES.

Testing the Data Challenges against the requirements is an excellent use of the portal. The narrowness of the stellar locus in color-color plots is a possible substitute for a quality assurance test which now looks at a truth table.

The Portal allows products from different SWG's to be easily accessed by each other. For example, the Trilegal AddStar catalogs are of interest to the QSO WG, the WL WG, and for quality assurance in the context of checks of star/galaxy separation.

It is important to be able to advertise concrete accomplishments; the development of the AddQSOs module is an example.

The Portal also provides a good environment to facilitate various algorithm and code comparison projects of interest to the SWGs. In particular the Clusters WG has successfully conducted a comparison of cluster finders, the Photo-z WG is using the Portal as part of its DC6 photo-z challenge, and star/galaxy separation may serve as another comparison project whose results will be of utility to many SWGs.

R: It is our understanding that some of the SWGs intend to use the portal to evaluate DC6B. Establishing a stable production platform and loading the DC6B data should be high priority.

R: We recommend that the suite of tasks currently being addressed by the Brazil IT team be reviewed to determine if there are things that can be dropped from the portfolio (or at least deferred) to tighten the focus and to maximize the impact of the available resources.

Charge Element #2

Based on the written material and the presentations, please comment on the merits of the technical approach, namely the infrastructure being put into place, the management plan and schedule, and the operations plan. Within the constraints of what has already been built, what could be improved?

R: The team will soon need to transition from R&D to the very different environment of production operations. Several things are now "95% done" and these should be brought to completion.

Besides contributing to the definition of the requirements, the SWG's are involved in many other ways besides being the principal users, for example they are invited to contribute to the development, installation, and testing of analysis tools. It is important to understand the likely number of users of the Portal in the various categories in order to design accordingly.

Concerning processing power for science analysis, the model is that the Portal does what it can with its local machines, and the architecture allows for additional processing to be done remotely. This capability has not yet been implemented, but remote processing has been solved by the Open Science Grid and expertise to do this exists at Fermilab. By tapping in to experience at Fermilab, the timescale for implementing remote processing could be accelerated.

Since authentication is a project-wide problem, we need to solve it as a group. We should adopt an existing tool if possible.

The Portal includes lots of web interfaces to services with pull-down menus, ways of entering parameters, etc. It is not clear how flexible or brittle these interfaces are - e.g., what is involved in adding a new SWG?

Interest was expressed in having the opportunity to run codes stand-alone on a small scale ("lap top") locally in order to speed up the development cycle, for example local analysis with subsets of code and/or data extracted from the Portal. This would enhance the effectiveness of remote scientists' time, but does create additional demands, for example the need to sync up or insert results back into the Portal environment. This raises the question of what can be provided to remote developers in this respect, and what level of support it would require.

What is the capability of PostgreSQL to handle many users and many queries?

Not all scientific codes or applications can be run in the Brazil portal, not even in remote orchestration mode.

The ease and reliability of access to documents on the Portal (twiki.linea.gov.br/bin/view/LINEA/Scope and http://testing.des-brazil.org/).should be comparable to docDB.

Charge Element #3

More specifically, we are requesting ideas concerning the repositories for code and for value-added catalogs. How can these facilities be made more visible, inviting, and useful?

The idea of requiring pipeline developers to use the Portal as a managed code repository is noble but will require buy-in from the relevant developers and perhaps from the DES Collaboration as a whole, and it may conflict with developers' uses of their own individual repositories. It was pointed out that CMS meets this requirement by permitting only results produced by managed code to be accepted for publication, but DES is not at that stage.

The code repository may provide a first step in using the Portal for some of the SWG's. Making this easy is important in order to maintain interest in the other capabilities of the Portal.

Concerning value-added catalogs, each WG needs to propose which would be useful to access via the Portal. (Note that the Portal actually generates VAC's, e.g. random catalogs for LSS, and the Trilegal code for AddStars.) Validation of a SWG-supplied VAC could be done via a second pipeline that runs within the Portal.

Charge Element #4

By design, the work of the Brazil and NCSA groups are complementary. Are there new interfaces or developments that could clearly enhance the sum of these two parts?

The Portal provides a way to demonstrate that the outputs of DES DM are meeting the Science Requirements, such that all in the Collaboration can see the status of the tests. New plots of quantities of interest are easy to add (but this raises the issue of change control).

R: It might be useful for the Portal to set up a small standard testbed of images/raw data which could be repeatedly run through the NCSA piplines as part of development and quality control. It could be combined with a facility to add synthetic objects to real DES images to test the pipelines' ability to recover objects (as already done for the SN pipeline).

Both DES DM and Brazil are requesting help from the SWG's. While this requested help is of somewhat different character, there is a potential for competition for the same resources. The DES Project (and the SC in particular) should be aware of this.

R: It will be productive to encourage meetings between the NCSA group and the Brazil team to discover synergies and to plan work that is complementary. In particular, a visit by Patricia Egeland to Fermilab would be welcome and would allow her to learn about the Fermilab group's experience with grid-based tools. It would also be useful for Don Petravick to travel to Rio de Janeiro for interactions.

The DES Project should define its expectations for the role of tertiary sties, such as the Brazil site, in terms of distribution of files from DES DM, the databases, and analysis computing.

Charge Element #5

Please comment on the data processing efforts, namely the Quick Reduce development and the PreCam pipeline: status and any suggested future directions.

R: The trial by potential users in February 2011 during the Mock Observing run was very productive and the QR developers have taken that experience seriously. Before the design progresses too much farther, it would be useful to have another such trial run, i.e. an opportunity for members of the Collaboration with observing experience to evaluate and propose changes. For example, it would be good to see how time series information can be visualized. QR is an important part of the DECam system and it deserves a commensurate level of attention.

QR can in principle provide more informative QA than is currently provided by SISPI's Quick Look facility. One can think of SISPI as providing on-line diagnostics that are useful for mountaintop quality monitoring, and QR is the corresponding off-line facility. The only outputs of QR that are archived in the SISPI database are alarms.

Some benchmarking tests may be needed as the QR processing time presented (62 CCDs in 20 min with 4 cores) seems a factor of a few slower than expected, based on experience on simulated images with basic detrending using IRAF and photometry using SExtractor.

R: the requirements for QR (features and performance) need to be worked out cooperatively with the DES Project and written down. It should be clear what QR provides that is not duplicated by monitoring undertaken by SISPI. A plan should be created for this additional development and testing.

R: Processing the first year of PreCam data from scratch stresses many of the key elements of the Portal system, but it also requires many resources and is largely redundant with efforts at Fermilab. The main utility of a PreCam reduction pipeline at Brazil would come from a) a need to re-reduce the data for some reason or b) the acquisition of new data. Whether the former is needed or the latter will occur is unclear, hence we recommend that this effort be wrapped up.