Report of the Review Panel of the DES-Brazil Science Portal

29 October 2010

Background

A review of the DES-Brazil Science Portal development effort and status was held at Fermilab on 18 October 2010. The review was organized on behalf of the Dark Energy Survey (DES) Project Office. Members of the DES Collaboration who were at Fermilab for the Collaboration meeting were invited to attend, and about 15 people did so (in addition to the presenters and the reviewers). The agenda (Appendix A) included presentations by both the development team (technical aspects and design) and by DES scientists (their experiences so far). The purpose of the review was described in the Charge (Appendix B) to the review team, namely to evaluate how the portal has already helped the DES Science Working Groups (SWG's), and how its features can be enhanced to serve the SWG's in the future, when on-sky data will be available. The review team consisted of Rich Kron (chair), Brian Yanny, Huan Lin, Steve Kent, Bhuv Jain, and Enrique Gaztanaga. The report is organized as follows. An Executive Summary contains the main conclusions. Next we answer the elements of the Charge, including Findings, Comments, and Recommendations. At the end we include additional findings and comments under the topics of User interface and documentation, Infrastructure, and Staffing.

Executive Summary

The reviewers were impressed with the current functionality of the portal and were pleased to hear the very positive reports by the SWG users. There is significant potential by the portal to help the SWG's undertake science analysis efficiently in ways that would stress conventional methods (e.g. cooperative work on large data sets with a broadly distributed group of researchers). It is clear that the SWG's could benefit from the services that the Portal promises to deliver, e.g. running parallel analysis codes on the data and comparing outputs. The next step is to continue to encourage the

SWG's to learn about the portal, experiment with its capabilities, provide feedback on its performance, and suggest new features. A prerequisite to do this is to provide a quick-start guide and to streamline the development environment (e.g. fewer steps to check in code).

The system is actually much more than a Science Portal, it is a comprehensive, web-based eScience Analysis Center for the Dark Energy Survey Collaboration. It is designed to be very general, and allow the sharing of code, data, and results of analysis world-wide, in a virtual environment. We don't know that anything quite like it exists yet - it is forward-looking and ambitious.

"Bring the analysis to the data" when the data set is large is a paradigm that has existed for over a decade and has been attempted by a number of large projects (e.g., Griphyn, iVDGL, NVO) with varying and often limited degrees of success. The limiting factors include trying to be all things to all people, lack of a coherent system design, and obsession with technology rather than a focus on the needs of users.

In some ways, Science Portal is trying to be everything to everybody, which raises concerns about scope and focus. As described below, defining a realistic set of goals, requirements, and technical specifications is important and should be done in concert with the SWG's. We suggest this should be accomplished by July 2011. It would be helpful to be more formal in terms of what is to be delivered to the DES Collaboration (e.g. a list with milestones), and criteria or metrics of success should be devised to help later evaluation. Establishing a clear set of priorities for development work is not simple because it requires an ongoing dialog between the development team and the SWG's. The SWG's can propose features that will be effective, but the prioritization depends on the skills and other resources that are available to the DES-Brazil team.

It would be helpful to have another review at a later date, perhaps including reviewers external to DES. Progress on the recommendations given here can be one topic of that next review.

Response to the Charge

Charge Element #1

1a) Do the stated goals of the development effort ... actually address the needs of the DES Science Working Groups? Are there goals that should be added?

The overall goals and purpose of the Science Portal were not presented during the review, nor were the actual needs of the SWG's presented in a systematic way. From the background document:

"The main design goal of the portal has been to have an integrated environment from where the user could have:

1. access to relevant information about the project;
2. monitor the progress of the survey;
3. reduce data (QR, Precam);
4. produce and validate catalogs;
5. perform scientific analyses;
6. manage, export, share, the resulting science products."

This collective goal is more than what is needed by the SWG's. The first two are vague and seem to overlap with what may be done by DES DM. (The project will benefit from having an explicit statement differentiating its goals from those of DES DM regarding access to Data Challenges and future data releases.) Reducing PreCam data (goal 3) was not discussed during the review, nor was the envisioned role of the Portal in Quality Assurance. The PreCam reduction is a different flavor of pipeline and may have a different set of concerns. The team should make clear what the Science Portal will not do - there are specialized applications which will not or cannot be supported.

Do not add additional goals. The group already understands the conflict between the desire to add new features versus stabilizing and creating a production system. This conflict becomes clarified when one has a schedule for delivering a product, but we did not see the schedule.

(1b) Which goals have been met?

It is too early to say that any single goal from the list above has been met.

Comments

The role as software repository should be discussed more widely within the SWG's. If key projects were defined by the Science Committee for which the relevant software were to be made available, it could have implications for the goals of the Portal. A clear statement about the functionality and support for the software repository would be valuable.

Recommendations

1. Clarify, in some short document, what the overall purpose and goal of the Science Portal is (and perhaps what it is not), making reference to what the DES DM system will provide and how the Science Portal goes beyond that.

2. Also, it would be good to have a clear statement of goals for the scale of the Portal (e.g. disk space, number of cyles, number of users, number/stability of support staff). Can the Portal be scaled in the future as more users appear? Is PostgresSQL good enough for the database and the number of users and jobs run? How are production needs going to be orchestrated with an anticipated heavy future load? How will priorities be established for using resources? It might help to have a set of standardized processes that can be used as benchmarks of performance and measures of scalability.

Charge Element #2

Have the Science Working Groups been using the portal productively? Based on experience so far with DC5, what features or conveniences could be added to make the portal more useful? In addition, please assess communications between the Science Working Groups and the DES-Brazil development team and suggest ways these communications could be improved.

The main areas where users felt the portal could be improved included:

a) A more streamlined process for gaining access to the portal and learning how to use it. Right now the learning curve is steep and only dedicated users are contributing.

b) Faster time to publish code. For testers, rapid turnaround is essential.

It is important not to create unnecessary barriers to users that frustrate them in getting work done. E.g., entering parameters for a given job via a web form might be fine for running a job once, but if there is a need to run it multiple times from a scripted environment, then it can be counterproductive. A user would never do such a thing if running on their own local machine.

Findings:

- SWG projects under development on the Portal include:
  - Cluster finder comparison and industrialization (Clusters WG)
  - "ArborZ" photo-z code (Photo-z WG)
  - Galaxy SED fitting code (Galaxy Evolution WG)
  - Angular correlation function code (LSS WG)

- However, up to now none of the SWG's have used the Portal in a production mode to get results from tests of the current Data Challenge outputs.

- A ticket system is in place to facilitate bug reporting by developers

- Implementation of SWG codes relies on involvement of both SWG code developers and Portal team members (wrappers, bug fixes)

- Detailed documentation for usage of the portal is available

Comments:

- Overall impression is that SWG code developers have been able to effectively use the Portal for DC5 science projects, albeit not yet in a production mode

- Communication between SWG developers and Portal team members appears to be good, as SWG developers universally commented positively on responsiveness of Portal team

- A frequent comment from SWG developers is that Portal documentation is detailed and thorough, but not so well suited for an initial quick start

- An impression, in particular from the detailed description of the Cluster WG experience with the Portal, is that the development process basically works, but there are a number of things that can be streamlined (e.g., ticketing system login, more web services for common tasks) in order to speed up or simplify the code development process

- Some consolidation of Portal expertise and experience from different SWG's may ease Portal learning curve and facilitate broader use of the Portal among the SWGs

Recommendations:

3. To get more involvement by the SWG's, it will greatly help (and perhaps be necessary) to write a quick-start guide to code development on the Portal, and also to streamline the process of code development and testing in the Portal environment.

4. We would like to suggest a process for enhancing the engagement of the SWG's once the previous recommendation has been addressed. The first step will happen inevitably: the SWG's continue to develop and refine codes for their science analysis. Given that, the SWG's are requested by the Science Committee to nominate at least one appropriate code to be implemented inside the portal. This code could provide a practical example to test e.g. performance within the portal environment and scalability to larger data sets or larger number of users. The actual implementation would in general require attention from both the respective SWG and the portal development team, and the process thus needs to be cooperative. Meanwhile the SWG is presumed to be running their codes also externally to the portal. This provides an opportunity to compare results of the same code running in the two different environments. The benefits of this process are:
1) increasing the visibility of the Portal without placing significant new demands on the SWG's; 2) learning what the SWG's actually need;
3) learning what infrastructure is still needed to streamline adding new codes to the Portal (and running them and analyzing their outputs); and
4) providing a more systematic way to evaluate performance and estimate scalings.

Charge Element #3

In the longer term, in what ways could the portal better support science analysis from future data challenges, and from actual DECam observations? How can the Collaboration's use of the portal be increased (number of people routinely using it, and number of things they do with it)? Besides science analysis, are there other DES needs (e.g. operations) that the portal can support?

It is important to engage the SWGs as much as possible to allow them to run analyses on the portal. They are clearly interested. Ideally each SWG would have a member from the Brazil DES group, although that may not be practical. It is also important that the SWG's be brought in sooner, rather than later, because otherwise they will develop their own mechanisms for accessing and processing data and the Portal will have lost their attention.

It may be worthwhile to establish some milestones with respect to one or more SWGs, for example demonstrating the running of five cluster-finding codes on the same input dataset by some agreed-on date. These milestones depend on the SWGs actually delivering working code but would serve to encourage collaboration.

In the longer term, it is unclear what the deliverables will be: is it a portal infrastructure as a product that someone else will run, or is it an operations center? The deliverables need to be better defined.

Findings:

- Portal team estimates that, with a data release procedure in place between NCSA and tertiary sites (like Brazil), it would require:
  - 5 hours to transfer (from NCSA) and ingest 200 deg^2 of DC5B data
  - 5 days to just transfer (from NCSA) 5000 deg^2 of coadd data

Comments:

- Overall positive, though somewhat limited, experience with DC5 suggests expansion of SWG projects for DC6; in particular, the various SWG DC6B acceptance tests of DES DM outputs provides a good opportunity for implementation in the Portal

- There is the possibility of having the Portal installed at different tertiary sites to balance out processing resoures across the collaboration. However, there is then a consequent tradeoff between the effort needed to maintain/install the Portal vs. the return to the scientists locally

- There is a question of the scalability of the response and turnaround time from Portal team members if there are many SWG developers in the future vs. the current situation with just a few team members.

Recommendations:

5. A protocol for bulk transfers of data to the tertiary sites, preferably automatically, needs to be developed in coordination with DES DM. This protocol could be implemented for DC5 and/or DC6 data sets as a test.

6. Work with the SWG's to develop plans for implementation in the Portal of acceptance tests and/or other projects for the DC6B data

7. Explore the scope of the task of installing and maintaining the Portal at other tertiary sites.

Additional findings and comments

User interface and documentation

- Users want to push out changes to their code without having to depend on human intervention from the Brazil team to proceed to the next iteration in the code development cycle. Need to allow developer-level users to do most things without human intervention to eliminate the delays in step-by-step development of a code.

- Need a simple worked example on how to install a code and push it and run it, and how to make a change.

- Rather than entering command-line or tunable parameters into a web page, perhaps have a way to allow save-able short ASCII text files to record input and command line parameters for running pipelines.

- Should be clear to users what computing resources are available: CPU cores/Disk Space/core memory available per node.

- If XML is used, give a step-by-step guide to using it, with a worked example.

- Set up standards for code contributed by the SWG's to be checked in and documented.

- Users have a concern that placing their code into a repository will make it broadly accessible. It was good to hear assurances that users can control who runs what code, and who has access to the source files.

- List what compilers (gcc, f77, etc) are allowed, and which libraries (standard Unix C libraries, any astronomical libraries like slalib, etc.) are allowed for running on the Brazil site. Also what are the allowed input and output formats for input catalogs, input parameter files, and output (FITS, ASCII). Support a small number of standard formats.

- List which databases are supported (Simulated Galaxy Catalog, Star Catalog, SDSS/2MASS/USNO-B/FIRST/ROSAT, etc), plus which bits of DES data catalogs will be available (DC5, DC6, early commissioning data

- It will also be helpful to users to know what images are available on what turnaround time after they come from the mountain.   The turnaround time (how fast to populate the tertiary database with processed images and catalogs) should be specified in a requirements document.

Infrastructure

- 'tawala' appears to supply the need for putting together a set of pipelines, each grouped as a product with their own versions and parameter settings.

- Project has almost completely avoided proprietary software (IDL and maybe astronomical library like slalib are the only known exceptions). This is an important accomplishment and greatly expands the usability of the system. How are IDL and slalib licences handled? Important to stick to non-proprietary compilers (gcc) and libraries and databases (postgres is fine) whenever possible. We support the open-source solution chosen, even for the database (with multiple copies as needed).

- The investigation of efficient spatial indicies (RA/DEC) and the running of multiple copies of a back-end database transparently are important areas of research.

- NVO has invested a lot of effort in developing visualization tools with varying degrees of success (e.g. Aladin). Is the Portal in danger of repeating this effort?

- Have the developers looked at how other infrastructures work (or don't work)? E.g., Open Science Grid has much infrastructure for distributed exectuables. Some is already being used at Fermilab (condor_g, globus) but the authentication stuff is not. The ssh-key methodology for identifying users would not be allowed at Fermilab, which relies on DOE-Grid certificates.

- Product management (e.g., versioning, dependency tracking) was not discussed in detail. Fermilab developed a system called UPS to handle this (and other projects such as BOSS and LSST are using a variant of it). Is there any need seen for such a product?

- It seems that all code gets compiled on the Portal machines, which alleviates some OS mismatches. Could changes in the operating system environment ever become a maintenance issue?

- Error handling was not discussed much but it can be a roadblock. The main issues include providing feedback on the causes of errors, e.g. resource exhaustion and job failures on remote nodes. (This is a problem for existing grid operations - jobs are launched into the void and disappear without a trace.)

Staffing

- The development team is to be commended for their willingness to help users get their code working within the Portal. The existence of this manpower needs to be advertised more aggressively to the SWG's. Help-desk support is hard (expensive) to keep up for the long term.

- Would like to see a work-breakdown structure with employee numbers which shows how many people on help-desk vs. development vs. operations support through early operations of DES.

- The positions of Project Manager, Systems Administrator, and Database Administrator are all open, suggesting a serious issue with manpower. However, a search is being conducted for all of these positions.

- The job of fetching/staging data from a Primary archive was discussed in one talk and it is clear that it could grow into a big issue (DC5b is only a small test). How much effort will be involved? Will this effort conflict with other planned activities?

Appendix A - Agenda

October 18, 2010

1:00 - 1:30 - general introduction (L. da Costa)

1:30 - 1:50 - Science Portal overview (B. Rossetto)

1:50 - 2:50 - SWG experiences (10 min each)

Donnacha Kirk - Weak Lensing
Claudia Maraston - galaxy SED analysis codes
Brian Gerke - Clusters
Wayne Barkhouse - Clusters
Fernando de Simoni - Large-Scale Structure
Beatriz Ramos - Photo-z and Galaxy Evolution

2:50 - 3:00 - break

3:00 - 3:20 - Science Portal design - L. Martelli

3:20 - 3:35 - handling data on the Science Portal - A. Fausti

3:35 - 4:30 - general discussion of features and priorities for development

4:30 - 5:00 - executive session: panel formulates recommendations

5:00 - 5:30 - close-out: recommendations are presented to development team

Appendix B - Charge

Charge to DES-Brazil Science Portal Review PanelA document (to be provided) outlines the goals of the DES-Brazil Science Portal in terms of serving the needs of the Science Working Groups. This document will be a starting point for the review. Specifically,

1a) Do the stated goals of the development effort, namely the functionality of the portal, actually address the needs of the DES Science Working Groups? Are there goals that should be added?

1b) Which goals have been met?

1c) For those goals that have not yet been met, is there a sound plan in place to complete the development work and are the priorities appropriate? Please contribute any practical advice concerning implementation of the work plan.

2) Have the Science Working Groups been using the portal productively? Based on experience so far with DC5, what features or conveniences could be added to make the portal more useful? In addition, please assess communications between the Science Working Groups and the DES-Brazil development team and suggest ways these communications could be improved.

3) In the longer term, in what ways could the portal better support science analysis from future data challenges, and from actual DECam observations? How can the Collaboration's use of the portal be increased (number of people routinely using it, and number of things they do with it)? Besides science analysis, are there other DES needs (e.g. operations) that the portal can support?