

# FY12 Science Analysis Project for LSST Data Management at Fermilab (updated Dec. 8, 2011)

## 1. Introduction

The LSST Data Management (DM) group has presented a proposed list of tasks for Fermilab participation in LSST beginning in FY12. The purpose of this document is to outline the vision and goals of this work and to discuss the benefits to LSST and Fermilab. LSST DM would like to have agreement on the work plan as soon as possible. This work plan will cover one year.

## 2. Vision

The vision is to establish an LSST presence at Fermilab. Members of the LSST community at Fermilab will gain experience in using LSST software and will develop the ability to run science analyses on Grid computing resources. By joining the LSST Data Management effort now, Fermilab will be able to help define Dark Energy (DE) science goals and contribute to the specification and development of data processing tools and capabilities.

## 3. Goals

This section provides a brief summary of goals for FY12 and FY13 for Fermilab participation in LSST DM. A more detailed discussion of the individual goals is presented in the “Proposal” section of this document. The goals are the following:

1. Gain understanding of the LSST software by first porting a subset of production applications to the Open Science Grid (OSG).
2. Define and establish a science data processing environment to run DE science analyses at Fermilab.
3. Establish, operate, and maintain an LSST Virtual Organization (VO) VOMS and submit host at Fermilab.
4. Collaborate with LSST and DES scientists to define and develop 1, 2, or 3 science analyses as “science drivers.” These science analyses will be used to help develop requirements for the Level 3 Science Analysis Toolkit, an important part of the data analysis environment that LSST scientists will use to do their science.
5. Take advantage of Fermilab expertise to run science analysis applications on OSG.

## 4. Stakeholders

The stakeholders included in this section represent the groups that have an interest in Fermilab participation in LSST Data Management.

- LSST Data Management (DM) Management
- DOE’s LSST Project Office (SLAC)
- Fermilab Management (e.g. Director, Associate Directors, Sector Heads)
- Fermilab Center for Particle Astrophysics (FCPA) Management
- Scientific Computing Division (SCD) Management

- Experimental Astrophysics Group (EAG)
- Computing Enabling Technologies (CET) Group
- Grid and Cloud Computing Department
- LSST Group at Fermilab

## 5. Benefits

This section describes benefits to LSST and Fermilab to begin collaborating on LSST DM for fiscal year FY12. The benefits are specific to the above-mentioned goals, which were developed during a one-day meeting at Fermilab on May 12, 2011. Participants at the meeting included Jeff Kantor (Project Manager for DM), LSST institutions working on DM (IPAC, NCSA, SLAC), several groups from Fermilab's Computing Division (CET, EAG), and Fermilab personnel involved in OSG user support.

LSST DM sees the following benefits by including Fermilab as a collaborating institution. First and foremost, the LSST Project will be able to take advantage of Fermilab expertise to reduce risk in Data Management in preparation for NSF and DOE reviews. LSST wants to establish solid institutional relationships in preparation for these reviews. By getting involved in FY12, Fermilab will be able to develop the necessary expertise to contribute to DM development efforts during the construction phase.

Fermilab sees benefits in participating in LSST DM by beginning to establish a local LSST community at the lab, and by forming a partnership with OSG. OSG involvement can make data and software available to users within the DOE community.

## 6. Selected Background Material

LSST DM has less than 2 years remaining in their R&D phase. The main emphasis has been on software development for Level One (L1) and Level Two (L2) production data processing. The creation and management of Level Three (L3) science data products is a new effort, and this is the area that Fermilab will be involved with. A work plan that accomplishes the above goals will cover approximately one year of effort. This effort should begin as early as possible so that Fermilab's role is established in time for the DOE CD-2 Review and corresponding NSF review.

LSST has clearly stated some of the major processing constraints for LSST data and the production and support of L1 and L2 data products. It is important to note that these data products do not include the science. Science data products are produced at L3 and are not defined or driven by the LSST experiment. L3 products are the results of science groups interested in processing LSST data. There is no exclusive access to LSST data and access to data is not based on LSST membership. A planned 10% of LSST resources (storage and computing) are allocated for user-driven science analyses. Science groups will apply for these resources, and L3 science products will have the advantage of federation with L2 products. LSST expects that Grid and other institutional resources will provide a substantial, and perhaps even the bulk of the computing and storage resources for L3 data processing. Science groups will apply for use of these resources.

LSST has defined Data Access Centers (DAC) to handle a portion of the science data processing load, make L1 and L2 products available to the community, and store selected L3 products. There are requirements for becoming a DAC, which include participating in data redundancy demands for the experiment, and overall query load balancing.

The L1/L2 production open software frameworks are anticipated by the DM software groups to be very suitable for L3 data processing needs and that these frameworks will be used in combination with researcher-provided code. This option needs to be explored to understand how well the products work

in this environment. It will be necessary to carefully package the toolkits and provide a runtime environment that can be used in a small setting, such as a scientist's laptop or workstation.

DOE is committed to doing a survey and this will be done as a science group operating in L3. They are committed to providing resources in support of it. A DAC does not need to be formed to support a copy or partial copy of the LSST data necessary to do DE science.

Production processing for LSST is unlike that of the LHC tier 0-3 centers, partly because LSST has more dedicated roles for centers and partly because of their "open access" policy to all data. Within LSST, the archive and base centers are strictly for data-reduction processing to produce L1/L2 data products, and the DACs are for L3 analyses.

An important question to consider when it comes to Grid resource use is the I/O to processing ratio. What is the overall time used in moving data versus processing it for different classes of jobs?

IPAC is assigned to work on the science user interface. They have a draft of scientist requirements for accessing LSST science-level data. At Fermilab we were shown a prototype of web interfaces for making data requests. It is likely that any tools or capabilities developed here at Fermilab would be candidates for inclusion or integration with the IPAC user workspace toolkit.

## **7.Proposal**

### **7.1.OSG production port**

Using the June 2011 software release (PT1.2), take appropriate production pipelines and port them so that they are capable of running within OSG. This will provide a good introduction to LSST software in preparation for further L3 work. Working on pipelines that rely on co-adding images would be a good first step, since LSST algorithms can be evaluated for use by DES.

One type of pipeline application runs on one CCD detector at a time, which is about 16MB of image data plus 30MB of calibration data. Current job packaging for processing on computing farms is 46MB total for one CCD. The output from a single job is < 200MB. Memory requirements are less than 2GB per process.

This work requires a developer who is already knowledgeable with porting applications to OSG. It also requires a physics-developer who can understand and quickly become familiar with the LSST pipeline applications. Most of this work can proceed without dependence on any of the other goals.

### **7.2.Science data processing environment**

The LSST software environment needs to be established at Fermilab. This includes product packaging, management, maintenance, and installation/deployment available through facilities such as UPS.

Documentation needs to be written to explain use of the software on general-purpose resources available at Fermilab. Expertise needs to be established at Fermilab by communicating with LSST core infrastructure developers so that Fermilab developers can have a point of contact for asking questions about setup and use of LSST software. Jeff Kantor points out that this includes such things as APIs, an operational model, user tools, and resource management.

This work could be the basis for the initial core software setup for movement into L3 science processing.

This work requires system administrator and computer system specialists to establish and manage the environment. It requires a physics-developer to act as liaison between LSST software groups and future

users at Fermilab.

### **7.3.LSST VO**

An LSST VO needs to be established at Fermilab using available Fermilab resources. The needs of LSST in this area should be small over the one-year period covered by this work plan. Fermilab will serve a similar purpose within DOE for OSG as NCSA did within NSF for TeraGrid. This facility would enable DOE scientists to take advantage of resources and usage policies that they are already familiar with.

The detailed costs and security issues associated with using Fermilab on-site resources for this will need to be understood.

The VO will need to be established with the help of the Grid and Cloud Computing Department. It will need to be maintained by administrators or computing system specialists assigned to the LSST group.

### **7.4.Science use-case development**

Fermilab will work to find DE scientists in the area of gravitational weak lensing, supernovae, and/or galaxy clusters. The work involves getting a firm commitment to work with the scientists to define completely 1, 2, or 3 L3 use cases and identify the processing tasks to carry them out. The use cases will help drive future development of the L3 science analysis toolkit. The purpose of this work is also to define the L3 science data products that are needed for these analyses and to learn how they products can be stored, manipulated, queried, and perhaps even federated with L1/L2 production-level data. Many of the Fermilab contacts are currently working on DES science needs, and many of them are also part of LSST or will become part of LSST science groups. We suggest focusing on DES science analyses using LSST software infrastructure and available tools.

During the use case development, we will need to classify the problems as catalog or image processing tasks, and if they are random access database queries or sequential access database queries. These use cases will need to be documented and reviewed by relevant LSST science collaborations.

This task requires a physics-developer and a software infrastructure developer/analyst so that both science needs and software design needs can be addressed.

### **7.5.Run science applications**

The L3 applications identified as part of use-case development can be integrated with or developed using the existing production software toolkit. This exercise will demonstrate the suitability (or lack of suitability) of these toolkits to solve the scientists' problems. This experience will help determine what the L3 science analysis toolkit should look like and what additional pieces might be necessary.

This work requires a physics developer and a software developer/analyst. The work also requires that some of the work on use cases be complete, and that applications that can implement a use case be identified. The science data processing environment will also need to be available.