Laboratório Interinstitucional de e-Astronomia: passado, presente, e futuro

Luiz A. Nicolaci da Costa[1,2], Angelo Fausti Neto[2,3], Marcio A. G. Maia[1,2], Ricardo L. C. Ogando[1,2], Riccardo Campisano[2,4]

[1]Observatório Nacional - MCTI

[2]Laboratório Interinstitucional de e-Astronomia – LIneA - MCTI

[3]LSST Corporation - USA

[4]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ

{ldacosta,ogando,maia,angelofausti,riccardo.campisano}@linea.gov.br

Abstract. *We describe how LIneA was created, what it has produced in terms of software and hardware infrastructure in the last 10 years of work on Dark Energy Survey and in preparation to many other projects, such as LSST, when an e-science center in Brazil will be needed to handle huge data volume, velocity, and variability. Parallelism, provenance, and visualization are some of the challenges faced at LIneA in order to generate products and achieve scientific results.*

Resumo. *Descrevemos como o LIneA foi criado, e o que produziu em termos de software e hardware nos últimos 10 anos de trabalho no Dark Energy Survey e em preparação para vários outros levantamentos, tal como o LSST, quando um centro de e-ciência no Brasil será necessário para atender seu grande volume, velocidade, e variabilidade de dados. Paralelismo, proveniência, e visualização são alguns dos desafios encarados pelo LIneA a fim de gerar produtos e alcançar os resultados científicos.*

1. Introdução

O desafio de poder atuar no levantamento *Dark Energy Survey* (DES, Flaugher et al. 2015) motivou, há quase 10 anos, a formação de uma rede de pesquisa composta por pesquisadores de diversas instituições e a formação de um time de tecnologistas para dar suporte técnico ao desenvolvimento de ferramentas que facilitassem o manuseio de grandes volumes de dados. Estas necessidades levaram a proposta da criação de um projeto estruturante no Observatório Nacional (ON) conhecido como Astrosoft e finalmente a criação do Laboratório Interinstitucional de e-Astronomia (LIneA), um centro de e-Ciência voltado para a Astronomia. Este centro permite que pesquisadores brasileiros se engajem na ciência de grandes levantamentos astronômicos. A missão do LIneA é criar soluções de software e montar uma infra-estrutura física para lidar com o problema de *Big Data* gerado por projetos de *Big Science*. Por *Big Data* entenda-se projetos que envolvam a aquisição de dados que satisfaçam um dos seguintes quesitos: velocidade, volume ou variedade. Por *Big Science* entenda-se problemas de ciência fundamental na fronteira do conhecimento, os quais devido à complexidade inerente requerem projetos de *Big Data* e tipicamente grandes colaborações científicas, como por exemplo, o estudo da Energia Escura. A evolução do LIneA nos últimos 5 envolveu o desenvolvimento de um sofisticado portal científico que está em constante aprimoramento possuindo instâncias de operação em vários centros internacionais como Cerro Tololo Inter-american Observatory (CTIO), Fermilab, e National Center for Supercomputing Applications (NCSA). O laboratório estabeleceu colaborações técnicas

com centros como Lawrence Berkeley National Laboratory (LBL), Fermilab, NCSA, Large Synoptic Survey Telescope (LSST, Abell et al. 2009) e Stanford Linear Accelerator Center (SLAC). Expandiu suas metas promovendo a entrada de pesquisadores brasileiros em novos projetos internacionais como Sloan Digital Sky Survey IV (SDSS-IV), Dark Energy Spectroscopic Instrument (DESI, Eisenstein et al. 2015), e LSST em colaboração com os principais centros de pesquisa no mundo, apoiando pesquisa de ponta em astronomia. O LIneA também promove a formação de um novo tipo de profissional capaz de atuar como protagonista em grandes colaborações internacionais e capacitado para a nova ciência de dados.

2. Infraestrutura física

Para atender diferentes necessidades computacionais dos projetos apoiados pelo LIneA, um centro de armazenamento, processamento, análise e distribuição de dados foi construído ao longo dos anos. Atualmente o centro tem mais de 100 equipamentos que fazem parte de uma arquitetura desenhada para atender as principais demandas das aplicações científicas, tais como: altas taxas de transferência de dados; armazenamento de grandes quantidades de dados; consulta eficiente de banco de dados; processamento paralelo; hospedagem de serviços críticos para a operação e colaboração. Destacam-se os seguintes elementos: *Cluster* de processamento para produção (38 nós, 912 núcleos), *Cluster* de processamento para testes de integração (4 nós, 72 núcleos), *Cluster* Lustre (6 nós, sistema de arquivos compartilhado de alto desempenho), Mass Storage (armazenamento de dados), Banco de dados do SDSS (MS SQL), Banco de dados do DES (PostgresSQL), Cluster de serviços (VMs) onde se encontram as ferramentas colaborativas git, twiki, repositório de documentos, máquina dedicada ao desenvolvimento, e sistema de Transferência de Dados (DTS).

3. Desenvolvimento de software

O desenvolvimento do portal teve início em 2007 e seu desenho básico evoluiu muito desde então. Cerca de 30 profissionais de TI contribuíram para o seu desenvolvimento, sem contar alunos, pós-doutorandos e pesquisadores. Vários produtos já foram entregues e estão disponíveis para diferentes comunidades. Por exemplo o *Quick Reduce* (QR), o *Data Server*, Criação de Catálogos Científicos, e *Workflows* Científicos. O QR é um sistema para diagnóstico em "tempo real" das exposições da câmera do DES (DECam) instalada no telescópio Blanco de 4 metros no CTIO, Chile (Fausti Neto et al. 2013). *Data Server* é um sistema disponível no LIneA e no Fermilab que permite: acompanhar o progresso do levantamento DES, verificar visualmente a qualidade das imagens, permitindo marcar defeitos, rejeitá-las e fazer comentários, sobrepor objetos de catálogos (internos e externos ao DES) às imagens do levantamento, realizar buscas nos catálogos de objetos produzidos a partir destas imagens, criar ou carregar listas de objetos de interesse (ex. lentes fortes), visualizar estes objetos e criar "recortes de imagens" para referência futura ou publicação. A criação de catálogos para análises científicas específicas é talvez um dos maiores desafios de levantamentos fotométricos de grandes áreas do céu tendo em vista o volume de dados, o grande número de atributos associados a uma fonte (~900), o grande número de decisões que devem ser tomadas ao longo do caminho (ex. cortes em magnitudes, *flags*, e regiões rejeitadas) e a necessidade de preservar sua memória de produção. O último estágio do sistema de análise *end-to-end* (da Costa et al. 2013) é feito pelos *workflows* científicos que usam os catálogos preparados no estágio anterior para realizar diferentes análises científicas. Para controlar a execução de todos esses passo foi criado uma tela de *Dashboard* (Figura 1). O *Dashboard* consulta o banco administrativo que registra processos executados, seu status, e proveniência para um dado *release* do DES. Nessa tela central também se pode acessar e compartilhar os resultados de cada processo.

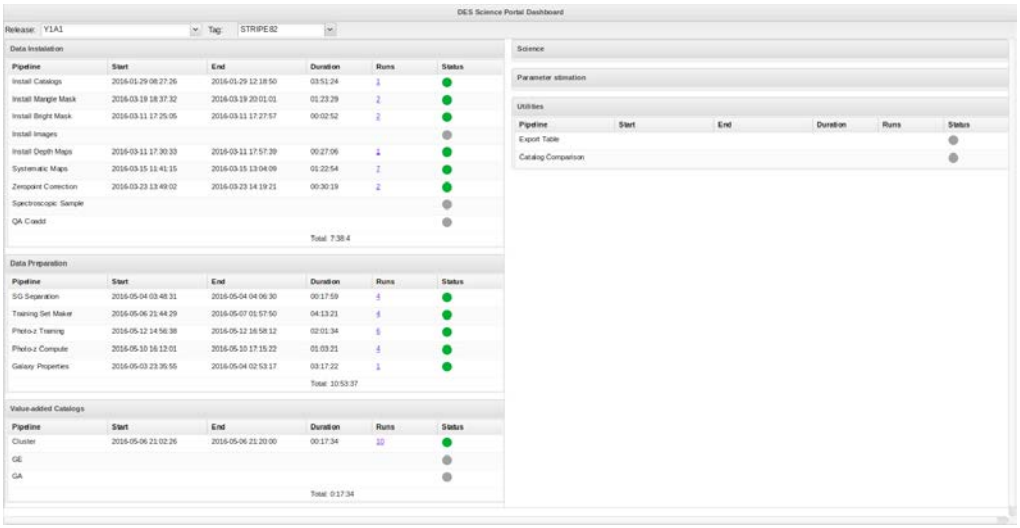

Figura 1. Dashboard usado para monitorar o número de execuções, início, fim, duração e status do último processo de cada pipeline executado no portal.

4. Resultados

Entre os resultados obtidos ao longo dos últimos cinco anos, destacamos a criação e a operação de um laboratório multi-usuário de médio porte para o processamento e distribuição de dados, dedicados aos grandes levantamentos astronômicos e integrados a vários centros internacionais. O laboratório tem mais de 100 usuários cadastrados, atendendo pesquisadores de universidades no Rio de Janeiro, São Paulo, Paraná e Rio Grande do Sul e algumas instituições no exterior.

O desenvolvimento de um complexo e abrangente sistema de software para atender de forma eficiente as necessidades de validação, análise e mineração do grande volume de dados envolvido. Estes sistemas estão em operação no observatório do CTIO no Chile atendendo a todos os usuários da DECam, no Fermilab com mais de 180 usuários registrados, no LIneA atendendo aos pesquisadores brasileiros e, em breve, no NCSA e em Berkeley. Como consequência dos termos acordados pelo LIneA com os vários projetos, o laboratório assumiu compromissos que se estendem pelos próximos cinco anos. Além disso, a experiência adquirida mostra o papel vital que o laboratório pode e deve ter na preparação da comunidade brasileira para a era do LSST. Além disso, temos a formação de pesquisadores produzindo até agora 12 teses de mestrado e seis de doutorado. Quatro pesquisadores jovens foram reconhecidos como “construtores” no projeto DES pela contribuição equivalente a 24 meses de trabalho dada para a infraestrutura do projeto, no caso o desenvolvimento de ferramentas de análise integradas ao portal científico desenvolvido pelo LIneA.

A produção de 95 publicações em revistas arbitradas. Com quase 20 artigos por ano e da ordem de 5.000 citações acumuladas, atestam a produtividade e impacto científico dos projetos.

5. Perspectivas

Do ponto de vista técnico, o LIneA desenvolveu, implantou e manteve dois complexos sub-sistemas do Portal que o validaram, consolidando a posição do laboratório como um parceiro confiável na entrega de produtos úteis e de qualidade à colaboração internacional do DES. Estes subsistemas já estão em uso pela comunidade internacional no CTIO e no Fermilab e, em breve, no NCSA. A qualidade do produto motivou, por exemplo, a colaboração do DESI em convidar o laboratório para participar do

desenvolvimento do *Quick Look Framework* para validar 15.000 espectros a serem obtidos durante cada exposição do instrumento. A assinatura do acordo com o LSST garante que o investimento feito pelo LIneA para o DES poderá ser estendido e melhorado, integrando novas tecnologias nos próximos 5 anos, dando a oportunidade única da equipe brasileira estar altamente preparada para capitalizar nos dados do LSST, já no período de verificação científica, previsto para começar em 2020-2021. Até lá, será possível usar o próprio DES para desenvolver os vários *workflows* científicos.

No momento, junto com o Laboratório Nacional de Astrofísica (LNA), uma grande prioridade é estabelecer o processo de montagem do *Brazilian Participation Group* do LSST, construir a infraestrutura de software, e montar a base do Centro Regional de dados do LSST – e se for o desejo da comunidade, trabalhar para estender o acordo para incluir toda a comunidade astronômica em um projeto que mudará o jogo da astronomia na próxima década.

References

Abell et al. (2009) “LSST Science Book, Version 2.0” ArXiv 0912.0201L

da Costa et al. (2013) “End-to-end scientific processing in the LIneA Science Portal” CSBC 2013 - BreSci - VII Brazilian e-Science workshop

Eisenstein et al. (2015) “The Dark Energy Spectroscopic Instrument (DESI): Science from the DESI Survey” AAS 225 336.05

Fausti Neto et al. (2013) “Quick Reduce: Dark Energy Survey Camera mountain-top Quality Assessment tool and its master calibration pipeline” CSBC 2013 - BreSci - VII Brazilian e-Science workshop

Flaugher, B. et al. (2015) “The Dark Energy Camera” AJ 150 150