NOAO Data Lab

Matthew J. Graham
NOAO/Caltech
Our current team

- Knut Olsen (Project Scientist)
- Mike Fitzpatrick (System Architect, Lead Developer)
- Matthew Graham (Developer, Scientist)
- Ken Mighell (Requirements Lead, Scientist)
- Betty Stobie (Project Manager)
- Pat Norris (Documentation and Testing Lead, Developer)
- Stephen Ridgway (Scientist)

but will also be hiring in next few months
Big Data @ NOAO
Big Data @ NOAO: the numbers

270 TB of imaging data currently from:
- Dark Energy Survey
- DECaLS and DESI Targeting Survey
- Community DECam programs and surveys

Hundreds of TBs more coming

Large catalogs coming:
- Dark Energy Survey – 45 TB
- DESI Targeting Survey - ~5 TB
- Community programs and surveys – up to several TBs each
DataLab enables

Catalog science:
- Search for Galactic substructure through photometric selection of candidate populations

Data exploration:
- Selection of a sample of large galaxies from a catalog, retrieving image cutouts, overlaying with catalog measurements

User-defined custom workflows:
- Use a large sample of galaxies to determine frequency of minor mergers, obtaining image cutouts and performing custom pixel analysis (e.g., PSF subtraction, image filtering, automated feature detection)

Collaborative research:
- 30 investigators in SMASH collaboration working on many aspects of the search for Magellanic Cloud populations all over the sky
An integrated VO-enabled framework supporting:

- (A)synchronous catalog access
- Virtual storage services (local/remote)
- Databases (local/remote)
- Image access tools
- Task automation tools
- Workflow composition and orchestration
- Variable resolution display tools
- Statistical analysis tools
DataLab targets

Science Users:
- **Experts** - Know SQL, scripted use of D/L, large data processing
- **Novice/Intermediate** - Exploratory science, web interfaces, use Data Lab for access with local tools
- **Collaborations** - Shared/remote access, data publishing, mixed experience

Application Developers:
- Implementing **new science** tools/capabilities
- **Automating** analysis workflows
- **Extending** the Data Lab framework for new projects
DataLab architecture

Astronomer’s Desktop
- Web Page
- Cmdline Tools
- User Code
- Legacy Apps

Data Lab Ops
- User Mgmt
- Monitoring

Public Services
- Authentication
- Job Manager
- Resource Resolver
- Query Manager
- Storage Mgr
- Public Repo

Private Services
- Ops Monitor
- Private Repo

Data Access Services
- SIA
- SSA
- SCS
- UWS
- VOSpace
- UWS
- TAP
- UWS
- SQL Service

Databases
- MyDB
- Large Cats
- Data Pub
- Ops DBs

Storage Resource
- User Space
- Virtual Space
- UWS
- Compute Jobs

Compute Resource
- UWS
- VO Data
- VO Svcs
- NSA

Presentation Layer
Services Layer
Data Access Layer
Resources Layer

15% Minimal Changes Needed
40% Existing Implementation
10% Ongoing Dev Required
35% Full Dev Required

IVOA Sesto - 16 June 2015
Technology highlights

• **VOSpace:**
  – Remote (Java) and local (Python) implementations
  – Remote can appear as local filesystem via FUSE layer
  – Capabilities are an integral part
    • launch arbitrary code when data is placed in a directory
    • use config files to turn capabilities on/off

• **Database:**
  – Distributed system based on QServ (LSST solution)

• **Authentication:**
  – Integrates with existing NOAO identity system

• **Computation:**
  – Docker containers provide lightweight task virtualization
  – Easy and shareable workflows
Schedule

• Phase I – Demo @ AAS June 2016:
  – User-ready virtual storage (FUSE)
  – Basic job control
  – Basic data query tools (web-based, MyDB)
  – Data services (SQL/TAP to SMASH, DECaLS and NSA)
  – Custom plotting tool

• Phase II – Summer 2017 release
  – Authentication
  – Management tools: users, storage, jobs, queries
  – Visualization framework
  – Large catalog support
  – Operations tools: system monitoring, help desk, user documentation
In the burgeoning era of data-intensive astronomy, the DataLab will enable and facilitate:

- Exploration of large data sets including interactive visualization
- Near-data processing/computation – only final products need retrieving
- Reuse of existing data sets (including those not in NSA) through federation for further data mining
- Collaborative work (particularly between domain experts)
- Development of scalable solutions for tera/petascale data
- No programming required so friendly to legacy code
- Primarily predicated on current and future NOAO data holdings