SPACE TELESCOPE SCIENCE INSTITUTE

STScI Science Cloud Evaluation

Exploring Astronomy Science Solutions in the Big Data Era

V. Acosta
10 Oct 2016

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

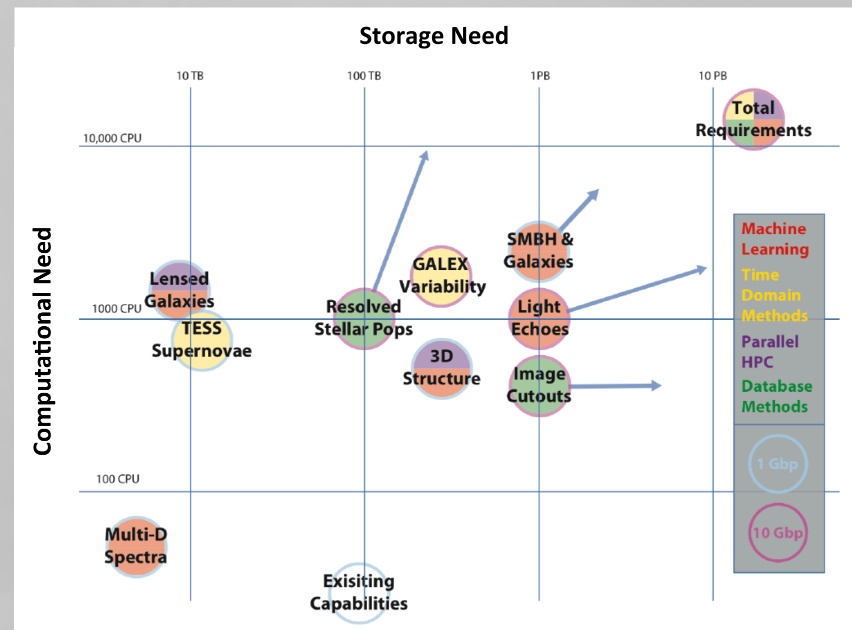
Contents

- ST Background
- ST Evaluation
  - ST SciServer Implementation
- ST Way Ahead
- Common Astronomer's Workspace
- Backup
  - Technologies Evaluated
  - Use Case: Globular Clusters Search around M87 (HSC & HLA query)
  - Big Data @ STScI Use Cases

Background

Big Data @ STScI (15 Mar 2016) recommended improved computing infrastructure and software tools

Computing Infrastructure	EOY 2016	EOY 2018	EOY 2021
Networking Bandwidth	External: • 1 Gbps Internal: • 10 Gbps	External: • 10 Gbps Internal: • 100 Gbps	TBD
Storage Capacity • Increase database server capacity for high volume query usage	1429 TB MAST (ops)	6 PB	30 – 60 PB
Computing Power • Strategic partnerships for HPC	~200 CPUs MAST (ops)	1000 CPU	10,000 CPU



ST Big Data Identified Needs

Software tools: (1) user environment, (2) data visualization tools, (3) Machine Learning architectures, (4) automated spectral feature extraction and classification software

http://newsletter.stsci.edu/big-data-strategy-at-stsci/ | http://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf

Evaluation

- In Jun 2016, we began a trade study to identify applicable solutions for ST

Evaluated in house

Funder	Organization	Science Cloud Solution	Summary	Technologies Summary
NSF	JHU IDIES	SciServer sciserver.org	Modular tools to search and process large (tera and peta byte) datasets allowing user defined queries, processing scripts, and sharable datasets	• Jupyter Notebooks, Docker • GitHub • OpenStack (KeyStone) • Microsoft SQL • Matlab, R, iPython, Python, Java, CLI • SkyServer supports VO: Cone Search, TAP+, SIAP
ESA	Parameter Space	Gaia Added Value Interface Platform (GAVIP) (portal).gavip.science	Platform that enables user-contributed code run next to Gaia archive (1-2 petabytes) and allow for reuse and sharing of this code; users provide AVIs	• Jupyter Notebooks, Docker • AVIs analysis pipeline • Anaconda 4, GitHub • REST-ful web pages; JavaScript/CSS • Python, Java, Fortran, CLI • Includes IVOA (TAP+), ADQL
AURA	NOAO	Data Lab datalab.noao.edu	Provide efficient exploration and analysis of large datasets generated by NOAO's wide-field telescopes	• Jupyter Notebooks, Docker • Python, but meant to be agnostic • Heavily IVOA architected: TAP+, SAMP, VO Space, SIA, SCS, SSA, UWS, SSO

- Remaining to explore: HERA (HEASARC), LSST, IPAC, Gemini, others?

SciServer

SciServer Altair 2+ Evaluation (Jul-Aug 2016)

Goals:

- ✓ Evaluate SciServer instance within STScI infrastructure
- ✓ Assess nominal MAST science use case
- ❖ Determine way ahead for ST science cloud solution

Use Cases Demonstrated:

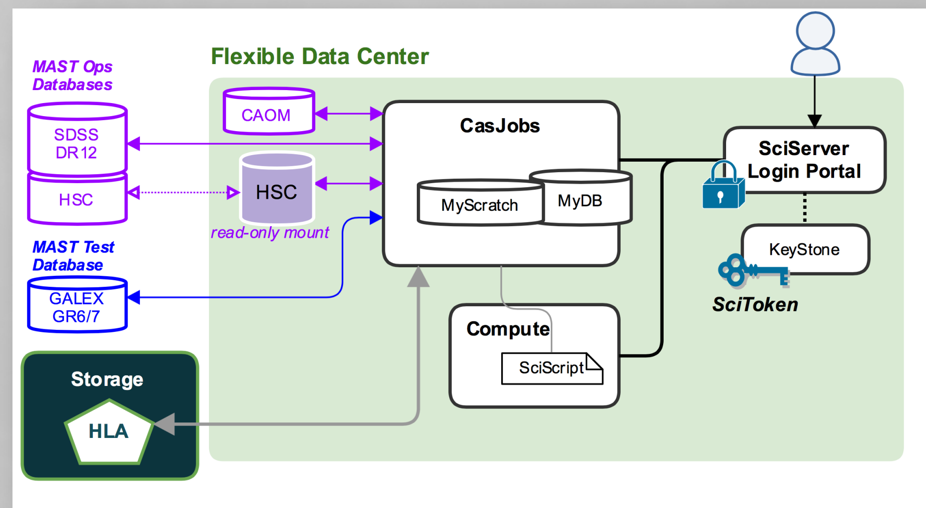
- ✓ Functional Use Case: CAOM access, GALEX, SDSS
- ✓ Science Use Case: Globular Clusters Search around M87

***SciServer Feedback:**

- Enterprise-level implementation, e.g.:
  - Administrative user rights
  - Installation guide
  - Documentation
- SSO Integration

ST Challenges Learned:

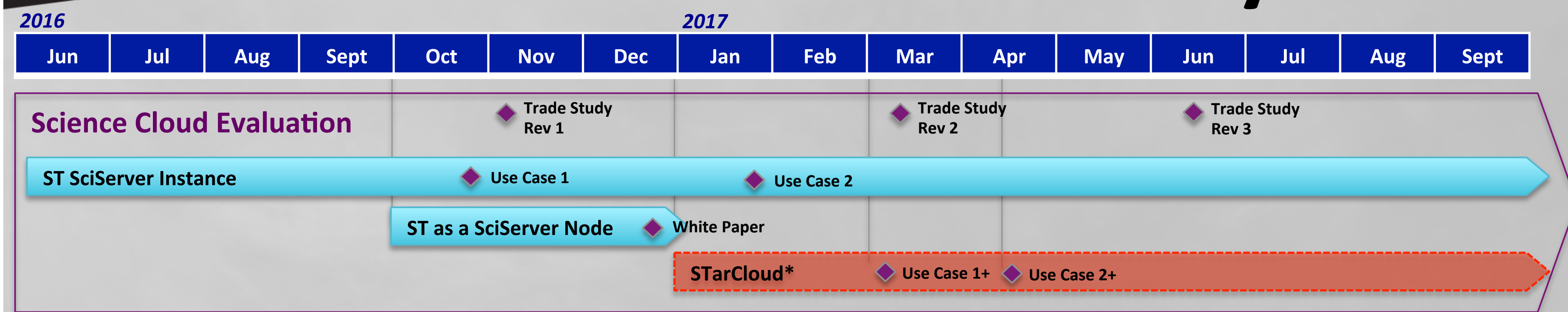
- Enterprise-ready technologies
- Not set up for distributed dev
- Invested in proprietary software, relational databases



Component	Application
CasJobs w. RESTful interface	JHU IDIES custom (MS SQL)
Compute	Jupyter Notebooks
Login Portal	OpenStack Keystone

**Current SciServer 2+ evaluated in house is still under development*

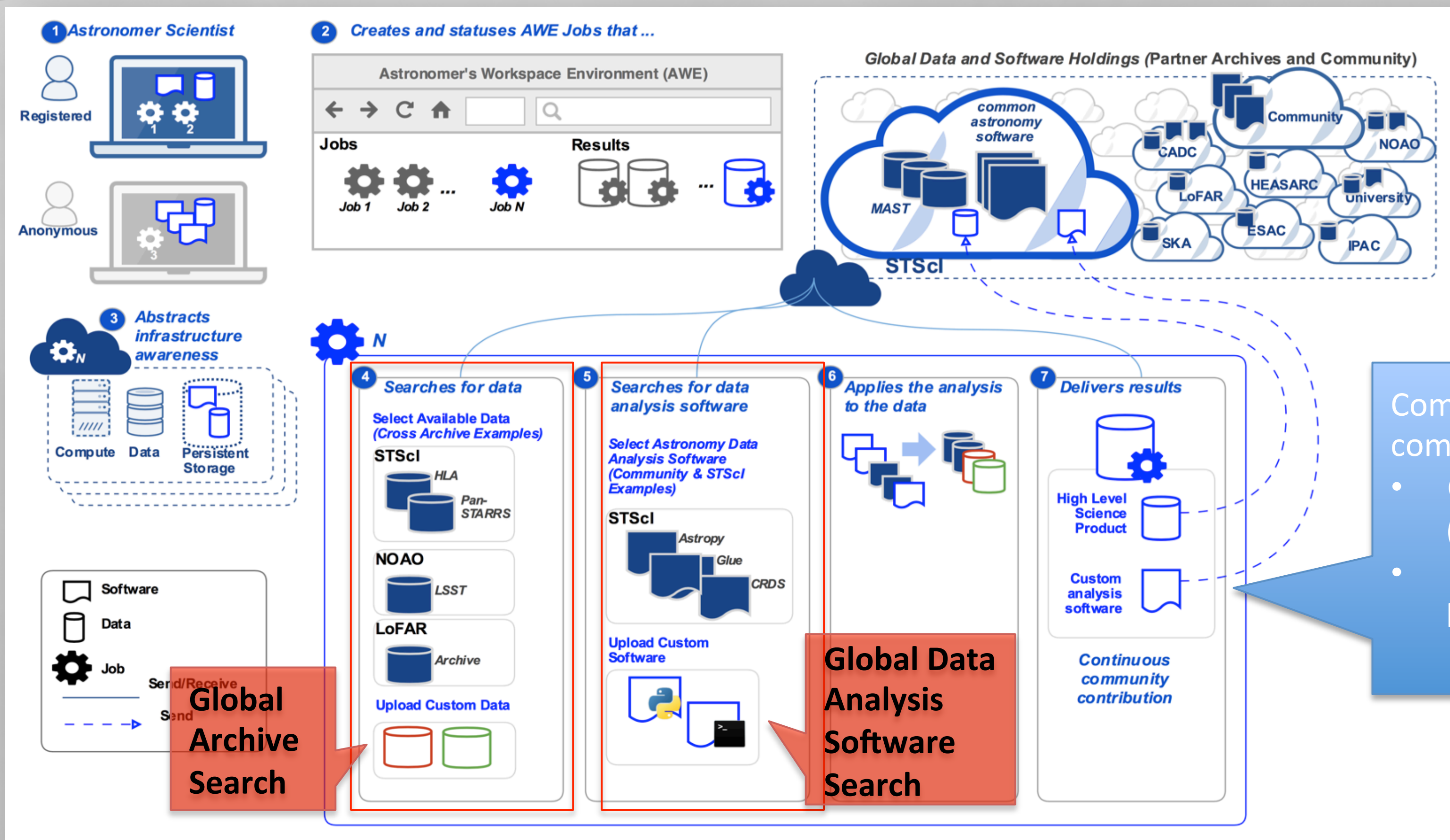
Way Ahead



	Science Cloud Evaluation	ST as a SciServer Node	ST SciServer Instance	STarCloud Solution
Summary	Comparative analysis across peer archive and industry solutions for data search, retrieval, and analysis	White Paper to define this concept and partnership between JHU and ST	Local installation updated per release; "feedback loop" environment for internal scientists' use and iterative feedback/development	Leverage lessons learned from evaluation; offer limited internal release (select data search/retrieval and select data analysis software) in Summer 2017
Deliverables	Trade Studies: Rev 1: initial 3 solutions Rev 2: + 3 more solutions … Rev *n*	White Paper	Use Cases via Jupyter Notebooks	• CONOPS • User Guide • Implementation Guide

***Pending Resource Availability**

Astronomer's Workspace Environment (AWE)



Global Archive Search

Global Data Analysis Software Search

Commonly used (and community-sourced):

- Catalog searches (VO Protocol XXX)
- Data-based search libraries

BACKUP

Technologies Evaluated

Trade Study Rev. 1

Technology	SciServer	GAVIP	Data Lab
Overall Cloud Technologies/Ecosystem			
OpenStack	x		
Platforms			
Windows	x		
Linux	x	x	x
Database Technologies			
PostgreSQL		x	
Microsoft SQL	x		
Storage			
OpenStack Swift			
VO Space			x

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Technologies Evaluated

Trade Study Rev. 1

Technology	SciServer	GAVIP	Data Lab
Job Execution			
Docker	x	x	x
Jupyter Notebooks	x	x	x
Django		x	
Luigi		x	
CLI	x	x	x
Distributed Processing			
OpenStack	x	x	
Celery		x	
RabbitMQ	x	x	
Authentication			
OpenStack Keystone	x		
CAS Oauth2		x	* Planned Support
IVOA SSO			* Planned support

Technologies Evaluated

Trade Study Rev. 1

Technology	SciServer	GAVIP	Data Lab
IVOA Supported Standards			
Cone Search	SkyServer		
SCS			x
SIA	SkyServer		x
SSA			x
SSO			x
SAMP		x	x
TAP +	SkyServer	x	x
VO Space		x	x
UWS			x

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Technologies Evaluated

Trade Study Rev. 1

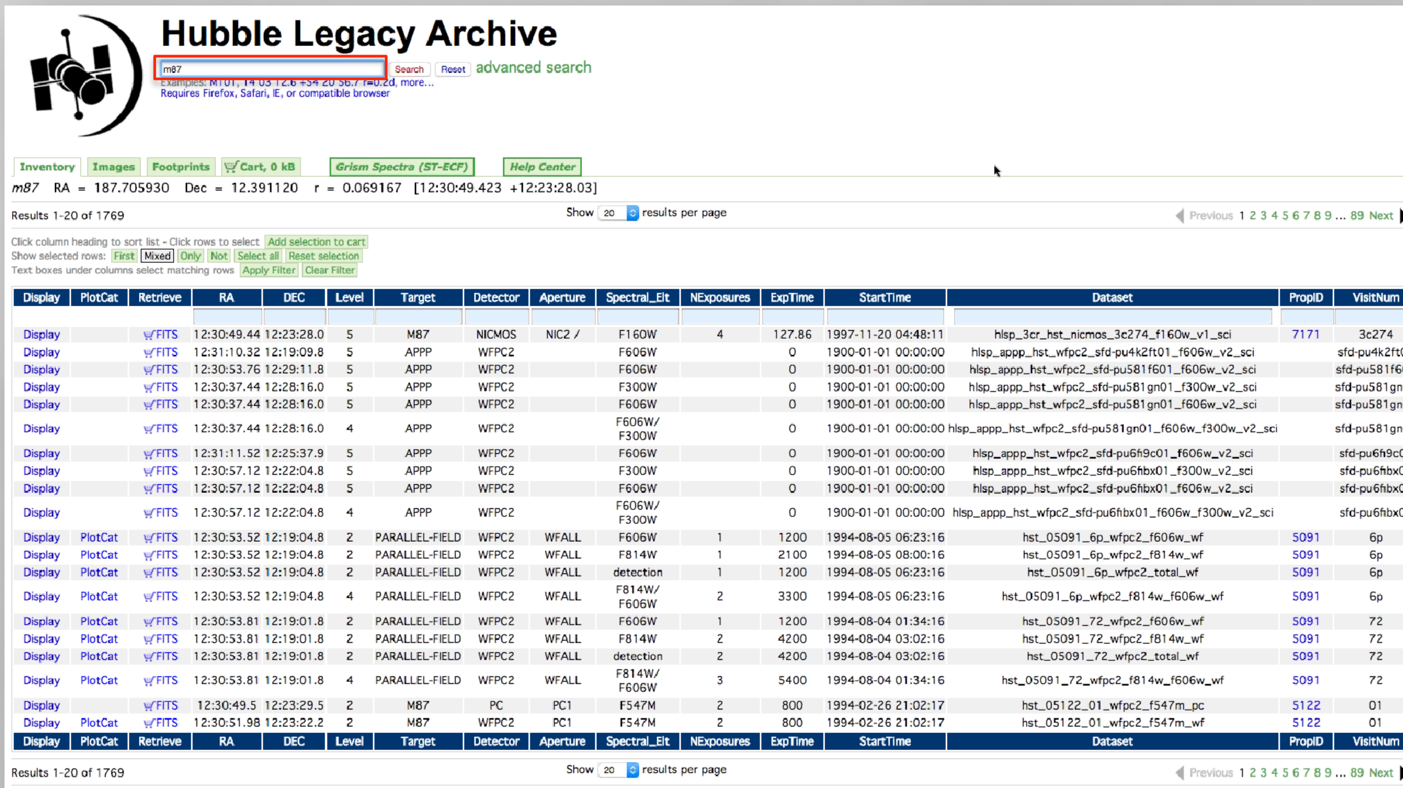
NOAO Data Lab

- Data Publication (i.e. hosted user datasets) and processing services associated with the data (e.g. image cutouts, catalog crossmatch,etc)
- Processing can be applied to published datasets (i.e. cutouts are the same whether coming from the main NOAO archive or a user's image collection) or as part of the VOSpace storage capability (e.g trigger a pipeline when a file is transferred to the virtual storage)
- Data Lab tools/services will be exportable for use outside NOAO and isn't tied to a particular hardware infrastructure (but naturally won't then have all the same capbilities)

Reference: NOAO Data Lab Architect Mike Fitzpatrick

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)



Hubble Legacy Archive

m87 Search Reset advanced search

Examples: M101, 14 03 12.6 +54 20 56.7 r=0.2d, more...
Requires Firefox, Safari, IE, or compatible browser

Inventory | Images | Footprints | Cart, 0 kB | Grism Spectra (ST-ECF) | Help Center

m87 RA = 187.705930 Dec = 12.391120 r = 0.069167 [12:30:49.423 +12:23:28.03]

Results 1-20 of 1769 Show 20 results per page Previous 1 2 3 4 5 6 7 8 9 ... 89 Next

Click column heading to sort list - Click rows to select Add selection to cart
Show selected rows: First Mixed Only Not Select all Reset selection
Text boxes under columns select matching rows Apply Filter Clear Filter

Display	PlotCat	Retrieve	RA	DEC	Level	Target	Detector	Aperture	Spectral_Elt	NExposures	ExpTime	StartTime	Dataset	PropID	VisitNum
Display		FITS	12:30:49.44	12:23:28.0	5	M87	NICMOS	NIC2 /	F160W	4	127.86	1997-11-20 04:48:11	hlsp_3cr_hst_nicmos_3c274_f160w_v1_sci	7171	3c274
Display		FITS	12:31:10.32	12:19:09.8	5	APPP	WFPC2		F606W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu4k2ft01_f606w_v2_sci		sfd-pu4k2ft0
Display		FITS	12:30:53.76	12:29:11.8	5	APPP	WFPC2		F606W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu581f601_f606w_v2_sci		sfd-pu581f60
Display		FITS	12:30:37.44	12:28:16.0	5	APPP	WFPC2		F300W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu581gn01_f300w_v2_sci		sfd-pu581gn0
Display		FITS	12:30:37.44	12:28:16.0	5	APPP	WFPC2		F606W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu581gn01_f606w_v2_sci		sfd-pu581gn0
Display		FITS	12:30:37.44	12:28:16.0	4	APPP	WFPC2		F606W/ F300W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu581gn01_f606w_f300w_v2_sci		sfd-pu581gn0
Display		FITS	12:31:11.52	12:25:37.9	5	APPP	WFPC2		F606W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu6fi9c01_f606w_v2_sci		sfd-pu6fi9c01
Display		FITS	12:30:57.12	12:22:04.8	5	APPP	WFPC2		F300W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu6fibx01_f300w_v2_sci		sfd-pu6fibx01
Display		FITS	12:30:57.12	12:22:04.8	5	APPP	WFPC2		F606W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu6fibx01_f606w_v2_sci		sfd-pu6fibx01
Display		FITS	12:30:57.12	12:22:04.8	4	APPP	WFPC2		F606W/ F300W		0	1900-01-01 00:00:00	hlsp_appp_hst_wfpc2_sfd-pu6fibx01_f606w_f300w_v2_sci		sfd-pu6fibx01
Display	PlotCat	FITS	12:30:53.52	12:19:04.8	2	PARALLEL-FIELD	WFPC2	WFALL	F606W	1	1200	1994-08-05 06:23:16	hst_05091_6p_wfpc2_f606w_wf	5091	6p
Display	PlotCat	FITS	12:30:53.52	12:19:04.8	2	PARALLEL-FIELD	WFPC2	WFALL	F814W	1	2100	1994-08-05 08:00:16	hst_05091_6p_wfpc2_f814w_wf	5091	6p
Display	PlotCat	FITS	12:30:53.52	12:19:04.8	2	PARALLEL-FIELD	WFPC2	WFALL	detection	1	1200	1994-08-05 06:23:16	hst_05091_6p_wfpc2_total_wf	5091	6p
Display	PlotCat	FITS	12:30:53.52	12:19:04.8	4	PARALLEL-FIELD	WFPC2	WFALL	F814W/ F606W	2	3300	1994-08-05 06:23:16	hst_05091_6p_wfpc2_f814w_f606w_wf	5091	6p
Display	PlotCat	FITS	12:30:53.81	12:19:01.8	2	PARALLEL-FIELD	WFPC2	WFALL	F606W	1	1200	1994-08-04 01:34:16	hst_05091_72_wfpc2_f606w_wf	5091	72
Display	PlotCat	FITS	12:30:53.81	12:19:01.8	2	PARALLEL-FIELD	WFPC2	WFALL	F814W	2	4200	1994-08-04 03:02:16	hst_05091_72_wfpc2_f814w_wf	5091	72
Display	PlotCat	FITS	12:30:53.81	12:19:01.8	2	PARALLEL-FIELD	WFPC2	WFALL	detection	2	4200	1994-08-04 03:02:16	hst_05091_72_wfpc2_total_wf	5091	72
Display	PlotCat	FITS	12:30:53.81	12:19:01.8	4	PARALLEL-FIELD	WFPC2	WFALL	F814W/ F606W	3	5400	1994-08-04 01:34:16	hst_05091_72_wfpc2_f814w_f606w_wf	5091	72
Display		FITS	12:30:49.5	12:23:29.5	2	M87	PC	PC1	F547M	2	800	1994-02-26 21:02:17	hst_05122_01_wfpc2_f547m_pc	5122	01
Display	PlotCat	FITS	12:30:51.98	12:23:22.2	2	M87	WFPC2	PC1	F547M	2	800	1994-02-26 21:02:17	hst_05122_01_wfpc2_f547m_wf	5122	01
Display	PlotCat	Retrieve	RA	DEC	Level	Target	Detector	Aperture	Spectral_Elt	NExposures	ExpTime	StartTime	Dataset	PropID	VisitNum

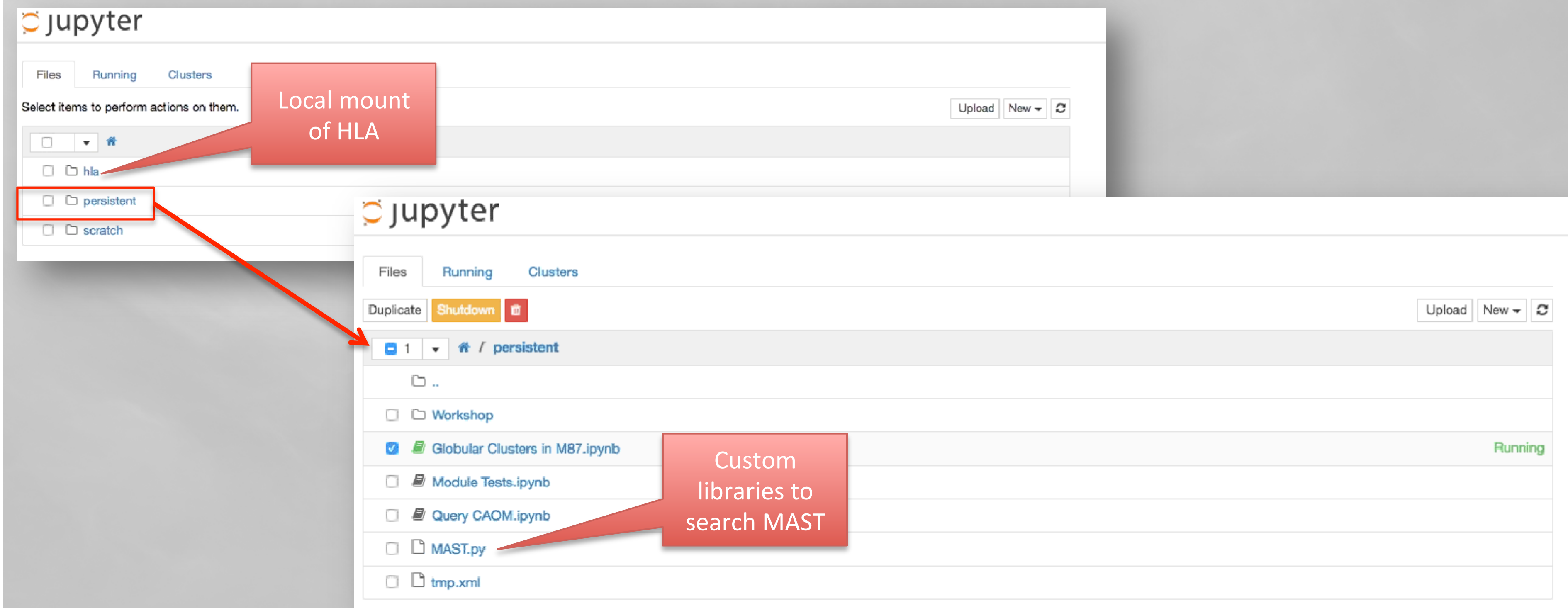
Results 1-20 of 1769 Show 20 results per page Previous 1 2 3 4 5 6 7 8 9 ... 89 Next

Today:
Use HLA site to search for M87

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)

Create Jupyter notebook to perform search, analysis, and display



STScI | SPACE TELESCOPE SCIENCE INSTITUTE

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)

jupyter Globular Clusters in M87 Last Checkpoint: 08/25/2016 (autosaved)

File Edit View Insert Cell Kernel Help | Python 3

Markdown | CellToolbar

1. Sign into SciServer

Load the login token and store it as a local variable.

In [1]:
```
# This code block defined your token and makes it available as a
#   system variable for the length of your current session.
#
# This will usually be the first code block in any script you write.
with open('/home/idies/keystone.token', 'r') as f:
    token = f.read().rstrip('\n')
# async queries require token to be in --ident system variable
import sys
sys.argv.append("--ident="+token)
```

2. Load the SciServer Libraries

These libraries provide communication functions with SciServer resources.

In [19]:
```
import SciServer.CasJobs as CasJobs                  # query with CasJobs
import SciServer.Authentication as Authentication    # controls/manages authentication
import MAST
#import SciServer.SciDrive                           # read/write to/from SciDrive
```

The authentication token can be retrieved as follows:

In [20]:
```
my_access_token = Authentication.getToken()
print("Your current token is",my_access_token)
```

```
Your current token is fb6d254f112548 aa42ac0833f20a407
```

Import the usual libraries:

In [4]:
```
from astropy.io.votable import parse,from_table,parse_single_table
from astropy.io.votable.tree import VOTableFile
from astropy.io import fits
from astropy.wcs import WCS
import astropy.visualization
from astropy.visualization.mpl_normalize import ImageNormalize
from wcsaxes import datasets

import numpy as np                     # standard Python lib for math ops
import pandas                          # data manipulation package
import matplotlib
import matplotlib.pyplot as plt        # another graphing package
import skimage.io                      # image processing library
import urllib                          # accessing resources thorugh remote URLs
import json                            # work with JSON files
from xml.dom import minidom
from xml.etree import ElementTree
import requests
import stsci
%matplotlib inline
```

In [23]:
```
# Apply some special settings to the imported libraries
# ensure columns get written completely in notebook
pandas.set_option('display.max_colwidth', -1)
# do *not* show python warnings
import warnings
warnings.filterwarnings('ignore')
```

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)

3. Query the HSC for Globular Clusters around M87

First we want to query the HSC for extended sources with the colors appropriate to globular clusters. We do this using CasJobs.

In [6]:
```
query="""
SELECT
      MatchRA, MatchDEC, MatchID, CI, W2_F606W, W2_F814W, V_I=W2_F606W - W2_F814W
FROM
      SearchSumCatalog(187.706, 12.391, 500.0, 1)
      where CI > 1.05 and CI < 1.5
      and (W2_F606W - W2_F814W) > 0.0 and (W2_F606W - W2_F814W) < 1.0
      and numimages > 50
ORDER BY matchID
"""
response = CasJobs.executeQuery(query, "HSC", token = my_access_token)
print(response[:10])

# Adjust the positions by a small amount
ra = response['MatchRA']
dec = response['MatchDEC']
```

	MatchRA	MatchDEC	MatchID	CI	W2_F606W	W2_F814W	V_I
0	187.687274	12.369296	250841	1.302192	22.735899	22.426411	0.309488
1	187.687805	12.372637	250963	1.205647	22.660417	22.371114	0.289303
2	187.684690	12.374775	250964	1.279722	22.161032	21.800969	0.360063
3	187.684618	12.375990	250982	1.237424	23.004038	22.618353	0.385685
4	187.681527	12.382457	251022	1.194007	23.272181	22.686539	0.585642
5	187.680604	12.378050	251325	1.355000	23.194697	22.650146	0.544552
6	187.692877	12.370444	251397	1.231290	22.252050	21.707011	0.545039
7	187.684363	12.374692	251476	1.280098	23.415614	22.973073	0.442541
8	187.683676	12.375708	251487	1.427002	23.529730	23.035396	0.494334
9	187.699635	12.366596	251499	1.329482	23.332398	22.912575	0.419823

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

4. Query the HLA

Next we want to overplot these positions on an HLA image. To do this we first need to query the HLA for images of M87. We do this using a pre-built routine in the MAST module.

In [7]:
```
# Query for all HLA images at the position of the first RA, Dec in the HSC query we ran previously.
table = MAST.HLA_executeQuery(
    RA  = ra[0],          \
    Dec = dec[0],         \
    instruments="",       \
    image_type="best", \
    format_type="image/fits"
    )
```

```
Calling: http://hla.stsci.edu/cgi-bin/hlaSIAP.cgi?pos=187.687273844,12.3692964212&size=0
h=0&format=image/fits,application/tar,text/html
Done.
```

Parse VO Table for display in notebook

Print some summary results of the query, including a list of detectors, file levels, etc./

In [8]:
```
# Number of results and list of columns
print("Number of results:", len(table))
print()
print("  Columns:",table.colnames)
print()
print(np.unique(table['Detector']))
print()
print(np.unique(table['Level']))
```

```
Number of results: 381

  Columns: ['URL', 'RA', 'DEC', 'Level', 'Target', 'Detector', 'Aperture', 'Spectral_Elt
 'StartTime', 'Dataset', 'PropID', 'VisitNum', 'PI_Name', 'DAOCat', 'SEXCat', 'Mode', 'D
velength', 'Format', 'ReleaseDate', 'NReleaseDate', 'Title', 'naxis', 'scale', 'naxes',
x', 'filename', 'regionSTCS']

 Detector
---------
  ACS/WFC
WFC3/UVIS
    WFPC2

 Level
 -----
     1
     2
     4
```

Next filter the broad query down into a list of images of more interest.

In [28]:
```
# Print the results in the notebook
table['RA','DEC','Level','Detector','Aperture','NExposures','ExpTime','Dataset'][mask].show_in_notebook()
```

Out[28]: <Table masked=True length=100>

Show 50 entries Search:

idx	RA	DEC	Level	Detector	Aperture	NExposures	ExpTime	Dataset
	deg	deg					s	
99	187.71218999999999	12.391106000000001	2	ACS/WFC	WFC	5	1960.0	hst_9401_02_acs_wfc_total
0	187.70935	12.390872	2	ACS/WFC	WFC	4	1440.0	hst_10543_01_acs_wfc_f814w
1	187.70935	12.390872	2	ACS/WFC	WFC	4	1440.0	hst_10543_01_acs_wfc_total
2	187.70844	12.391050999999999	2	ACS/WFC	WFC	4	1440.0	hst_10543_02_acs_wfc_f814w
3	187.70844	12.391050999999999	2	ACS/WFC	WFC	4	1440.0	hst_10543_02_acs_wfc_total
4	187.70937000000001	12.390943999999999	2	ACS/WFC	WFC	4	1440.0	hst_10543_03_acs_wfc_f814w
5	187.70937000000001	12.390943999999999	2	ACS/WFC	WFC	4	1440.0	hst_10543_03_acs_wfc_total

In [13]:
```
# Get the image with the larget exposure time
print("Longest exposure time in subselection:", np.amax(table['ExpTime'][mask]))

# Find out which element in the array this corresponds to and set
# the name of the dataset to retrieve
idx = np.argmax(table['ExpTime'][mask])
dataset = table['Dataset'][mask][idx]
print(" -> Dataset:", dataset)
```

```
Longest exposure time in subselection: 1960.0
 -> Dataset: hst_9401_02_acs_wfc_total
```

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)

5. Retrieve an image from the Isilon.

Next we want toretrieve and display an HLA image from the Isilon. We begin by opening it up as a FITS header data unit (HDU).

ST storage

In a "real-world" implementation of SciServer, we would want custom Python modules which return an Isilon path given a dataset name. For example:

```
path = MAST.HLA_getImage(dataset)
print("Path:",path)
```

For now, just hard-code the path to demonstrate that file retrieval actually works:

In [14]:
```
# For now, simply hard-code the path to the file
path = "/home/idies/workspace/hla/acs/V9.0/9401/9401_02/hst_9401_02_acs_wfc_total_drz.fits"
#path = "/home/idies/workspace/hla/acs/V9.0/10543/10543_01/hst_10543_01_acs_wfc_total_drz.fits"
#path = "/home/idies/workspace/hla/acs/V9.0/10543/10543_01/hst_10543_01_acs_wfc_f814w_drz.fits"

# Open the FITS image and print out file information
hdu_list = fits.open(path)
#hdu_list.info()

# Extract the image data you are interested in
image_data = hdu_list['SCI'].data

# Set all the values of NaN to zero
image_data = np.nan_to_num(image_data)

# Get the WCS information
wcs = WCS(hdu_list['SCI'].header)

# Now close the FITS file
hdu_list.close()
```

Now display the image data. We want to do a z-scale stretch.

In [15]:
```
# First change the data interval by rescaling
interval = astropy.visualization.ZScaleInterval()
#interval = astropy.visualization.PercentileInterval(99)

image_data = interval(image_data)

# Create a normalizer object to do the stretch
minmax   = interval.get_limits(image_data)
norm = ImageNormalize(vmin=minmax[0], vmax=minmax[1]*.90, stretch=astropy.visualization.LinearStretch())
```

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

SciServer Use Case

Globular Clusters Search around M87 (HSC & HLA query)

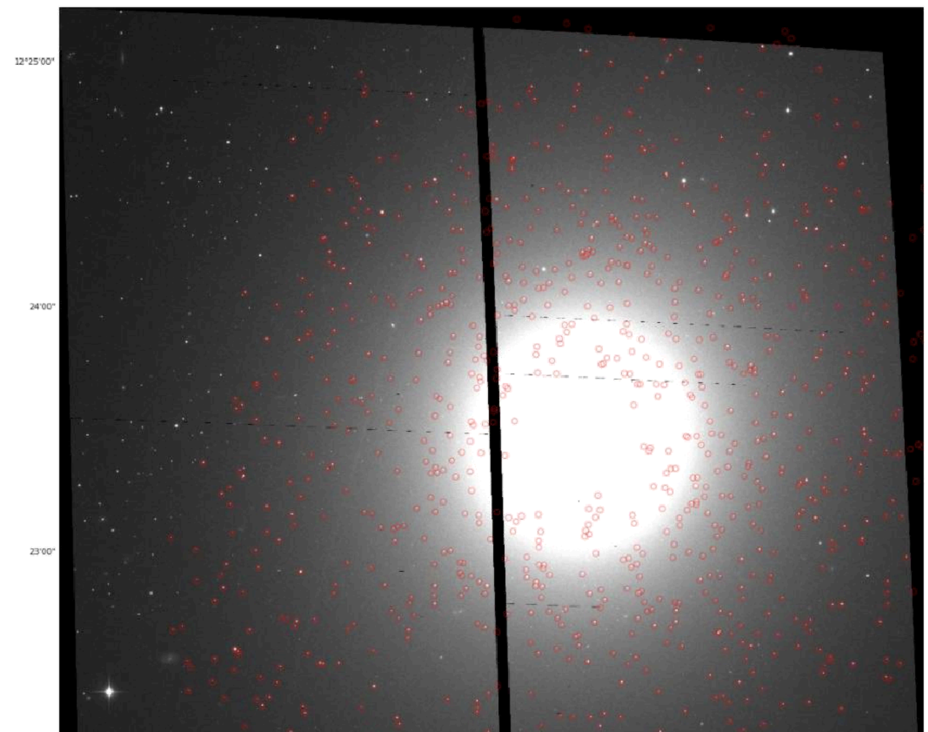
6. Display the image.

We will use a series of calls to setup a WCS projection of the image.

In [16]:

```
# Setup the figure and add the WCS axes as the projection
fig = plt.figure(figsize=(15,15))
ax = fig.add_axes([1., 1., 1., 1.],projection=wcs)

# Set the axis limits
ax.set_xlim(0, image_data.shape[1])
ax.set_ylim(0, image_data.shape[0])

#ax.tick_params(axis='both', which='major', labelsize=100)
#ax.tick_params(axis='both', which='minor', labelsize=8)

# Plot
cax = ax.imshow(image_data, cmap='gray', origin='lower',norm=norm)

# Overplot coordinates
ax.scatter(ra,dec,transform=ax.get_transform('icrs'),s=50,edgecolor='red', facecolor='none', alpha=0.5)
```

Out[32]: <matplotlib.collections.PathCollection at 0x7ff9f98a60f0>



STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Big Data @ STScI Report

Big Data Computing and Software Capabilities Needed

#	Science Case	Computing	Storage	Bandwidth	Software
1	Light echo detection and classification	>1000 CPU core cluster	~ 1 to 10 PB	~ 10GB citizen science and data transfer	• Machine Learning (ML) classification, • feature vectors, • citizen science user interface
2	Lensed galaxy detection	~1000 CPU core cluster	~ 10 TB (if image cutout service is available for large surveys)	~ 1 Gbps	• Parallelizable code management • ML algorithms
3	Resolved stellar populations	> 1000 CPU core cluster (~10K core ultimately needed)	Few 100 TB	~10 Gbps (uses non-local data)	• Automated pipeline management software • Efficient database query tools
4	3D Structure in the cosmos	~500 core cluster	>200 TB	>1 Gbps	• Parallelizable code management • Clustered algorithms • Photo-z algorithms
5	SMBH / Galaxy Co-evolution	> 2000 core CPU	> 1 PB	> 1 Gbps	• ML • Feature vectors • Efficient cross correlation algorithms • High dimension data visualization tools

Postman, et. al. Science Definition Team (SDT) Report. *Big Data @ STScI*, 2016, pg. 61

Big Data @ STScI Report

Big Data Computing and Software Capabilities Needed

#	Science Case	Computing	Storage	Bandwidth	Software
6	TESS supernova search	~ 1000 CPU core	~ 10 TB for raw data	~ 1 Gbps (if data local)	• Transient detection algorithms for highly undersampled data
7	GALEX Variability catalog	High performance computing required Exact specs TBD	Few hundred TB, solid state disks for fast access	High bandwidth to local db (~10 Gbps)	• Algorithms to detect and classify transients from archive of 10^8 light curves
8	Multidimensional spectroscopic datasets	Current capabilities sufficient	~ 100 GB	Current capabilities sufficient	• Feature vectors • ML algorithms • Automated spectral feature detection software • Efficient cross-correlation algorithms
9	Image cutout service for wide area sky surveys	Dedicated high-user volume server needed	> 1 PB	10 Gbps or more	• Must support thousands of users and many simultaneous queries

Postman, et. al. Science Definition Team (SDT) Report. *Big Data @ STScI*, 2016, pg. 61

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

STScI | SPACE TELESCOPE SCIENCE INSTITUTE