SciServer — Compute Brings Analysis to Big Data in the Cloud

idies — The Institute for Data Intensive Engineering and Science

Jordan Raddick, Dmitry Medvedev, Gerard Lemson, Barbara Souter
Institute for Data Intensive Engineering and Science, Johns Hopkins University

How to compute - a step-by-step guide

1 Go to http://compute.sciserver.org and click Sign In.

2 Sign in (or register) to the SciServer login portal to get access to all our resources.

Log in or Register to access **SciServer Services**.

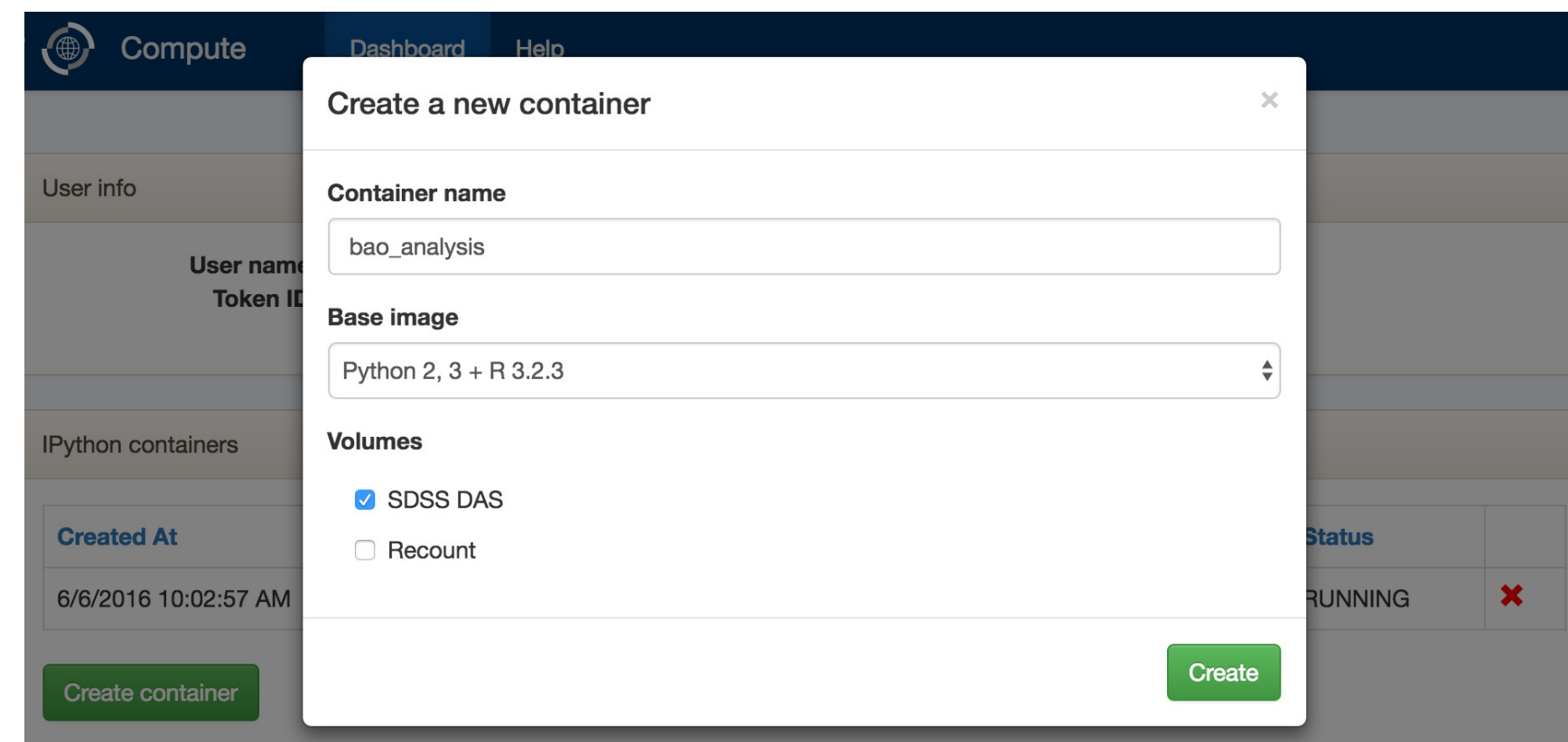If you have an existing CasJobs account and are using the Login Portal for the first time, please **Register here**.

Log in to an existing SciServer account.

User name: jordan
Password: •••••••••

[Submit]

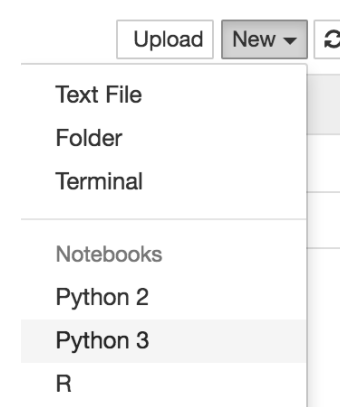**3** Create a container. Name it, select your base image, and check off any data sources to mount. Click Create.

Create a new container
Container name: bao_analysis
Base image: Python 2, 3 + R 3.2.3
Volumes: ☑ SDSS DAS ☐ Recount
[Create]

4 Click on your container name to open it in a new window.

Created At	Name	Docker ID	Image	Host	Status
6/9/2016 2:17:35 PM	bao_analysis	5359ca710a42...	mro_323	blackbox3	RUNNING

5 To save data between sessions, use the persistent folder.

6 Click New to create a new notebook, and select your scripting language. OR, drag any .ipynb file into the window and click Upload.

Upload | New
Text File
Folder
Terminal
Notebooks
Python 2
Python 3
R

SciServer Compute is a new resource that lets you run Python and R scripts on big data online in Jupyter notebooks. Compute **brings the analysis to the data**.

Data
We offer access to the complete SDSS dataset to the world through SkyServer.
You can store any data you like in a private database in CasJobs, or as files in SciDrive.

Query
You can find SDSS data with SQL queries through SkyServer or CasJobs.
You can run queries through scripts using Compute.

SkyServer

SciDrive

CasJobs

Compute

Collaboration
You can share your data and results with colleagues while you are Logged In.
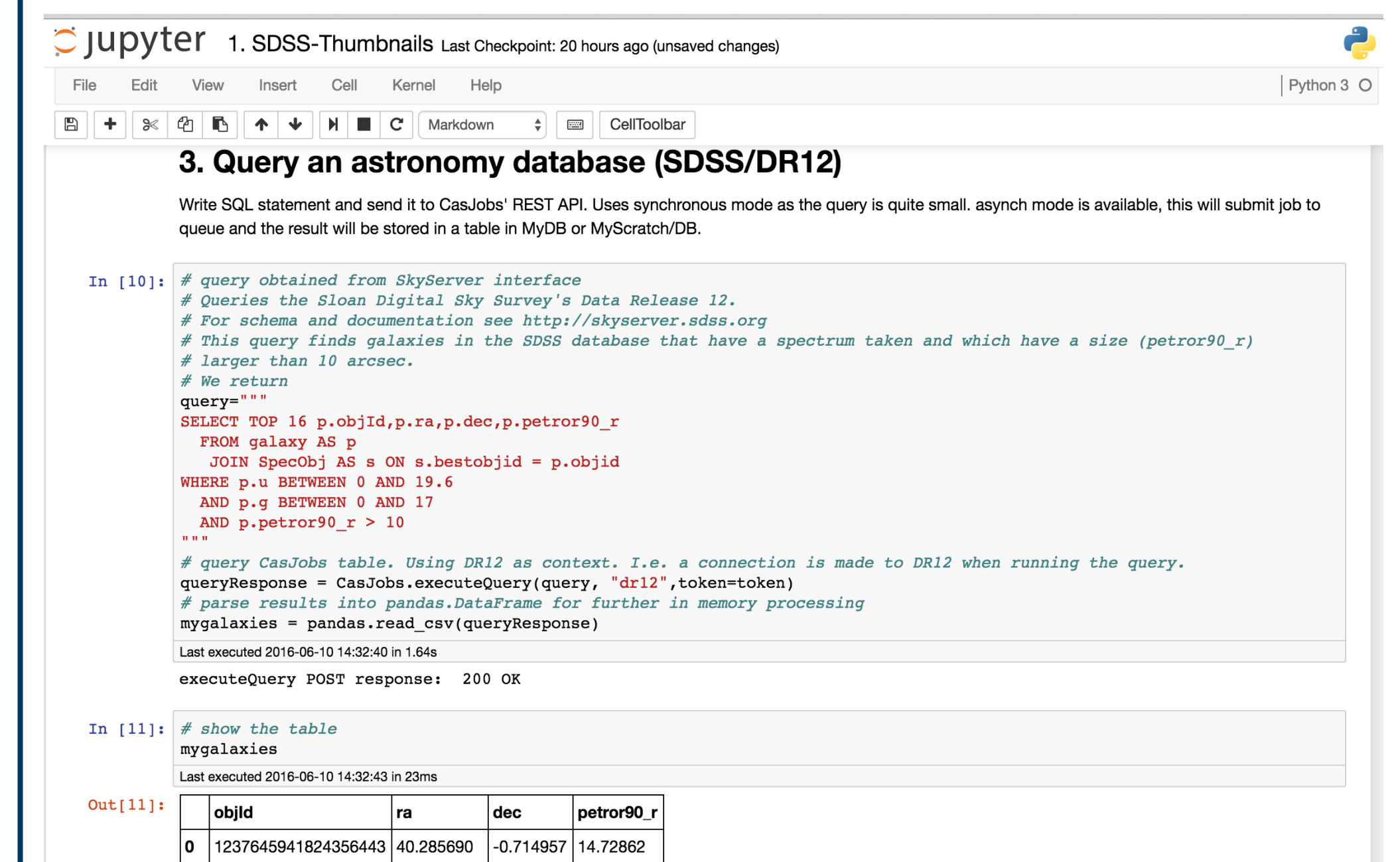You can share file data in any format with SciDrive.

Analysis
You can analyze huge datasets with Python or R scripts through Compute.
You can share scripts and results while Logged In.

Login Portal

How SciServer Compute works

SciServer Compute uses Jupyter notebooks running inside a Docker container on a virtual machine on servers at Johns Hopkins University. The SciServer login portal provides single-sign on to Compute and all SciServer resources. You can upload your scripts into our virtual machines and run any analysis commands within a self-contained, independent and reproducible environment.

Online analysis with iPython notebooks

SciServer Compute lets you run computations online. You no longer need to download anything to run sophisticated analyses, such as FFTs, on entire big datasets. Just go to the Compute website (see instructions on the left) and upload your R or Python (2 or 3) scripts into a Jupyter notebook. You can read and write data from/to your CasJobs MyDB or SciDrive.

Example - query SDSS DR12 in Python

3. Query an astronomy database (SDSS/DR12)

Write SQL statement and send it to CasJobs' REST API. Uses synchronous mode as the query is quite small. asynch mode is available, this will submit job to queue and the result will be stored in a table in MyDB or MyScratch/DB.

```
In [10]:  # query obtained from SkyServer interface
          # Queries the Sloan Digital Sky Survey's Data Release 12.
          # For schema and documentation see http://skyserver.sdss.org
          # This query finds galaxies in the SDSS database that have a spectrum taken and which have a size (petror90_r)
          # larger than 10 arcsec.
          # We return
          query="""
          SELECT TOP 16 p.objId,p.ra,p.dec,p.petror90_r
            FROM galaxy AS p
            JOIN SpecObj AS s ON s.bestobjid = p.objid
          WHERE p.u BETWEEN 0 AND 19.6
            AND p.g BETWEEN 0 AND 17
            AND p.petror90_r > 10
          """
          # query CasJobs table. Using DR12 as context. I.e. a connection is made to DR12 when running the query.
          queryResponse = CasJobs.executeQuery(query, "dr12",token=token)
          # parse results into pandas.DataFrame for further in memory processing
          mygalaxies = pandas.read_csv(queryResponse)

          Last executed 2016-06-10 14:32:40 in 1.64s
          executeQuery POST response:  200 OK

In [11]:  # show the table
          mygalaxies
          Last executed 2016-06-10 14:32:43 in 23ms
```

Out[11]:
	objId	ra	dec	petror90_r
0	1237645941824356443	40.285690	-0.714957	14.72862

This notebook uses a Python 3 script to search for data from the Sloan Digital Sky Survey Data Release 12. It makes a simple plot, then retrieves thumbnails as JPGs and store them in SciDrive.

What Compute can do for you

SciServer Compute puts powerful Big Data analysis at your fingertips. Here are some of the things you can do:
• interactively build complex data analysis and extraction code
• run complex analysis right next to the data to minimize data movement
• store the results of your analysis to SciDrive, MyScratch or MyDB
• share your results and code (as a Jupyter notebook) with collaborators

Astronomical applications of Compute include:
• create postage stamp images of your favorite SDSS objects
• create co-added images with various stretches to highlight features
• derive and evaluate photometric redshift estimates
• compare cosmological simulations (e.g. Millennium) to observations
• assign students homework and labs to work with real data online

---

Questions? Email us at sciserver-webmaster@jhu.edu!