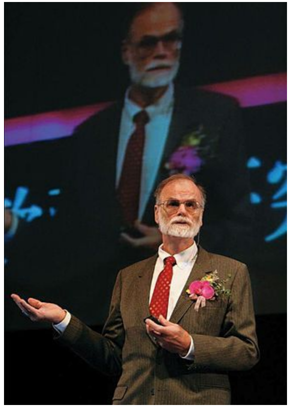
# SciServer

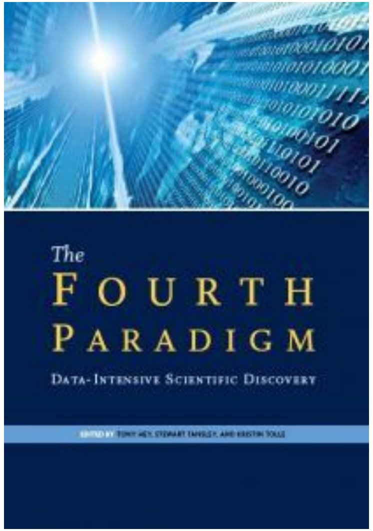
A Collaborative Science Platform developed by IDIES

# Outline

- Background
- SciServer overview
- Use case highlights
- Architecture & Deployments
- New/hidden features/accelerated computing
- Future directions
- Questions

# Beginnings



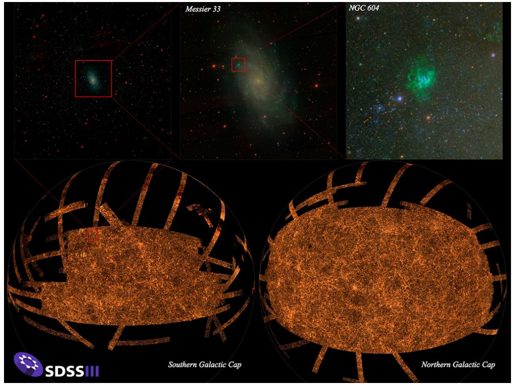


The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

SDSS

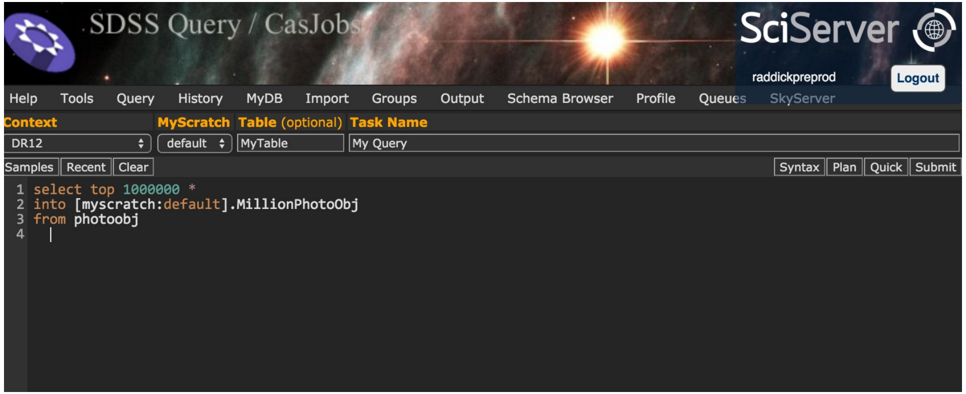


The fourth paradigm: **Data exploration**.

More data being produced than can be analyzed, increasing dependence on tools and machine learning to help make sense.

# SDSS / CasJobs

- SQL interface to large catalog of object data
- Server-side execution, download small result-set of interest
- Personal DB space
- Sharing features



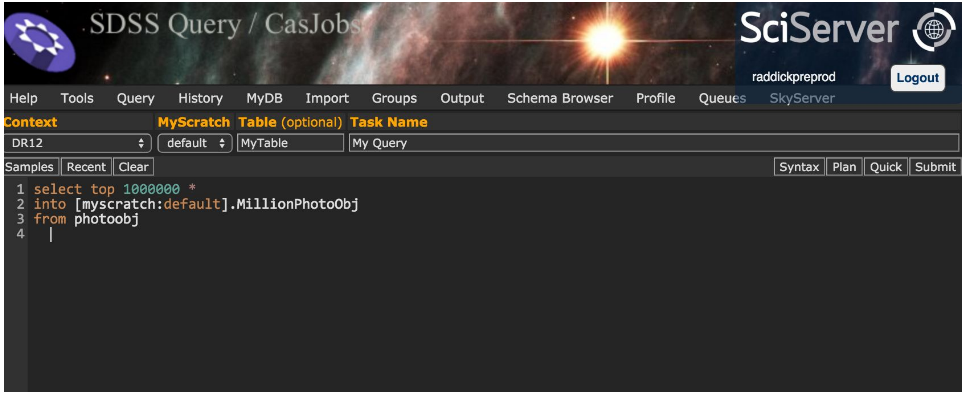
Context: DR12
MyScratch: default
Table (optional): MyTable
Task Name: My Query

```
select top 1000000 *
into [myscratch:default].MillionPhotoObj
from photoobj
```

# SDSS / CasJobs

- SQL interface to large catalog of object data
- Server-side execution, download small result-set of interest
- Personal DB space
- Sharing features

- Still need to download
- No solution for file-based data
- limited/no visualization support



Context: DR12

MyScratch: default

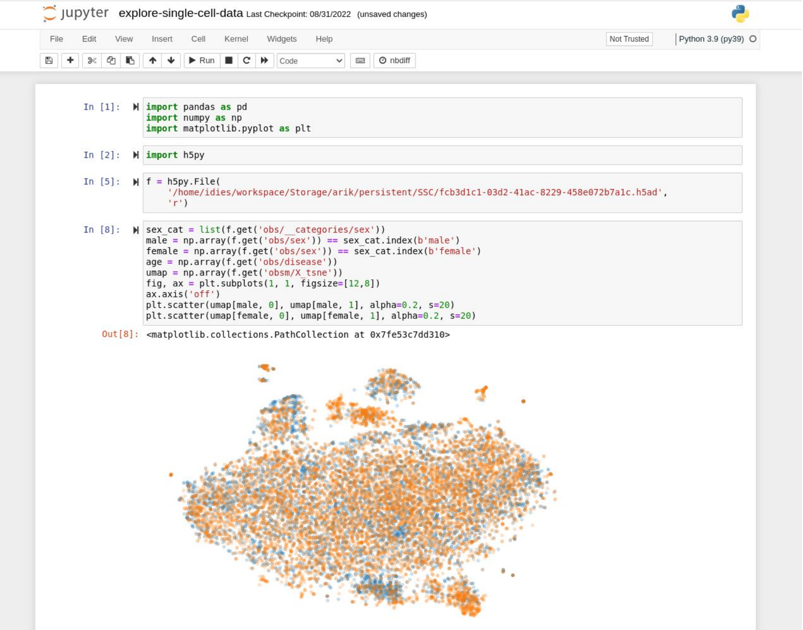
Table (optional): MyTable

Task Name: My Query

```
select top 1000000 *
into [myscratch:default].MillionPhotoObj
from photoobj
```

# Beyond SQL

- Add Jupyter
- Docker makes scheduling simple
- Jupyter can run arbitrary codes
- Inline visualization
- Common dependencies included
- SDK for CasJobs (get a pandas DF)

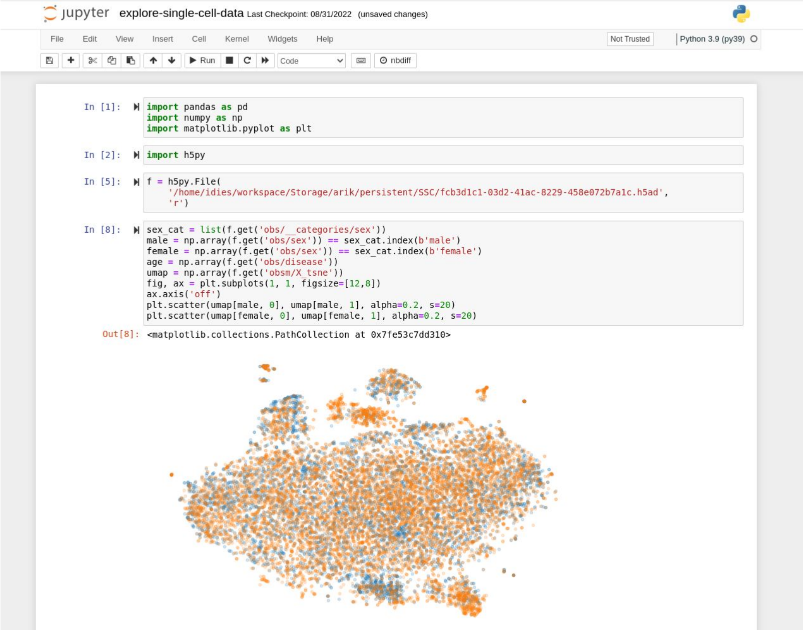


```
In [1]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt

In [2]:  import h5py

In [5]:  f = h5py.File(
             '/home/idies/workspace/Storage/arik/persistent/SSC/fcb3d1c1-03d2-41ac-8229-458e072b7a1c.h5ad',
             'r')

In [8]:  sex_cat = list(f.get('obs/__categories/sex'))
         male = np.array(f.get('obs/sex')) == sex_cat.index(b'male')
         female = np.array(f.get('obs/sex')) == sex_cat.index(b'female')
         age = np.array(f.get('obs/disease'))
         umap = np.array(f.get('obsm/X_tsne'))
         fig, ax = plt.subplots(1, 1, figsize=[12,8])
         ax.axis('off')
         plt.scatter(umap[male, 0], umap[male, 1], alpha=0.2, s=20)
         plt.scatter(umap[female, 0], umap[female, 1], alpha=0.2, s=20)

Out[8]:  <matplotlib.collections.PathCollection at 0x7fe53c7dd310>
```

# Beyond SQL

- Add Jupyter
- Docker makes scheduling simple
- Jupyter can run arbitrary codes
- Inline visualization
- Common dependencies included
- SDK for CasJobs (get a pandas DF)

- Server-side collaboration?
- Expanding scope?

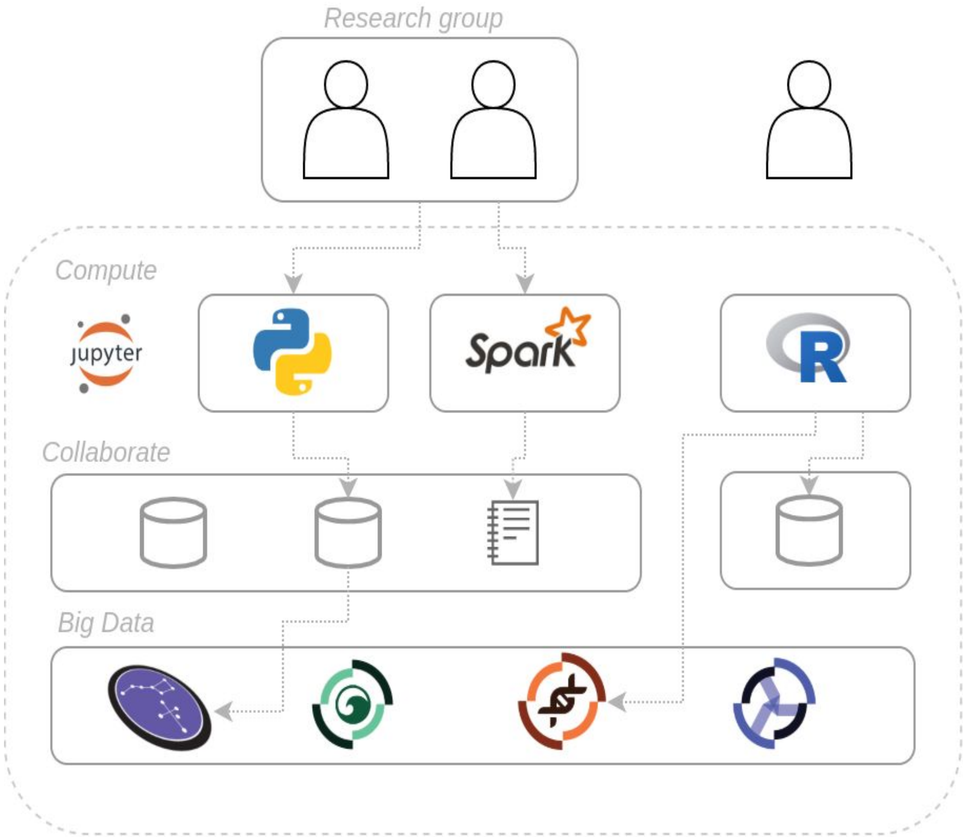


```
In [1]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt

In [2]:  import h5py

In [5]:  f = h5py.File(
             '/home/idies/workspace/Storage/arik/persistent/SSC/fcb3d1c1-03d2-41ac-8229-458e072b7a1c.h5ad',
             'r')

In [8]:  sex_cat = list(f.get('obs/__categories/sex'))
         male = np.array(f.get('obs/sex')) == sex_cat.index(b'male')
         female = np.array(f.get('obs/sex')) == sex_cat.index(b'female')
         age = np.array(f.get('obs/disease'))
         umap = np.array(f.get('obsm/X_tsne'))
         fig, ax = plt.subplots(1, 1, figsize=[12,8])
         ax.axis('off')
         plt.scatter(umap[male, 0], umap[male, 1], alpha=0.2, s=20)
         plt.scatter(umap[female, 0], umap[female, 1], alpha=0.2, s=20)

Out[8]:  <matplotlib.collections.PathCollection at 0x7fe53c7dd310>
```

# SciServer

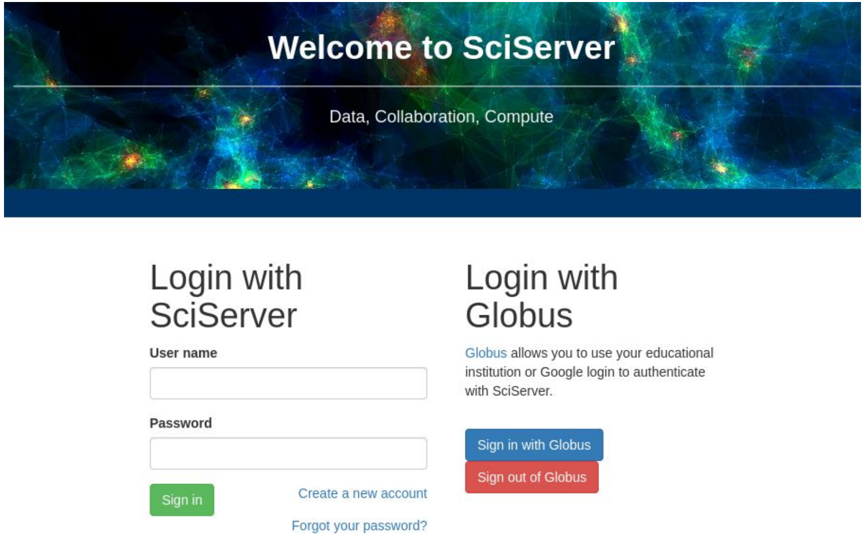
- Add in an ecosystem
- Sharing/resource access control
- Web based portals
- Curated data
  - Astronomy
  - Genomics
  - Earth Science
  - Simulation data,...

*A Science Platform*



# Basic Usage

- Go to https://apps.sciserver.org
- Create an account
    - You can sign in with your institution via globus, your google account or orcid by using “sign in with Globus”



Welcome to SciServer

Data, Collaboration, Compute

Login with SciServer

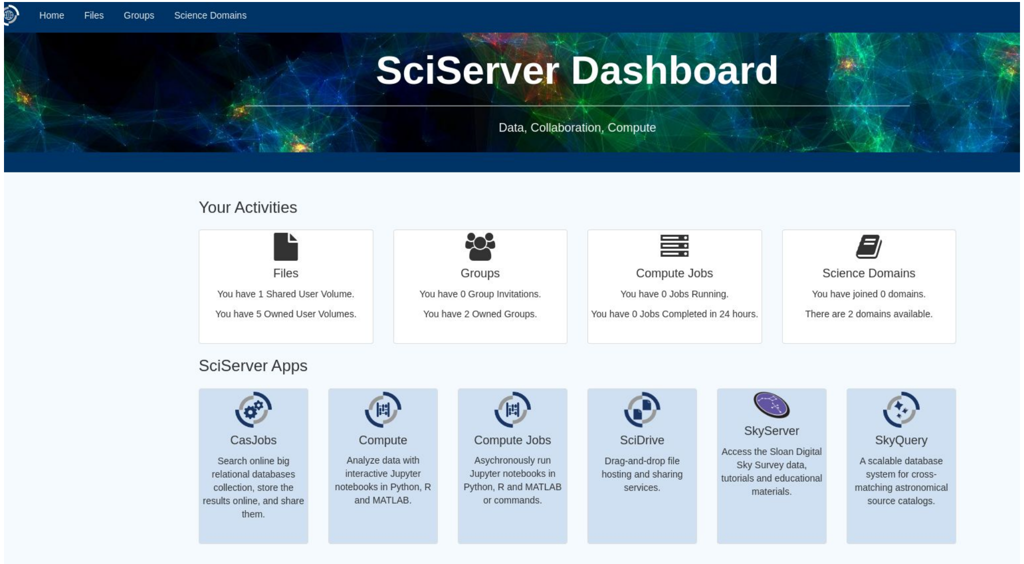
User name

Password

Globus allows you to use your educational institution or Google login to authenticate with SciServer.

# Basic Usage

- Arrive at dashboard
- A number of links to services
  - Files
  - Groups
  - Compute
  - Casjobs



Home Files Groups Science Domains

SciServer Dashboard

Data, Collaboration, Compute

Your Activities

Files
You have 1 Shared User Volume.
You have 5 Owned User Volumes.

Groups
You have 0 Group Invitations.
You have 2 Owned Groups.

Compute Jobs
You have 0 Jobs Running.
You have 0 Jobs Completed in 24 hours.

Science Domains
You have joined 0 domains.
There are 2 domains available.

SciServer Apps

CasJobs
Search online big relational databases collection, store the results online, and share them.

Compute
Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.

Compute Jobs
Asychronously run Jupyter notebooks in Python, R and MATLAB or commands.

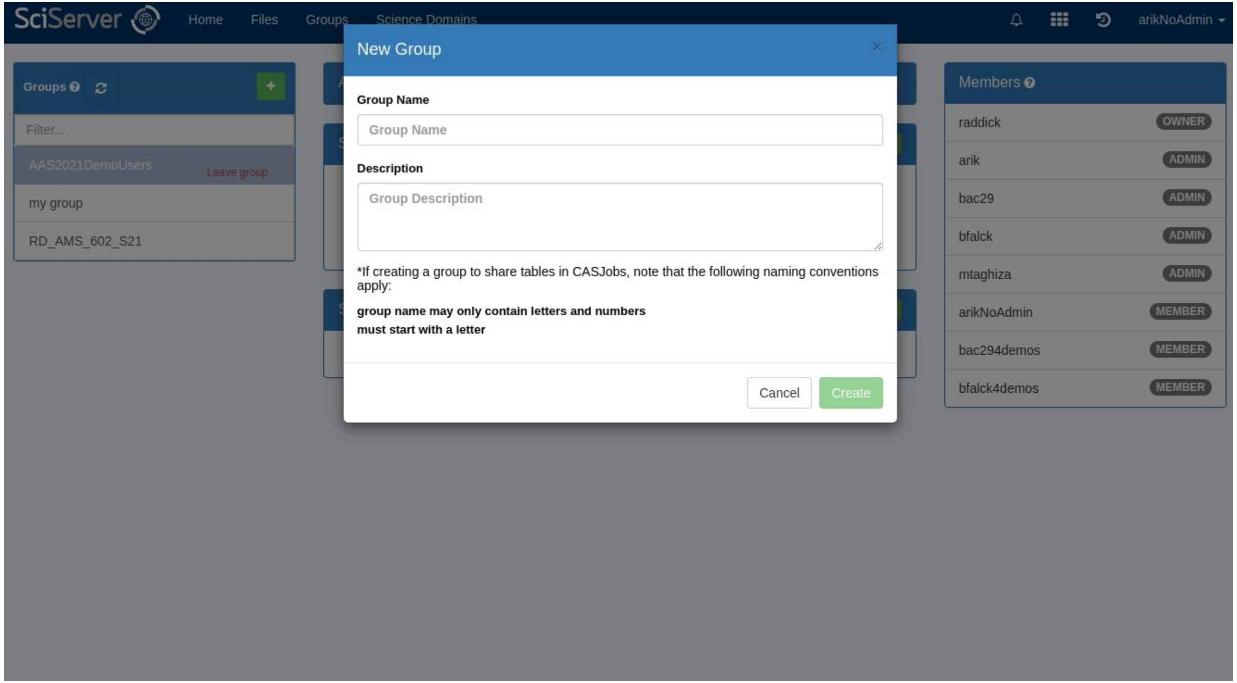
SciDrive
Drag-and-drop file hosting and sharing services.

SkyServer
Access the Sloan Digital Sky Survey data, tutorials and educational materials.

SkyQuery
A scalable database system for cross-matching astronomical source catalogs.

# Basic Usage - Groups

- Create groups
- Groups are a unit of sharing
- Invite other users



New Group

Group Name

Group Name

Description

Group Description

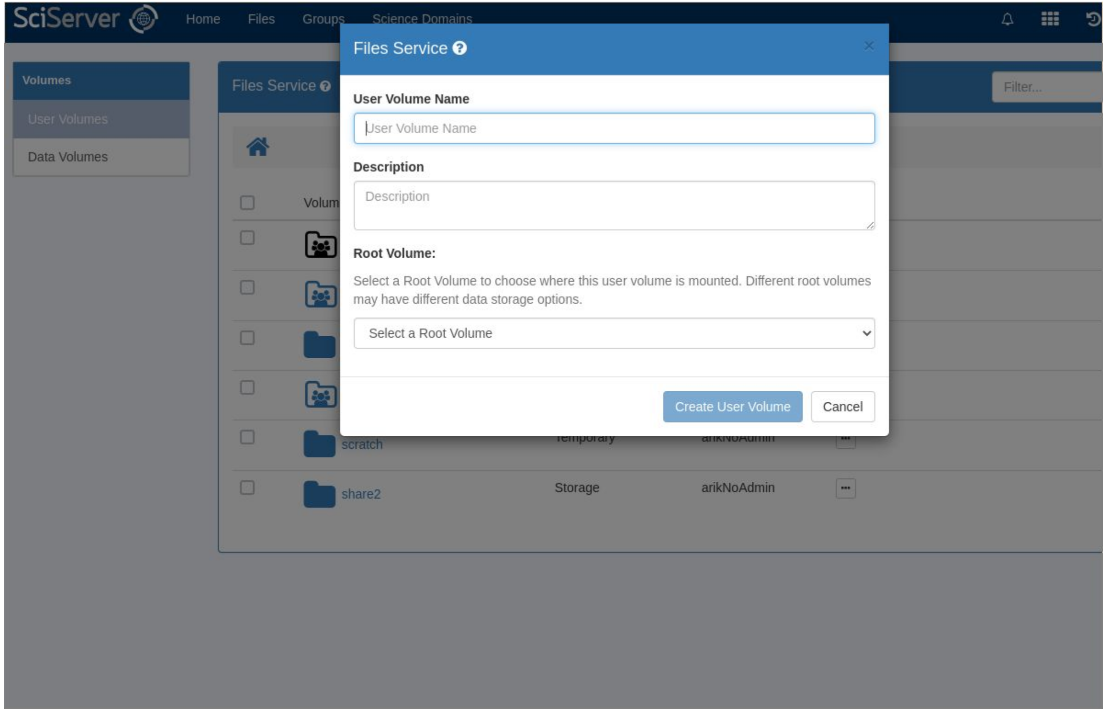
*If creating a group to share tables in CASJobs, note that the following naming conventions apply:

**group name may only contain letters and numbers**
**must start with a letter**

Cancel | Create

# Basic Usage - User Volumes

- Create “User Volumes”
- Volumes can be shared
- Store your own data
- Default limited space
- Temporary space:
  - For intermediate results
  - “Unlimited”
  - time-to-live



Files Service

User Volume Name

User Volume Name

Description

Description

Root Volume:

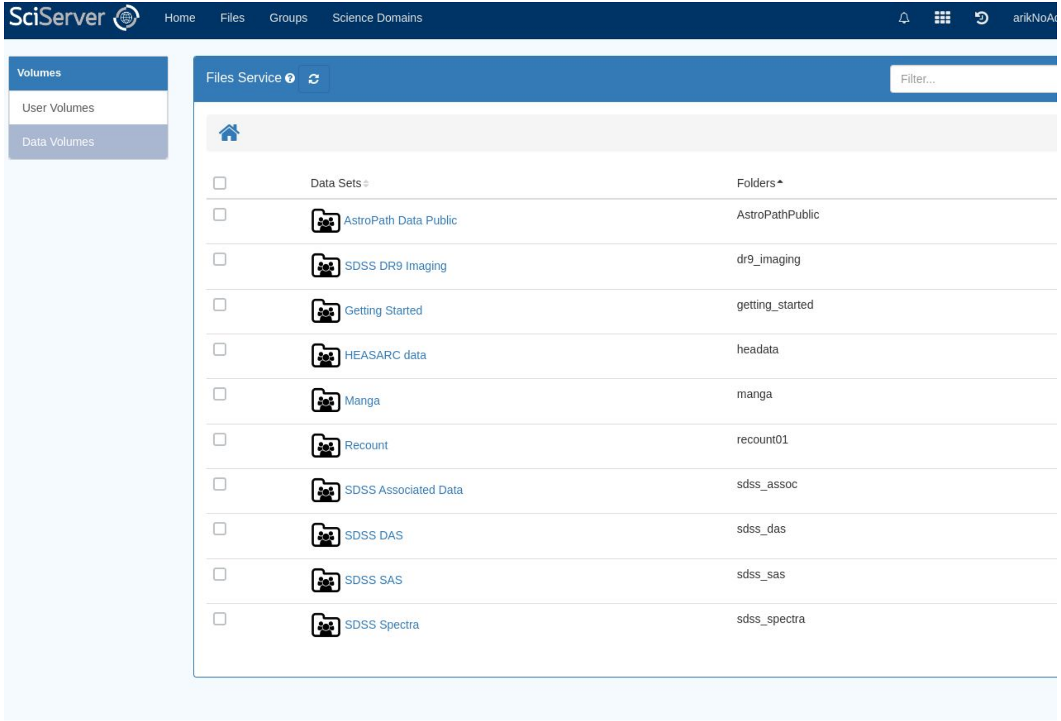
Select a Root Volume to choose where this user volume is mounted. Different root volumes may have different data storage options.

Select a Root Volume

Create User Volume | Cancel

# Basic Usage - Data Volumes

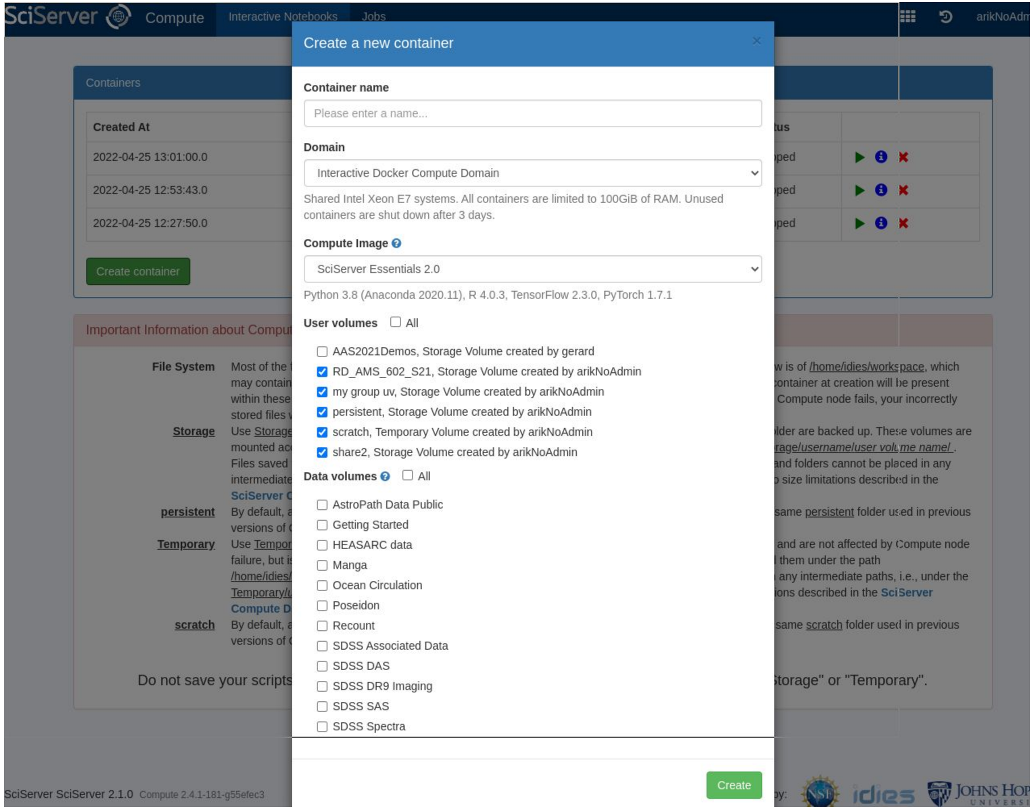
- “Data Volumes”
- Read-only
- Curated data (typically)
- Shared by “admins”



| □ | Data Sets | Folders |
|---|---|---|
| □ | AstroPath Data Public | AstroPathPublic |
| □ | SDSS DR9 Imaging | dr9_imaging |
| □ | Getting Started | getting_started |
| □ | HEASARC data | headata |
| □ | Manga | manga |
| □ | Recount | recount01 |
| □ | SDSS Associated Data | sdss_assoc |
| □ | SDSS DAS | sdss_das |
| □ | SDSS SAS | sdss_sas |
| □ | SDSS Spectra | sdss_spectra |

# Basic Usage - Compute

- Create a “container”
- Backed by docker
- Web service listening at specific port (Jupyter)
- Mount volumes needed



**Create a new container**

**Container name**

Please enter a name...

**Domain**

Interactive Docker Compute Domain

Shared Intel Xeon E7 systems. All containers are limited to 100GiB of RAM. Unused containers are shut down after 3 days.

**Compute Image**

SciServer Essentials 2.0

Python 3.8 (Anaconda 2020.11), R 4.0.3, TensorFlow 2.3.0, PyTorch 1.7.1

**User volumes** □ All

- □ AAS2021Demos, Storage Volume created by gerard
- ☑ RD_AMS_602_S21, Storage Volume created by arikNoAdmin
- ☑ my group uv, Storage Volume created by arikNoAdmin
- ☑ persistent, Storage Volume created by arikNoAdmin
- ☑ scratch, Temporary Volume created by arikNoAdmin
- ☑ share2, Storage Volume created by arikNoAdmin

**Data volumes** □ All

- □ AstroPath Data Public
- □ Getting Started
- □ HEASARC data
- □ Manga
- □ Ocean Circulation
- □ Poseidon
- □ Recount
- □ SDSS Associated Data
- □ SDSS DAS
- □ SDSS DR9 Imaging
- □ SDSS SAS
- □ SDSS Spectra

Create

**File System** Most of the f … w is of /home/idies/workspace, which may contain … container at creation will be present within these … Compute node fails, your incorrectly stored files …

**Storage** Use Storage … lder are backed up. These volumes are mounted ac … rage/username/user volume name/ . Files saved … and folders cannot be placed in any intermediate … size limitations described in the SciServer C…

**persistent** By default, a … same persistent folder used in previous versions of …

**Temporary** Use Tempor … and are not affected by Compute node failure, but is … them under the path /home/idies/ … any intermediate paths, i.e., under the Temporary/u … ions described in the SciServer Compute D…

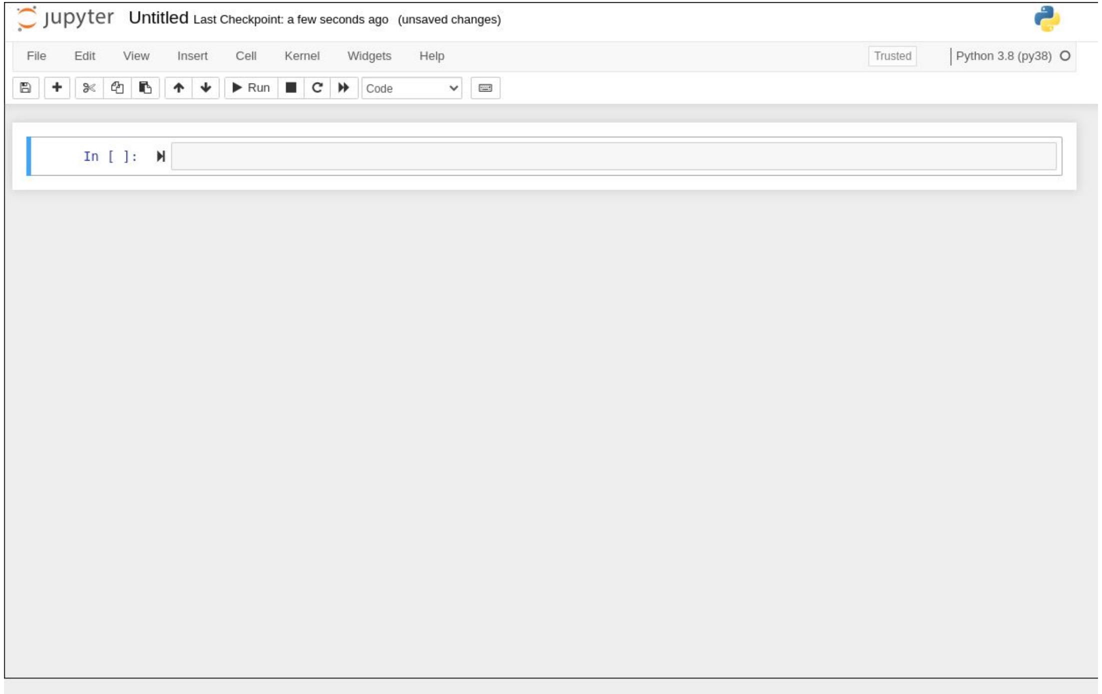
**scratch** By default, a … same scratch folder used in previous versions of …

Do not save your scripts … Storage" or "Temporary".

SciServer SciServer 2.1.0 Compute 2.4.1-181-g55efec3

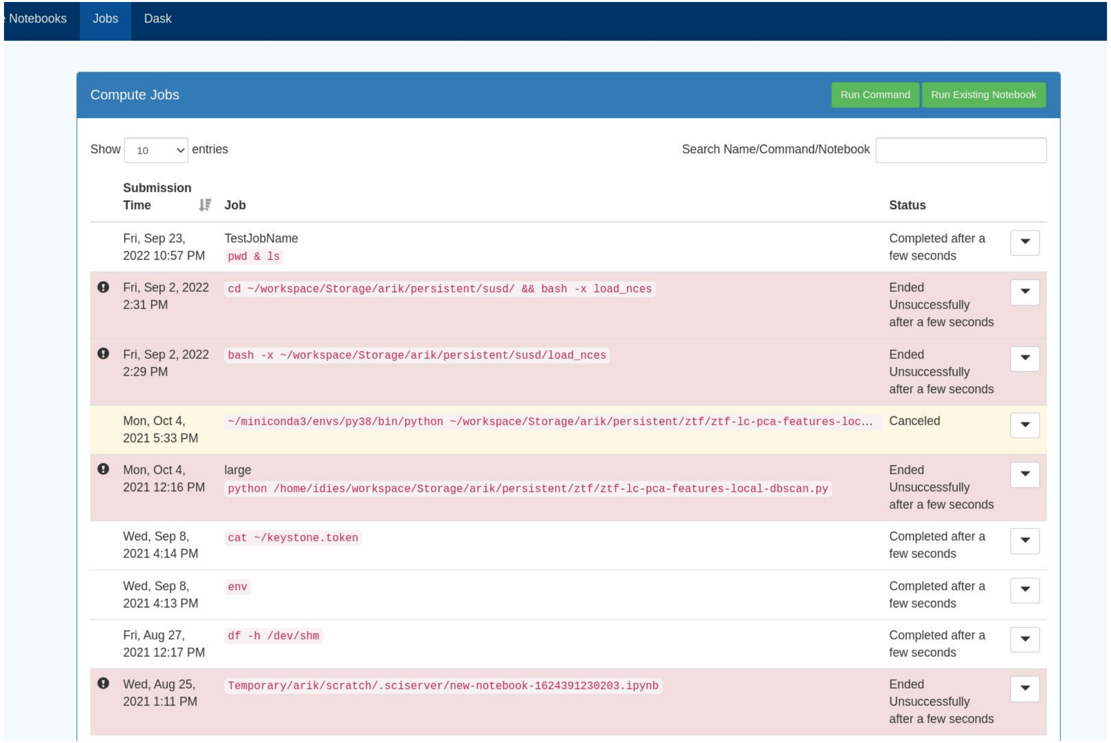
# Basic Usage - Compute

- Do analysis
- Volumes are mounted
  - Posix access to regular files
- Lots of software already included
- Token “injected” for API access to SciServer and CasJobs



# Basic Usage - Jobs

- Asynchronous Jobs
- Time-limited
- Large/specialized resources
- Queued, fair scheduling
- Clearer provenance



Compute Jobs

Show 10 entries

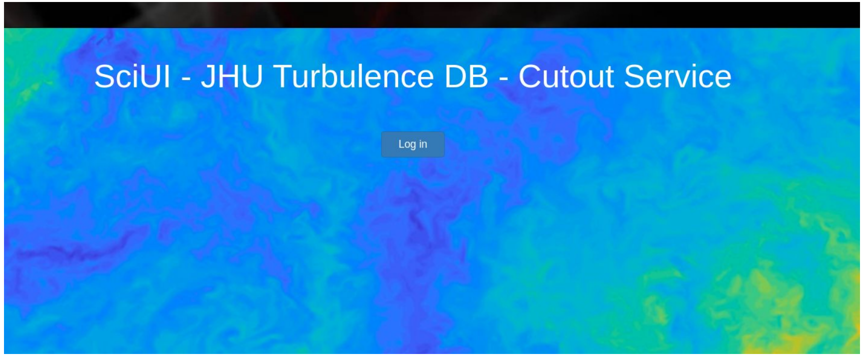
Search Name/Command/Notebook

| | Submission Time | Job | Status |
|---|---|---|---|
| | Fri, Sep 23, 2022 10:57 PM | TestJobName<br>`pwd & ls` | Completed after a few seconds |
| ! | Fri, Sep 2, 2022 2:31 PM | `cd ~/workspace/Storage/arik/persistent/susd/ && bash -x load_nces` | Ended Unsuccessfully after a few seconds |
| ! | Fri, Sep 2, 2022 2:29 PM | `bash -x ~/workspace/Storage/arik/persistent/susd/load_nces` | Ended Unsuccessfully after a few seconds |
| | Mon, Oct 4, 2021 5:33 PM | `~/miniconda3/envs/py38/bin/python ~/workspace/Storage/arik/persistent/ztf/ztf-lc-pca-features-loc...` | Canceled |
| ! | Mon, Oct 4, 2021 12:16 PM | large<br>`python /home/idies/workspace/Storage/arik/persistent/ztf/ztf-lc-pca-features-local-dbscan.py` | Ended Unsuccessfully after a few seconds |
| | Wed, Sep 8, 2021 4:14 PM | `cat ~/keystone.token` | Completed after a few seconds |
| | Wed, Sep 8, 2021 4:13 PM | `env` | Completed after a few seconds |
| | Fri, Aug 27, 2021 12:17 PM | `df -h /dev/shm` | Completed after a few seconds |
| ! | Wed, Aug 25, 2021 1:11 PM | `Temporary/arik/scratch/.sciserver/new-notebook-1624391230203.ipynb` | Ended Unsuccessfully after a few seconds |

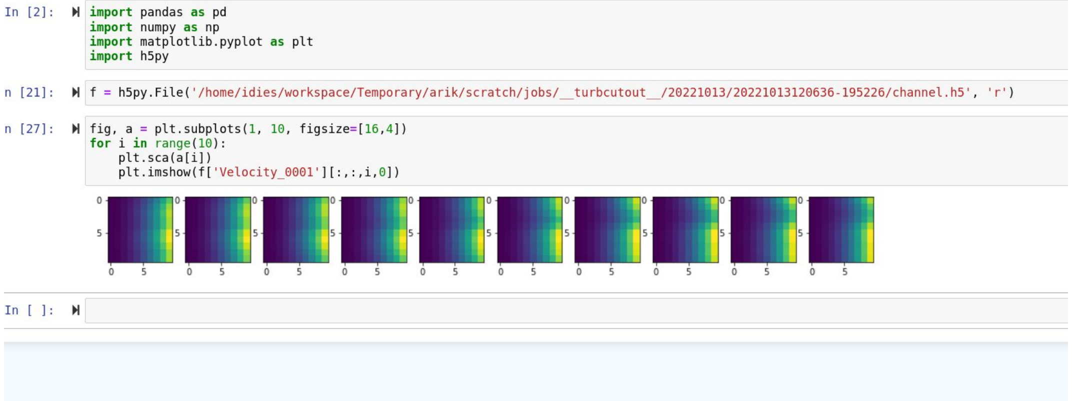
# Use Case Highlights

# Use Case: Turbulence

- Get cutouts of a large-scale turbulence simulation (~½ PB)
- Web interface, actions backed by SciServer jobs
- Use notebooks to analyze results



SciUI - JHU Turbulence DB - Cutout Service

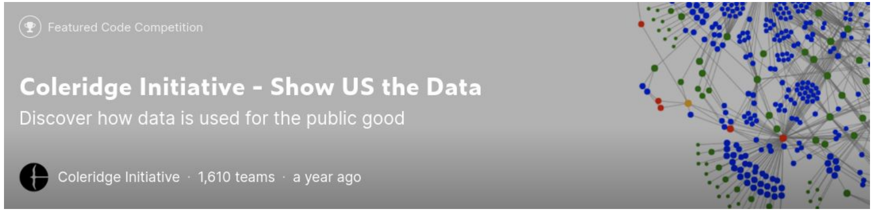


```
In [2]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import h5py

n [21]:  f = h5py.File('/home/idies/workspace/Temporary/arik/scratch/jobs/__turbcutout__/20221013/20221013120636-195226/channel.h5', 'r')

n [27]:  fig, a = plt.subplots(1, 10, figsize=[16,4])
         for i in range(10):
             plt.sca(a[i])
             plt.imshow(f['Velocity_0001'][:,:,i,0])

In [ ]:
```

http://turbulence.pha.jhu.edu/

# Use Case: Democratizing Data (Show Us The Data)

- Context: track dataset usage for gov agencies
- Rich data model, SQL backed
- Simple APIs for Tableu (hosted at TACC), but…
- SciServer enables deeper dive for those interested

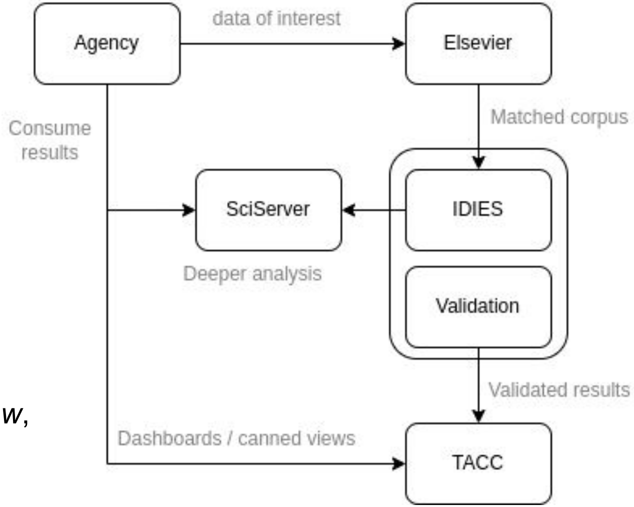


Featured Code Competition

**Coleridge Initiative - Show US the Data**

Discover how data is used for the public good

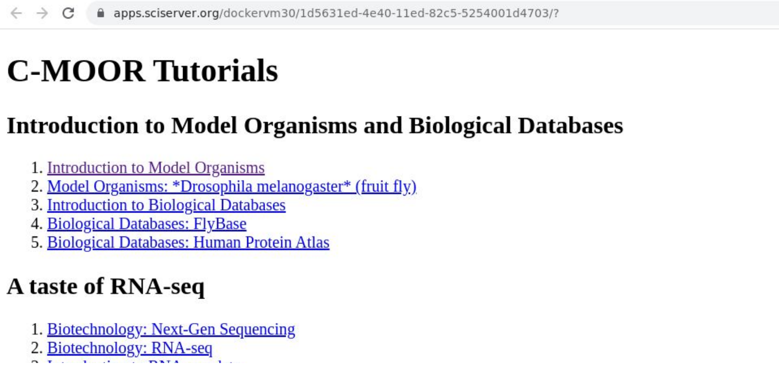
Coleridge Initiative · 1,610 teams · a year ago



See aso: Potok, N. (2022). Show US the Data. *Harvard Data Science Review*, *4*(2). https://doi.org/10.1162/99608f92.9d13ba15

# Use Case: Teaching

- Focus on course materials, not setup
- Share examples with students
- Immediate feedback and experimentation
- Collect materials via share


apps.sciserver.org/dockervm30/1d5631ed-4e40-11ed-82c5-5254001d4703/?

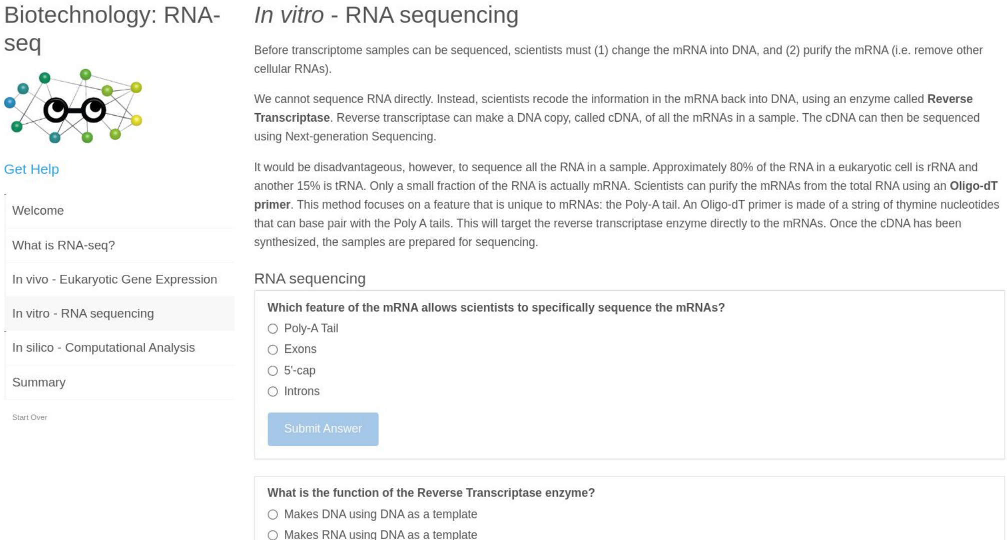
## C-MOOR Tutorials

### Introduction to Model Organisms and Biological Databases

1. Introduction to Model Organisms
2. Model Organisms: *Drosophila melanogaster* (fruit fly)
3. Introduction to Biological Databases
4. Biological Databases: FlyBase
5. Biological Databases: Human Protein Atlas

### A taste of RNA-seq

1. Biotechnology: Next-Gen Sequencing
2. Biotechnology: RNA-seq


Biotechnology: RNA-seq
Get Help
Welcome
What is RNA-seq?
In vivo - Eukaryotic Gene Expression
In vitro - RNA sequencing
In silico - Computational Analysis
Summary
Start Over

### *In vitro* - RNA sequencing

Before transcriptome samples can be sequenced, scientists must (1) change the mRNA into DNA, and (2) purify the mRNA (i.e. remove other cellular RNAs).

We cannot sequence RNA directly. Instead, scientists recode the information in the mRNA back into DNA, using an enzyme called **Reverse Transcriptase**. Reverse transcriptase can make a DNA copy, called cDNA, of all the mRNAs in a sample. The cDNA can then be sequenced using Next-generation Sequencing.

It would be disadvantageous, however, to sequence all the RNA in a sample. Approximately 80% of the RNA in a eukaryotic cell is rRNA and another 15% is tRNA. Only a small fraction of the RNA is actually mRNA. Scientists can purify the mRNAs from the total RNA using an **Oligo-dT primer**. This method focuses on a feature that is unique to mRNAs: the Poly-A tail. An Oligo-dT primer is made of a string of thymine nucleotides that can base pair with the Poly A tails. This will target the reverse transcriptase enzyme directly to the mRNAs. Once the cDNA has been synthesized, the samples are prepared for sequencing.

RNA sequencing

**Which feature of the mRNA allows scientists to specifically sequence the mRNAs?**

○ Poly-A Tail
○ Exons
○ 5'-cap
○ Introns

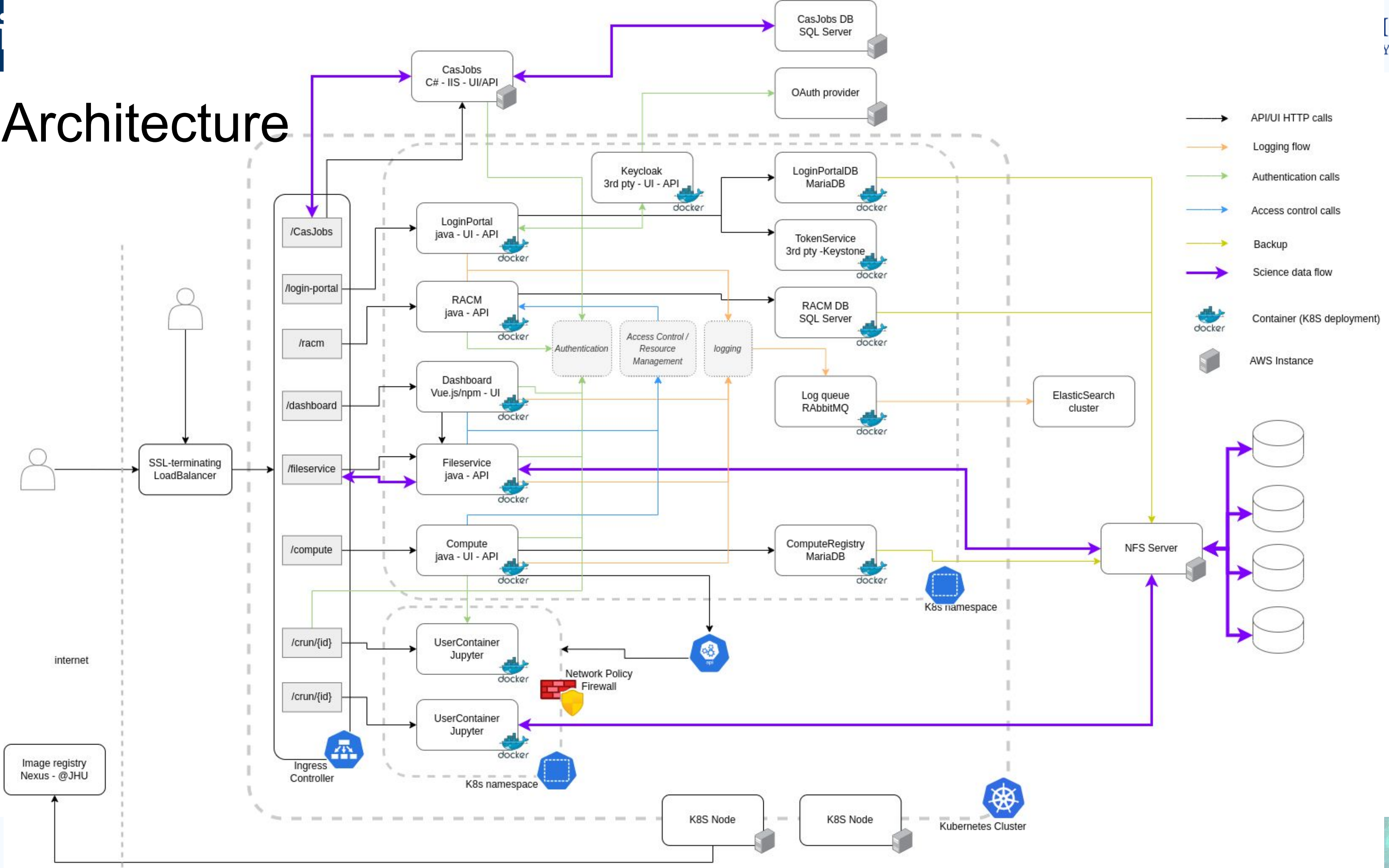
Submit Answer

**What is the function of the Reverse Transcriptase enzyme?**

○ Makes DNA using DNA as a template
○ Makes RNA using DNA as a template

# SciServer as a service

# Architecture



CasJobs DB SQL Server

CasJobs C# - IIS - UI/API

OAuth provider

Keycloak 3rd pty - UI - API

LoginPortalDB MariaDB

TokenService 3rd pty -Keystone

RACM DB SQL Server

Log queue RAbbitMQ

ElasticSearch cluster

LoginPortal java - UI - API

RACM java - API

Dashboard Vue.js/npm - UI

Fileservice java - API

Compute java - UI - API

ComputeRegistry MariaDB

UserContainer Jupyter

UserContainer Jupyter

Authentication

Access Control / Resource Management

logging

/CasJobs

/login-portal

/racm

/dashboard

/fileservice

/compute

/crun/{id}

/crun/{id}

Ingress Controller

SSL-terminating LoadBalancer

internet

Image registry Nexus - @JHU

Network Policy Firewall

K8s namespace

K8s namespace

K8S Node

K8S Node

Kubernetes Cluster

NFS Server

docker

API/UI HTTP calls

Logging flow

Authentication calls

Access control calls

Backup

Science data flow

Container (K8S deployment)

AWS Instance

# Deployments

- All the same features:
  - In another datacenter on own hardware
  - Behind a VPN
  - For specific insiders
  - In a HIPAA-safe environment
- A number of partners
- Active development

@JHU (sciserver.org)

*On premise*

@NIST

*AWS VPC, k8s*

@NAOJ

*On premise, k8s*

@MPE

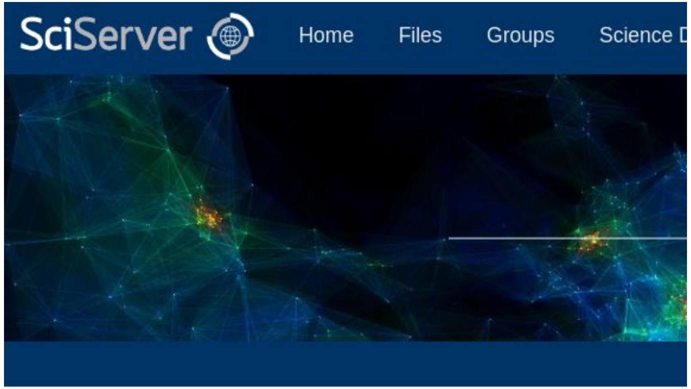
*On premise, k8s*

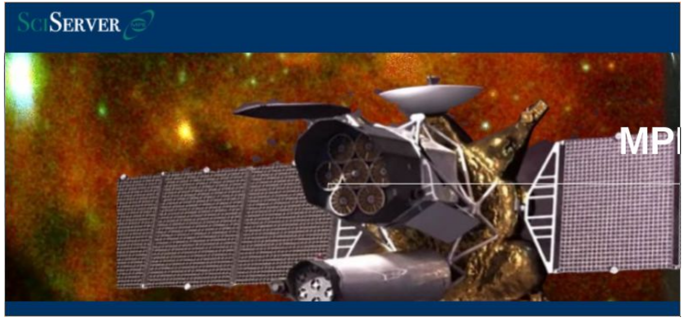
@Precision-Medic

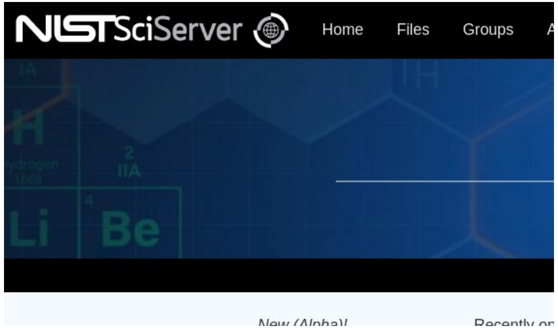
*Azure, secure-desktop*

# Deployments

- Branding
  - Theme customizations
  - links
  - etc







# Deployments

- Leverage Kubernetes and helm
- Bundle SciServer + 3rd party dependencies
- Configurable with a single file

```
$ helm install sciserver -f config.yaml sciserver-k8s/charts/sciserver
```

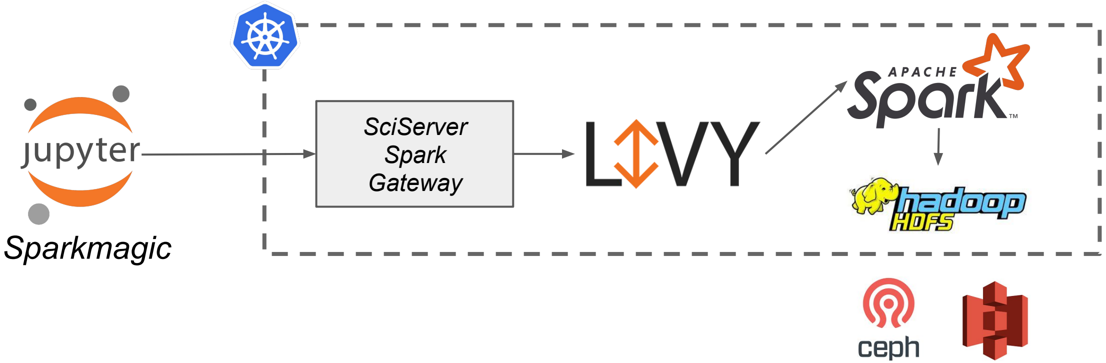
# Extra Features

# Extra features

- Accelerated computing
  - GPU
  - Spark
  - Dask
- Interactive and/or Jobs

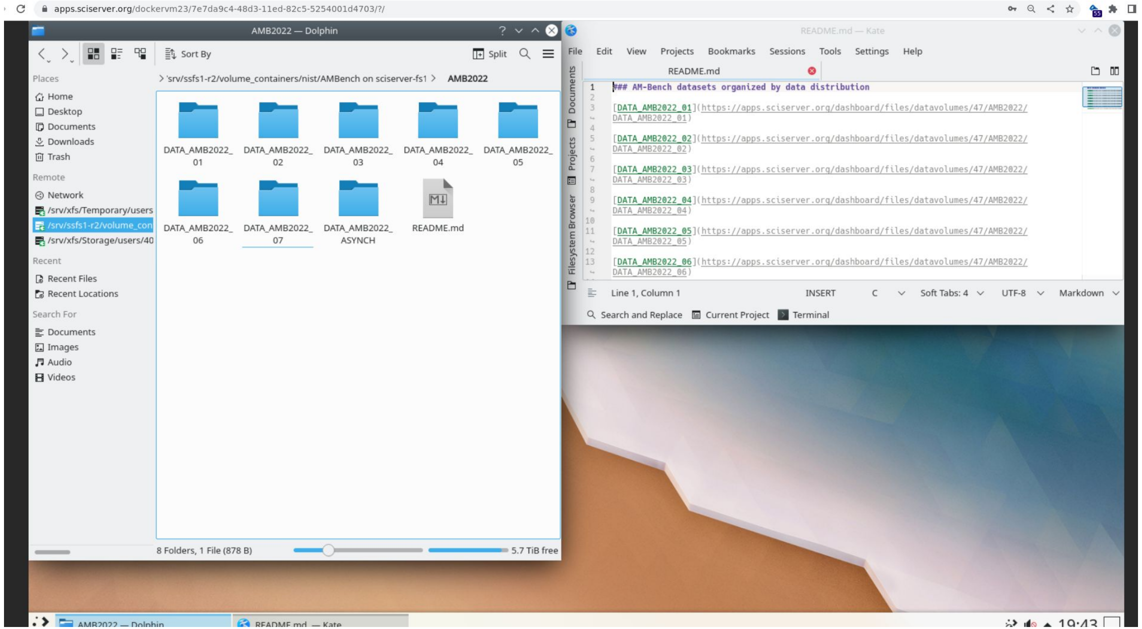
PyTorch TensorFlow

DASK



# Extra Features

- Not Just Jupyter
- Compute exposes service that listens to single port
- VNC, for example



apps.sciserver.org/dockervm23/7e7da9c4-48d3-11ed-82c5-5254001d4703/?/

AMB2022 — Dolphin

'srv/ssfs1-r2/volume_containers/nist/AMBench on sciserver-fs1 > AMB2022

8 Folders, 1 File (878 B)    5.7 TiB free

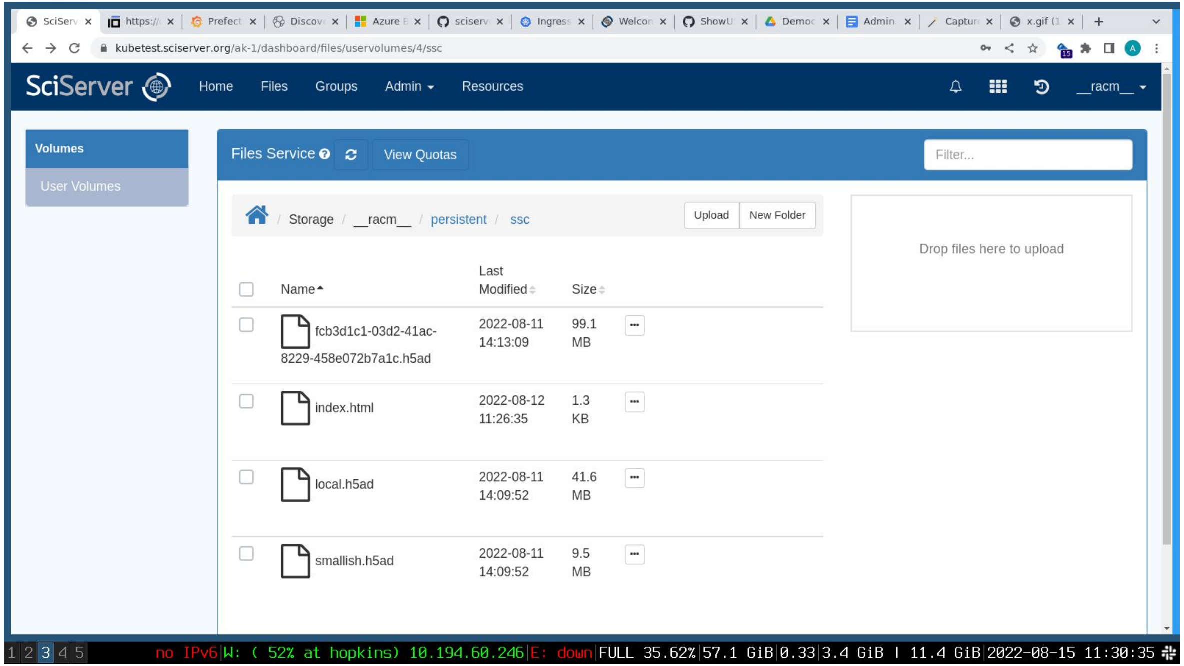
README.md — Kate

```
### AM-Bench datasets organized by data distribution

[DATA_AMB2022_01](https://apps.sciserver.org/dashboard/files/datavolumes/47/AMB2022/DATA_AMB2022_01)

[DATA_AMB2022_02](https://apps.sciserver.org/dashboard/files/datavolumes/47/AMB2022/DATA_AMB2022_02)

[DATA_AMB2022_03](https://apps.sciserver.org/dashboard/files/datavolumes/47/AMB2022/DATA_AMB2022_03)

[DATA_AMB2022_04](https://apps.sciserver.org/dashboard/files/datavolumes/47/AMB2022/DATA_AMB2022_04)

[DATA_AMB2022_05](https://apps.sciserver.org/dashboard/files/datavolumes/47/AMB2022/DATA_AMB2022_05)

[DATA_AMB2022_06](https://apps.sciserver.org/dashboard/files/datavolumes/47/AMB2022/DATA_AMB2022_06)
```

Line 1, Column 1    INSERT    C    Soft Tabs: 4    UTF-8    Markdown

# Extra Features

- "One-click apps"
- Shareable links
- Argument passing
  - Can do in one-click what would take steps using generic compute containers
- Operate on files
- Embedded feel



kubetest.sciserver.org/ak-1/dashboard/files/uservolumes/4/ssc

Storage / __racm__ / persistent / ssc

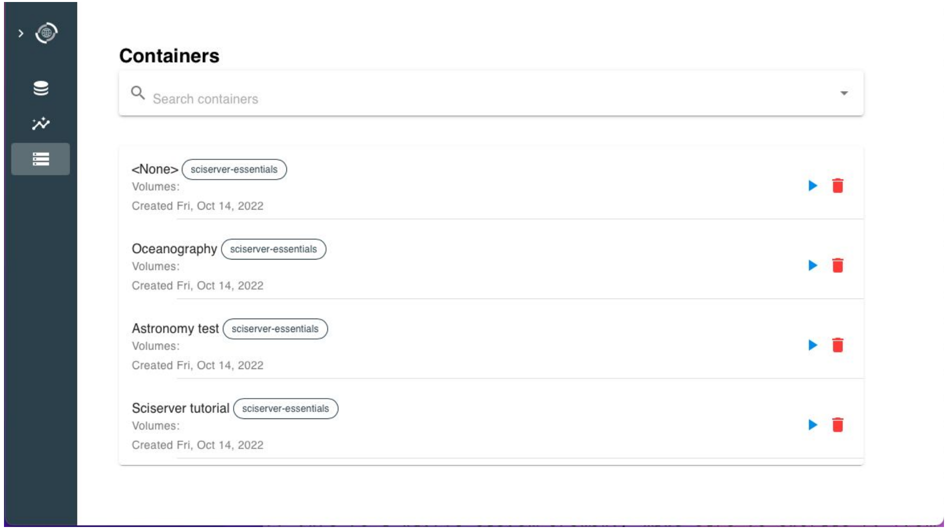
| Name | Last Modified | Size |
|---|---|---|
| fcb3d1c1-03d2-41ac-8229-458e072b7a1c.h5ad | 2022-08-11 14:13:09 | 99.1 MB |
| index.html | 2022-08-12 11:26:35 | 1.3 KB |
| local.h5ad | 2022-08-11 14:09:52 | 41.6 MB |
| smallish.h5ad | 2022-08-11 14:09:52 | 9.5 MB |

```
1 2 3 4 5   no IPv6|W: ( 52% at hopkins) 10.194.60.246|E: down|FULL 35.62%|57.1 GiB|0.33|3.4 GiB | 11.4 GiB|2022-08-15 11:30:35
```

# Future Directions

# Future Directions

- UI consolidation / improvements
  - Simplify development
  - Simplify branding and customizations
  - Improved layout and features



Containers

Search containers

<None> sciserver-essentials
Volumes:
Created Fri, Oct 14, 2022

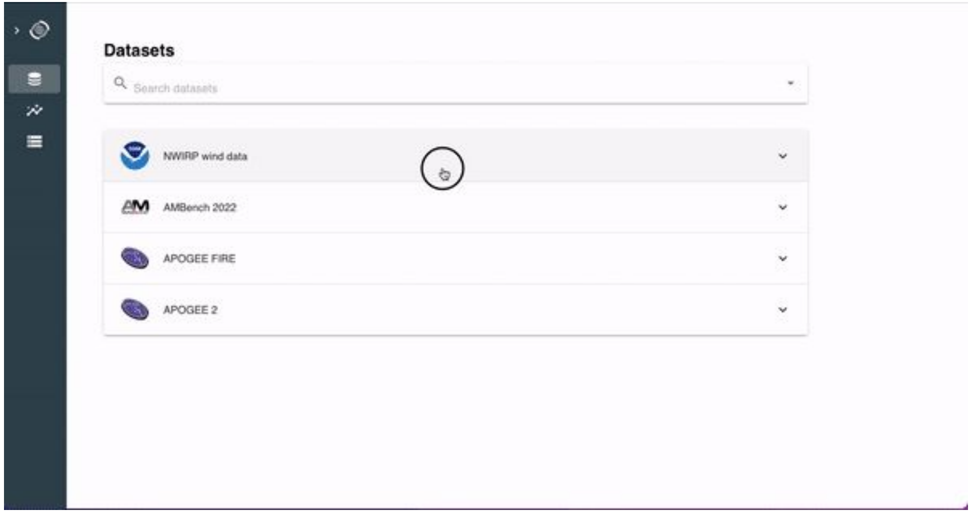
Oceanography sciserver-essentials
Volumes:
Created Fri, Oct 14, 2022

Astronomy test sciserver-essentials
Volumes:
Created Fri, Oct 14, 2022

Sciserver tutorial sciserver-essentials
Volumes:
Created Fri, Oct 14, 2022

# Future Directions

- UI consolidation / improvements
- Data discoverability
    - ○ “Documents close to the data”
    - ○ Leverage built-in access control
    - ○ Human and machine writable
    - ○ Just documentation, not full ontology



Datasets

Search datasets

NWIRP wind data

AMBench 2022

APOGEE FIRE

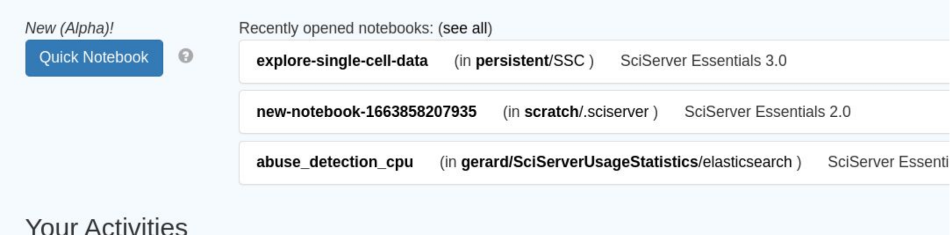
APOGEE 2

# Future Directions

Open in Colab

sciserver

- UI consolidation / improvements
- Data discoverability
- One-click simplicity / link-share
  - Combine with data discoverability
  - Notebooks, quicker start
  - Jupyter widgets based apps
  - Sends links, not wordy instructions



*New (Alpha)!*

Quick Notebook

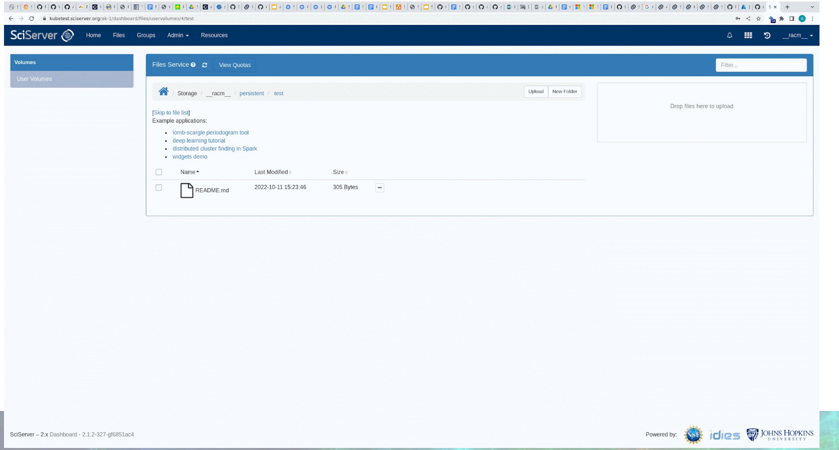
Recently opened notebooks: (see all)

**explore-single-cell-data** (in **persistent**/SSC ) SciServer Essentials 3.0

**new-notebook-1663858207935** (in **scratch**/.sciserver ) SciServer Essentials 2.0

**abuse_detection_cpu** (in **gerard**/SciServerUsageStatistics/elasticsearch ) SciServer Essenti

Your Activities


kubetest.sciserver.org/ak-1/dashboard/files/uservolumes/4/test
SciServer
Home
Files
Groups
Admin
Resources
__racm__
Volumes
User Volumes
Files Service
View Quotas
Filter...
Storage / __racm__ / persistent / test
Upload
New Folder
Drop files here to upload

[Skip to file list]
Example applications:

- lomb-scargle periodogram tool
- deep learning tutorial
- distributed cluster finding in Spark
- widgets demo

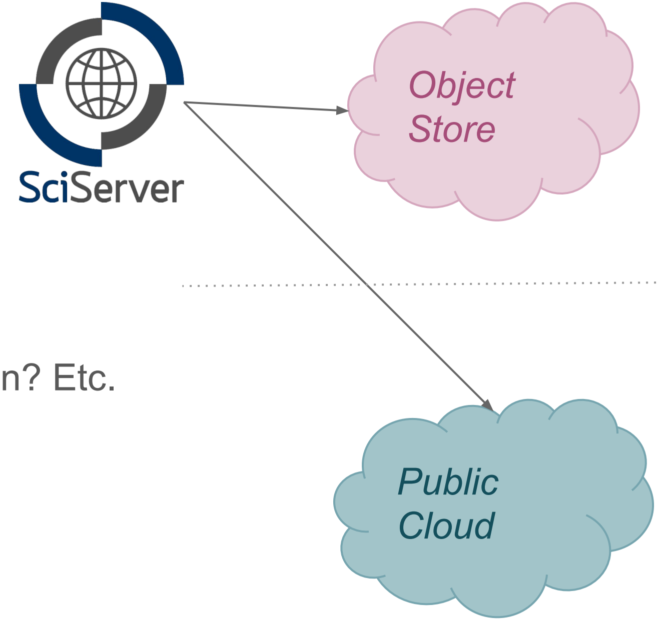
| | Name | Last Modified | Size |
|---|---|---|---|
| □ | README.md | 2022-10-11 15:23:46 | 305 Bytes |

SciServer – 2.x Dashboard - 2.1.2-327-gf6851ac4

Powered by:

# Future Directions

- UI consolidation / improvements
- Data discoverability
- One-click simplicity / link-share
- Cloud integrations
  - Object storage integration
  - Compute: site-wide? Research group? Bring-your-own? Etc.
  - Cloud-only datasets



# Future Directions

- UI consolidation / improvements
- Data discoverability
- One-click simplicity / link-share
- Cloud integrations
- Others on the radar
  - Extended login sessions / api keys
  - Scheduled / cron-like jobs
  - Anonymous file sharing

# Thanks!

SciServer