

Scientific Platforms and Astronomical Data Access in the Era of (Scientific) Cloud Computing

Matias Carrasco Kind
Senior Research Scientist, NCSA/UIUC
Data Release Scientist, DES

Outline



- What does Data Access mean?
- Scientific Platforms and Gateways
- The Notebook revolution
- Scientific Cloud computing
- Containerization
- Kubernetes
- Applications

What is a Data Release?

Data Products

Interfaces

Documentation

Support

What is a Data Release?

Data Products

Preparation
Vetting
Versioning
Consistency
Integrity
Redundancy
Data Model
Storage
Backups
Recovery
Hardware

Interfaces

Development
Version control
Licenses
Data Access
Languages
Sustainability
Guidelines
Scalability
Deployment
Hardware
Maintenance

Documentation

Papers
Web
Code
Data Model
Data Access
Data Format
Guidelines
Accessible
Maintenance
Contributions

Support

Short Term
Long Term
Forum
Help
Understanding
Deployment
Privacy
Maintenance
Focused
Distributed

What is a Data Release?

Data Products

Preparation
Vetting
Versioning
Consistency
Integrity
Redundancy
Data Model
Storage
Backups
Recovery
Hardware

Interfaces

Development
Version control
Licenses
Data Access
Languages
Sustainability
Guidelines
Scalability
Deployment
Hardware
Maintenance

Documentation

Papers
Web
Code
Data Model
Data Access
Data Format
Guidelines
Accessible
Maintenance
Contributions

Support

Short Term
Long Term
Forum
Help
Understanding
Deployment
Privacy
Maintenance
Focused
Distributed

What is Data Access?

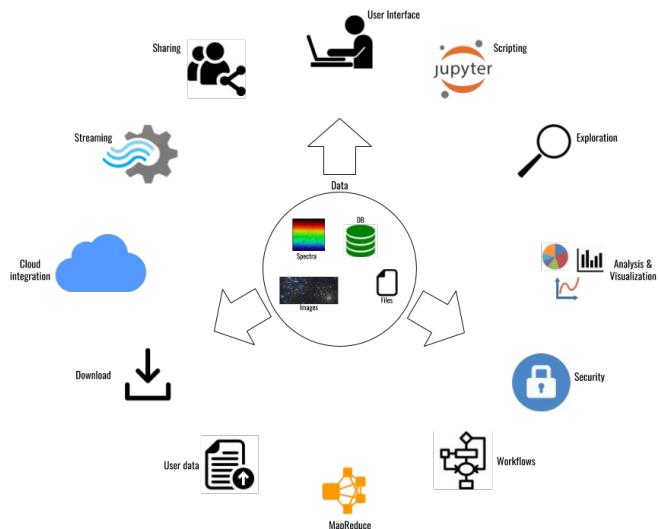


Several meanings around a central data archive, a.k.a “data lake”, repository with common components

- Storage
- Security
- Retrieving
- Interacting
- Modifying
- Understanding

Scientific Platforms and Gateways

... and many of these concepts are also associated with Scientific Platforms and Gateways (and Science portals, Science servers, etc.)



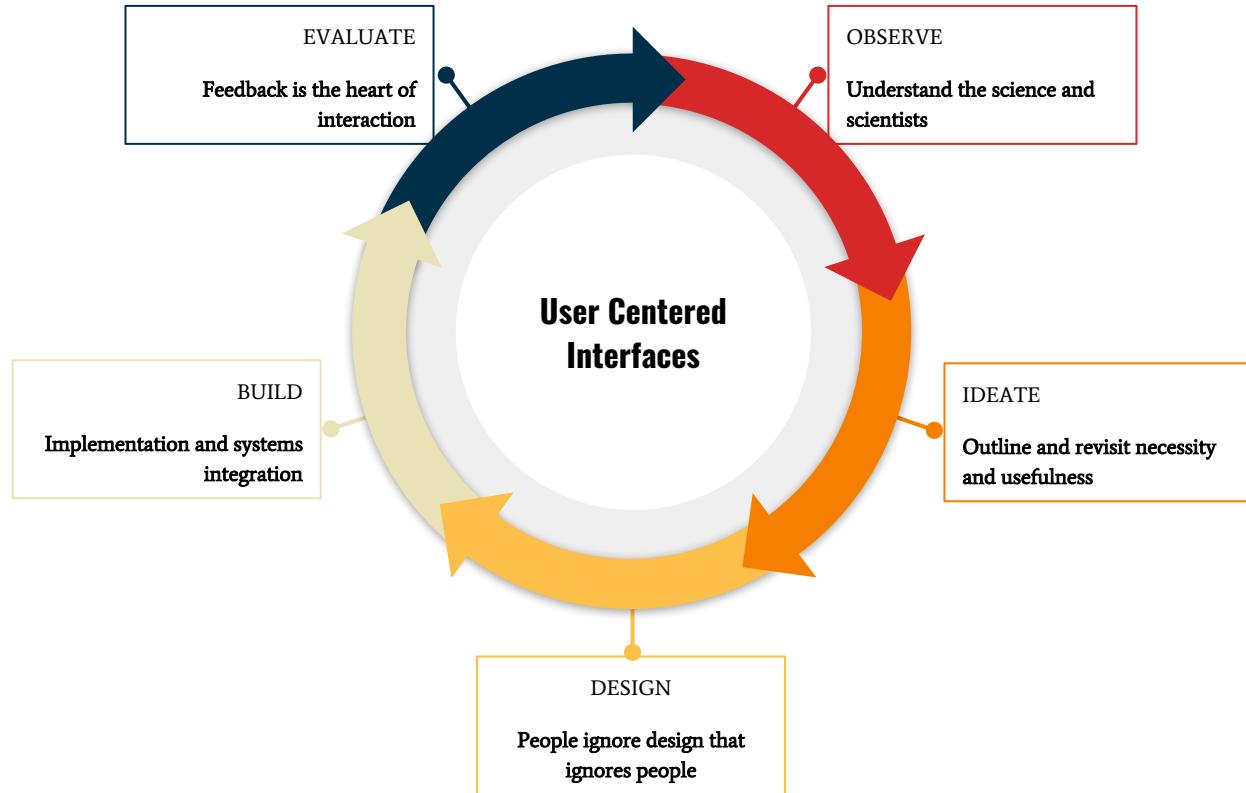
“Science gateways allow science & engineering communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines.”
(Science Gateways Institute)

“Science gateways is a place to do collaborative scientific related activities” (Me)

User (Scientist) Centered Design

Data Access would not exists without a user interface, but will only succeed if it is user driven.

“... In an ideal world, a user would remember every function after only a single use, but we do not live in idealism. The reality is that familiarity and intuition must be consciously designed into the interface”

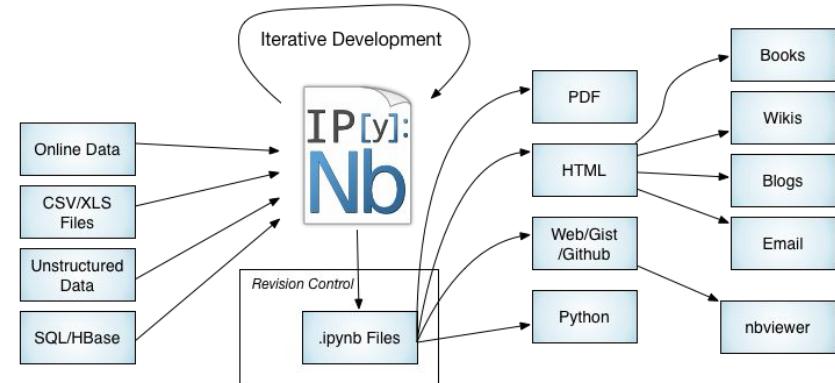
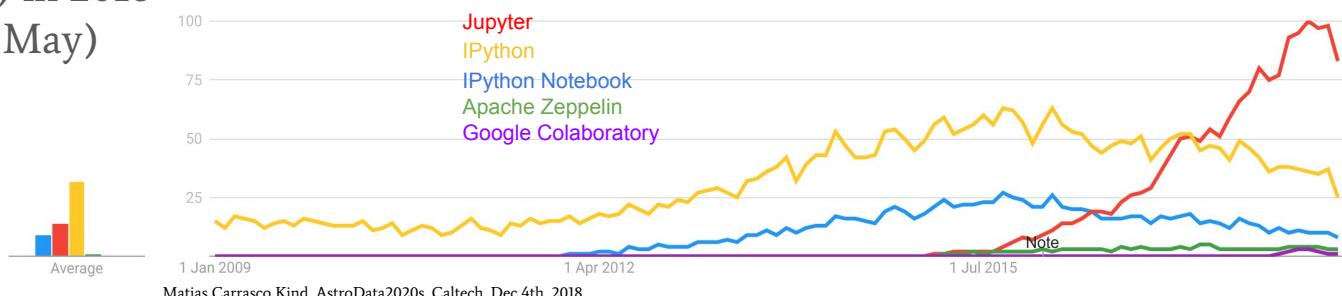


The Notebook Revolution



The Notebook Development

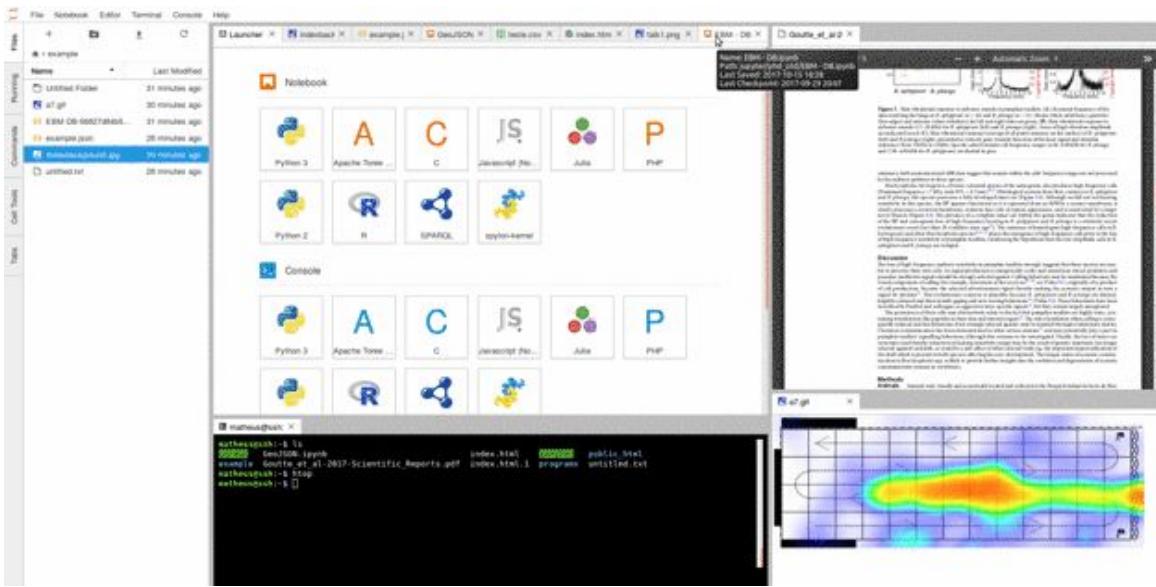
- Started from ideas like Matlab, Maple or Mathematica ~1988
- IPython has been around since 2001
- Sage Notebook released in 2005 (uses IPython)
- IPython Notebook was released in 2011
- IPython Notebook moved to Jupyter in 2014
- Apache Zeppelin created in 2015 (JVM and integrated with Apache Products)
- Beaker Notebook 2015 (moved to BeakerX)
- Google Colaboratory released in Oct 2017 (from ideas back in 2014)
- Cocalc (by SageMath) in 2018
- Jupyter Lab Beta 2.0 (May)



The Jupyter Notebook



- Computational narrative
- Scripting interface
- Scientific oriented interface
- Customizable
- Collaborative
- Adopted by many projects in scientific fields
- Widgets
- Big Data Integration (Spark)
- Interactive plots
- Multiple Kernels (Python, R, Julia, Scala, etc.)



The Jupyter Notebook



- Computational narrative
- Scripting interface
- Scientific oriented interface
- Customizable
- Collaborative
- Adopted by many projects in scientific fields
- Widgets
- Big Data Integration (Spark)
- Interactive plots
- Multiple Kernels (Python, R, Julia, Scala, etc.)

A screenshot of the Jupyter Notebook interface. On the left, a sidebar shows a file tree with several notebooks and a folder named 'Untitled Folder'. The main area contains a code cell with the following CSS and JavaScript code:

```
1 <!DOCTYPE html>
2 <html>
3 <head>
4 <meta charset="utf-8">
5 <meta name="viewport" content="width=device-width">
6 <title>Interactive Media with canvas effects</title>
7 <link rel="stylesheet" href="css/htmlbeautify.css">
8 <script src="js/htmlbeautify.js"></script></head>
9
10 <section id="wrapper">
11   <script type="text/javascript" src="http://cdn.carbonads.com/carbon.js?zoneid=1&serve=ck4DkXNgIaCmz"></script>
12 </section>
13
14 <style>
15   body {
16     background: #FFF;
17     color: #333;
18     padding: 10px;
19   }
20
21   /* I'm using CSS3 to translate the video on the R axis to give it a mirror effect */
22   #video {
23     display: block;
24     margin: 20px;
25     max-width: 100px;
26   }
27
28 .supported &video {
29   -webkit-transform: rotate(180deg) rotate(360deg, 0, 0, 0deg);
30   -o-transform: rotate(180deg) rotate(360deg, 0, 0, 0deg);
31   -ms-transform: rotate(180deg) rotate(360deg, 0, 0, 0deg);
32   -moz-transform: rotate(180deg) rotate(360deg, 0, 0, 0deg);
33   transform: rotate(180deg) rotate(360deg, 0, 0, 0deg);
34 }
```

Below the code cell is a terminal window titled 'mathesar@mathesar' showing command-line output:

```
1 [mathesar@mathesar ~]$ python3 test.py
2 [mathesar@mathesar ~]$ TaskID: 44_29_07_0_0_0
3 Load Average: 0.00 0.00 0.00
4 CPU: 0.00:0.00
5 GPU: 0.00:0.00
6 Memory: 0.00:0.00
7 Total: 0.00:0.00
8 [mathesar@mathesar ~]$
```

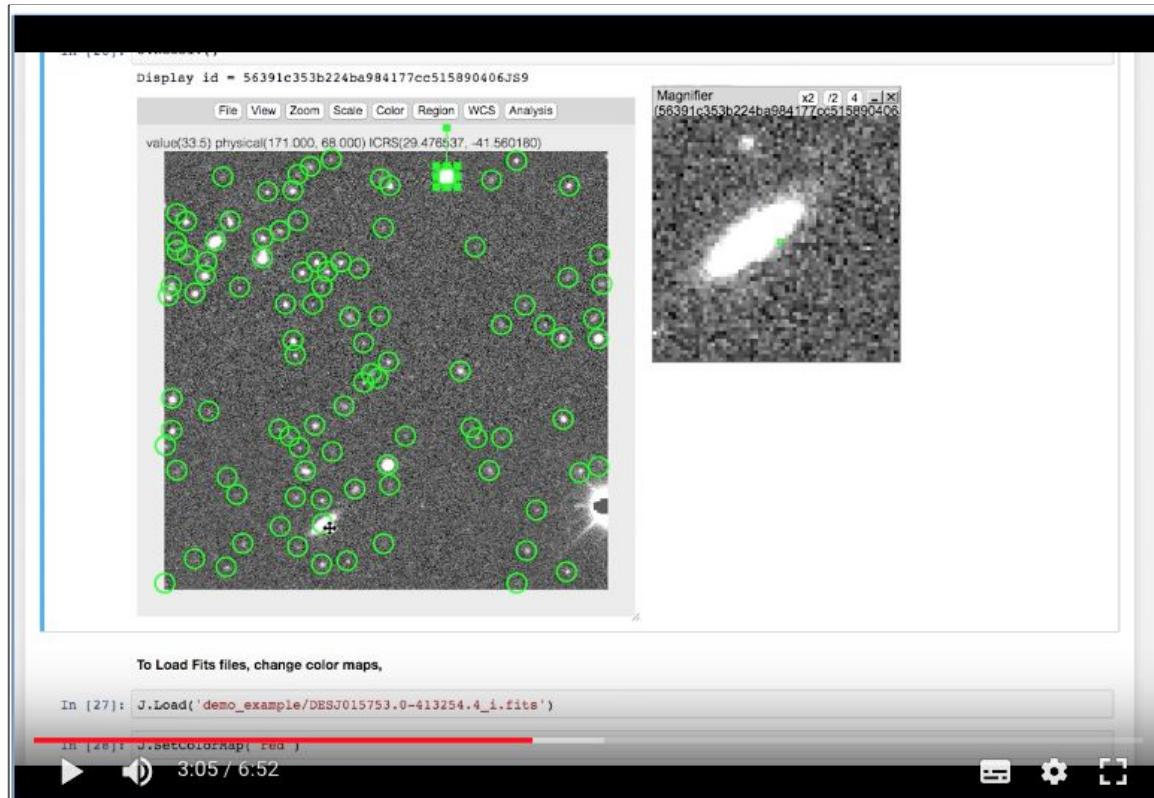
To the right of the code cell is a heatmap plot showing a red and blue color gradient over a grid.

Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers

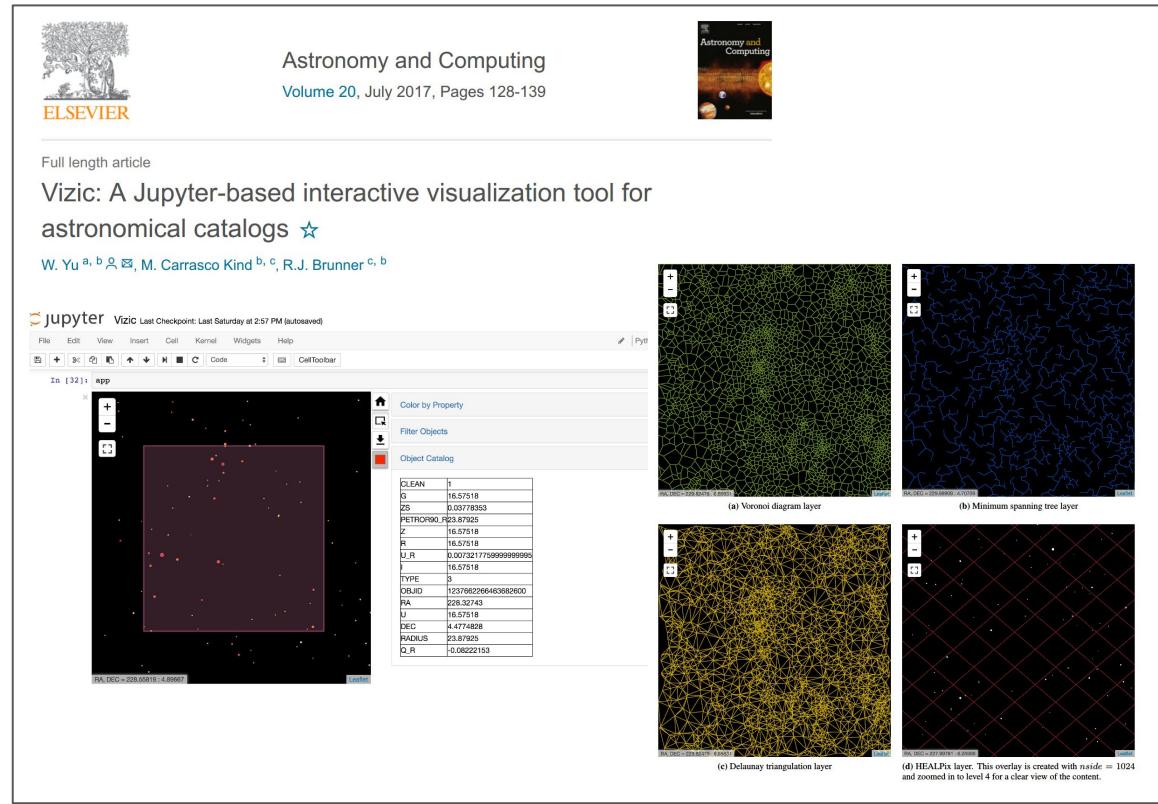
Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers



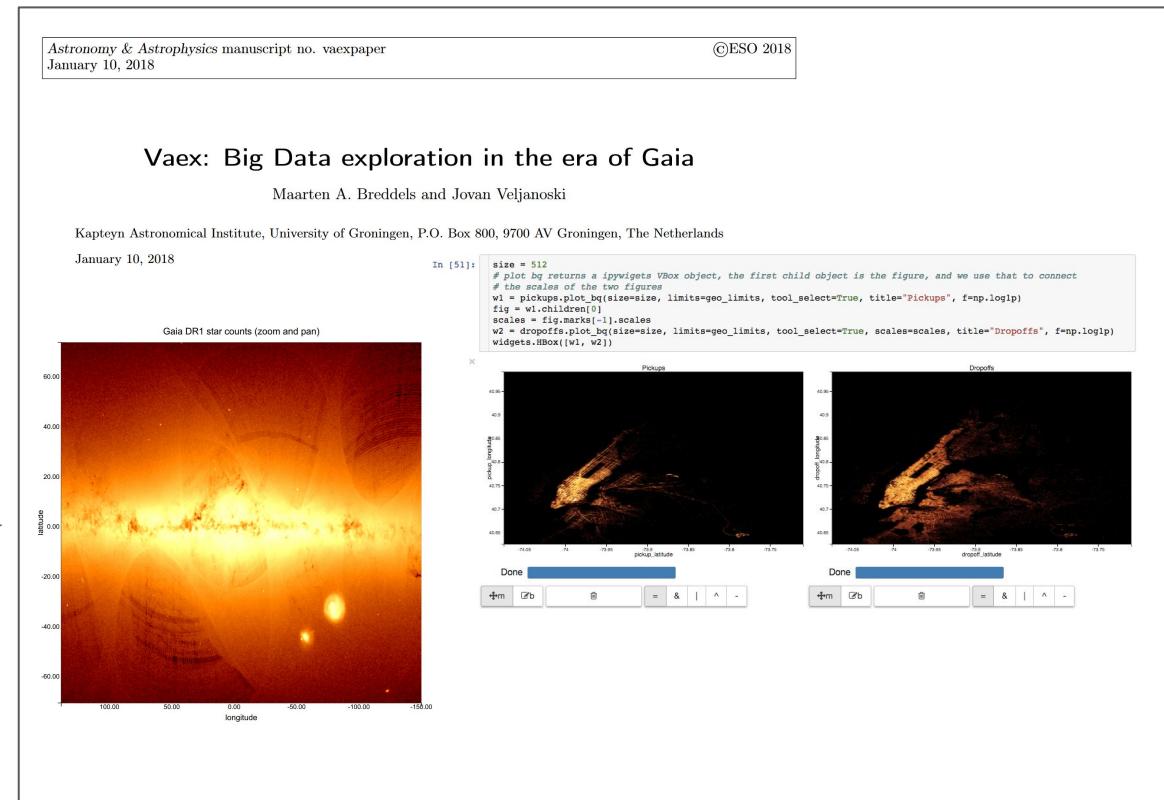
Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers



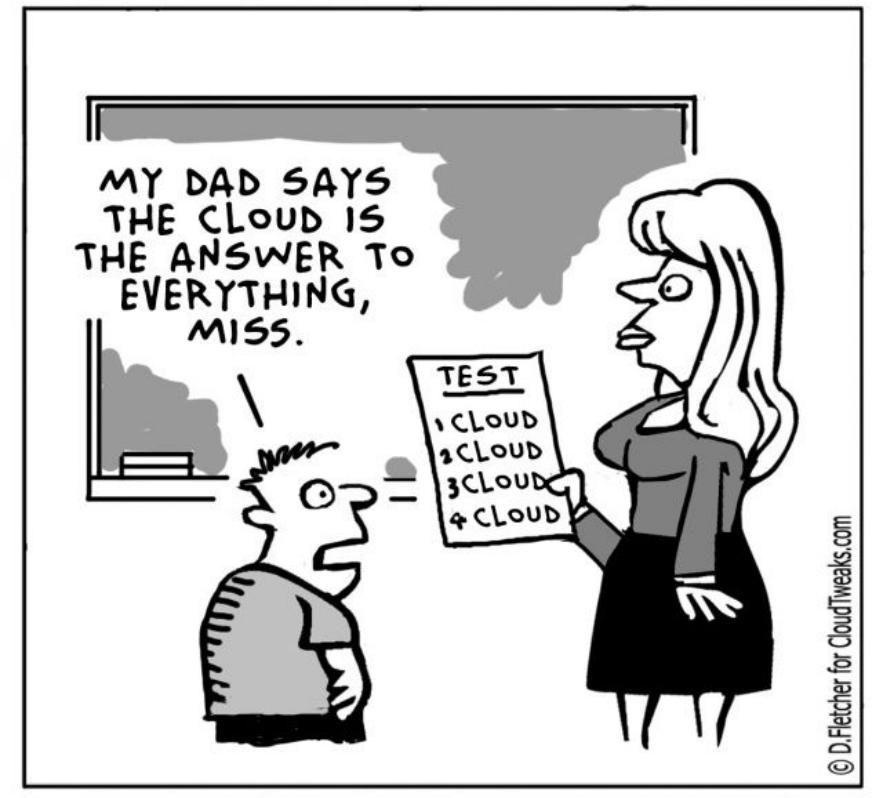
Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers



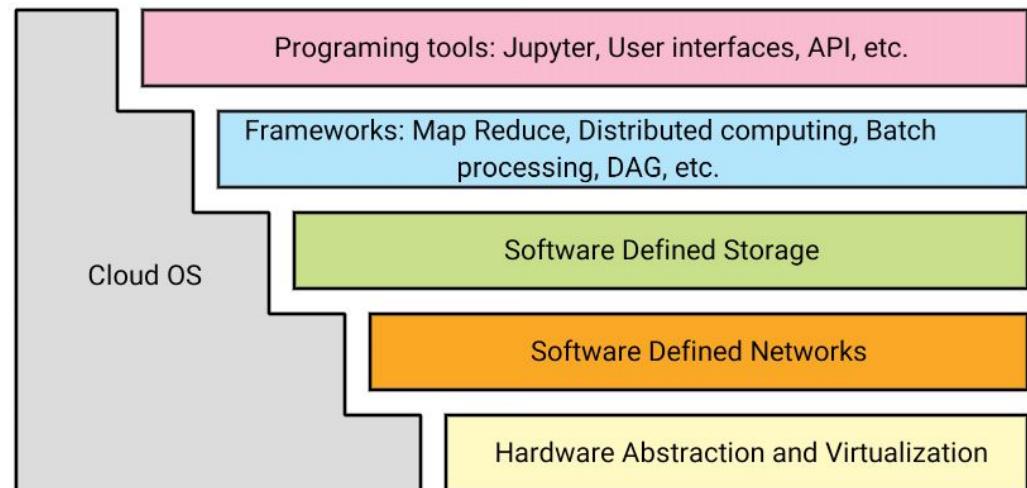
Scientific Cloud Computing

Cloud is about how you do computing, not where you do computing.



Why we should be doing science on the cloud

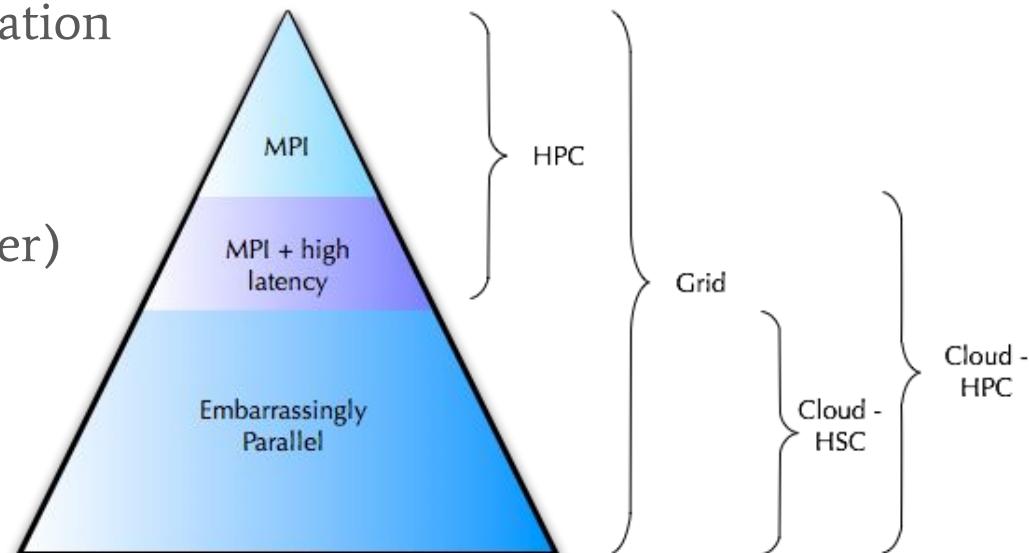
- Remote and dynamic data (\neq Big data)
- Big data \Rightarrow Data Gravity, Data Lake, etc.
- Remote software/server
- Easy to deploy*
- Asynchronous
- Web applications / Shareable
- Serverless applications
- Federation of Services
- Tablets/ChromeOS
- more...



*arguable

Why we shouldn't be doing science on the cloud

- Because there is no a real reason for it[^]
- HPC is not there yet, large latencies and bad bisection bandwidth
... but HPC is adopting cloud technologies
- Full control on data and application
- Security concerns*
- Faster development*
- Billing (if a commercial provider)
- more ...

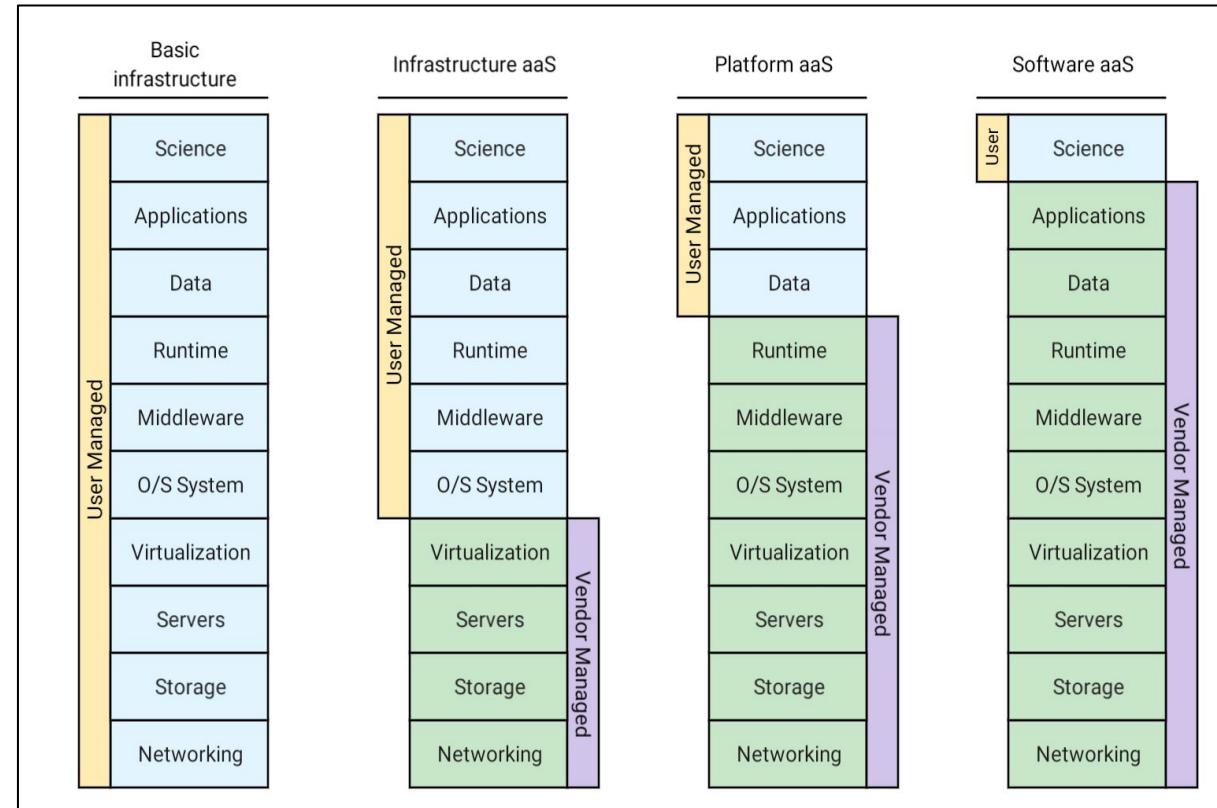


[^]arguable

*arguable (CI, CD)

What kind of science/projects? → Which model

- HTC vs HPC vs HSC
- Interactive
- Small projects
- Visualizations
- Short term projects*

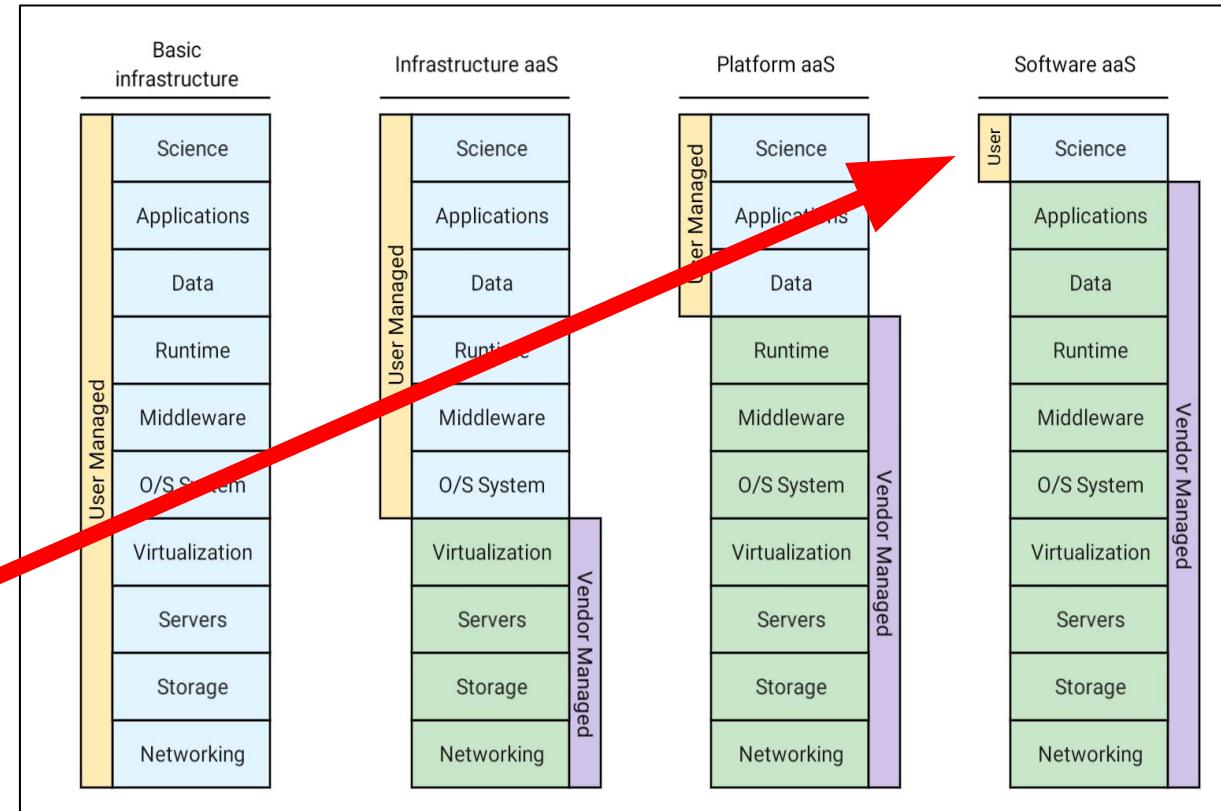


*arguable

What kind of science/projects? → Which model

- HTC vs HPC vs HSC
- Interactive
- Small projects
- Visualizations
- Short term projects*

Will we get to have Science as a Service (SClaaS?)



*arguable

Which Clouds?

Amazon Web Services (AWS) – 40%

Microsoft Azure – about 50% of AWS

Google Cloud – 3rd place

IBM Bluemix – growing VERY fast

Salesforce, DigitalOcean, Rackspace,

1&1, UpCloud, CityCloud, CloudSigma,

CloudWatt, Aruba, CloudFerro, Orange,

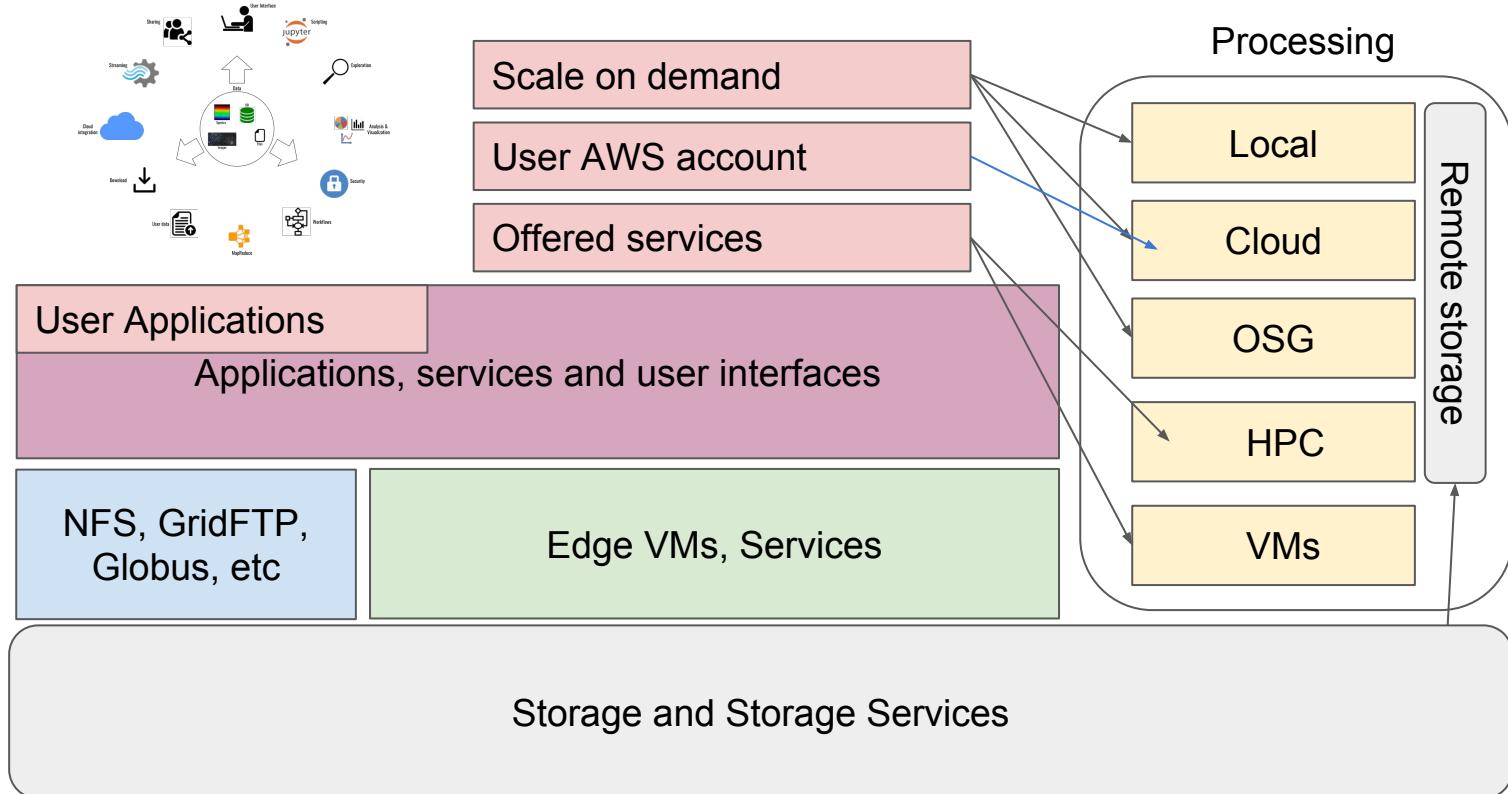
OVH, T-Systems



Cloud for Research: Aristotle,
Bionimbus, Jetstream, Chameleon, RedCloud



Why not both?



What kind of users/groups?

- Runaway/Abusive users (sleep infinity / over subscribed)
- Superuser a. → Data-In Intensive Jobs (e.g., ML)
- Superuser b. → Data-Out Intensive Jobs (e.g., Sims)
- Superuser c. → Resource Intensive Jobs (e.g., Modeling/Fitting)
- Super user d. → Any combination of the above
- User with no exposure or experience (Need good documentation/tutorial, Outreach)
- Users provisioning hardware
- Users with money/cloud credits to provide
- Ephemeral users (workshops, demos)
- “Normal” users

CLOUD NATIVE TRAIL MAP

The Cloud Native Landscape *Landscape* has a large number of options. This Cloud Native Trail Map is a recommended process for leveraging open source, cloud native technologies. At each step, you can choose a vendor-supported offering or do it yourself, and everything after step #3 is optional based on your circumstances.

HELP ALONG THE WAY

A. Training and Certification

Consider training offerings from CNCF and then take the exam to become a Certified Kubernetes Administrator or a Certified Kubernetes Application Developer cncf.io/training

B. Consulting Help

If you want assistance with Kubernetes and the surrounding ecosystem, consider leveraging a Kubernetes Certified Service Provider cncf.io/kcsp

C. Join CNCF's End User Community

For companies that don't offer cloud native services externally cncf.io/enduser

WHAT IS CLOUD NATIVE?

Cloud native technologies empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs exemplify that approach.

These techniques enable loosely coupled systems that are resilient, manageable, and observable. Combined with robust automation, they allow engineers to make high-impact changes frequently and predictably with minimal toil.

The Cloud Native Computing Foundation seeks to drive adoption of this paradigm by fostering and sustaining an ecosystem of open source, vendor-neutral projects. We democratize state-of-the-art patterns to make these innovations accessible for everyone.

l.cncf.io



v20181206

1. CONTAINERIZATION

- Commonly done with Docker containers
- Any size application and dependencies (even PDP-11 code running on an emulator) can be containerized
- Over time, you should aspire towards splitting suitable applications and writing future functionality as microservices



3. ORCHESTRATION & APPLICATION DEFINITION

- Kubernetes is the market-leading orchestration solution
- You should select a Certified Kubernetes Distribution, Hosted Platform, or Installer cncf.io/certified
- Helm Charts help you define, install, and upgrade even the most complex Kubernetes applications



5. SERVICE PROXY, DISCOVERY, & MESH

- CoreDNS is a fast and flexible tool that is useful for service discovery
- Envoy and Linkerd each enable service mesh architectures
- They offer health checking, routing, and load balancing



7. DISTRIBUTED DATABASE & STORAGE

When you need more resiliency and scalability than you can get from a single database, Vitess is a good option for running MySQL at scale through sharding. Rook is a storage orchestrator that integrates a diverse set of storage solutions into Kubernetes.



9. CONTAINER REGISTRY & RUNTIME

Harbor is a registry that stores, signs, and scans content. You can use alternative container runtimes. The most common, all of which are OCI-compliant, are containerd, rkt and CRI-O.



2. CI/CD

- Setup Continuous Integration/Continuous Delivery (CI/CD) so that changes to your source code automatically result in a new container being built, tested, and deployed to staging and eventually, perhaps, to production
- Set up automated rollouts, roll backs and testing

4. OBSERVABILITY & ANALYSIS

- Pick solutions for monitoring, logging and tracing
- Consider CNCF projects Prometheus for monitoring, Fluentd for logging and Jaeger for Tracing
- For tracing, look for an OpenTracing-compatible implementation like Jaeger



6. NETWORKING

To enable more flexible networking, use a CNI-compliant network project like Calico, Flannel, or Weave Net.



8. STREAMING & MESSAGING

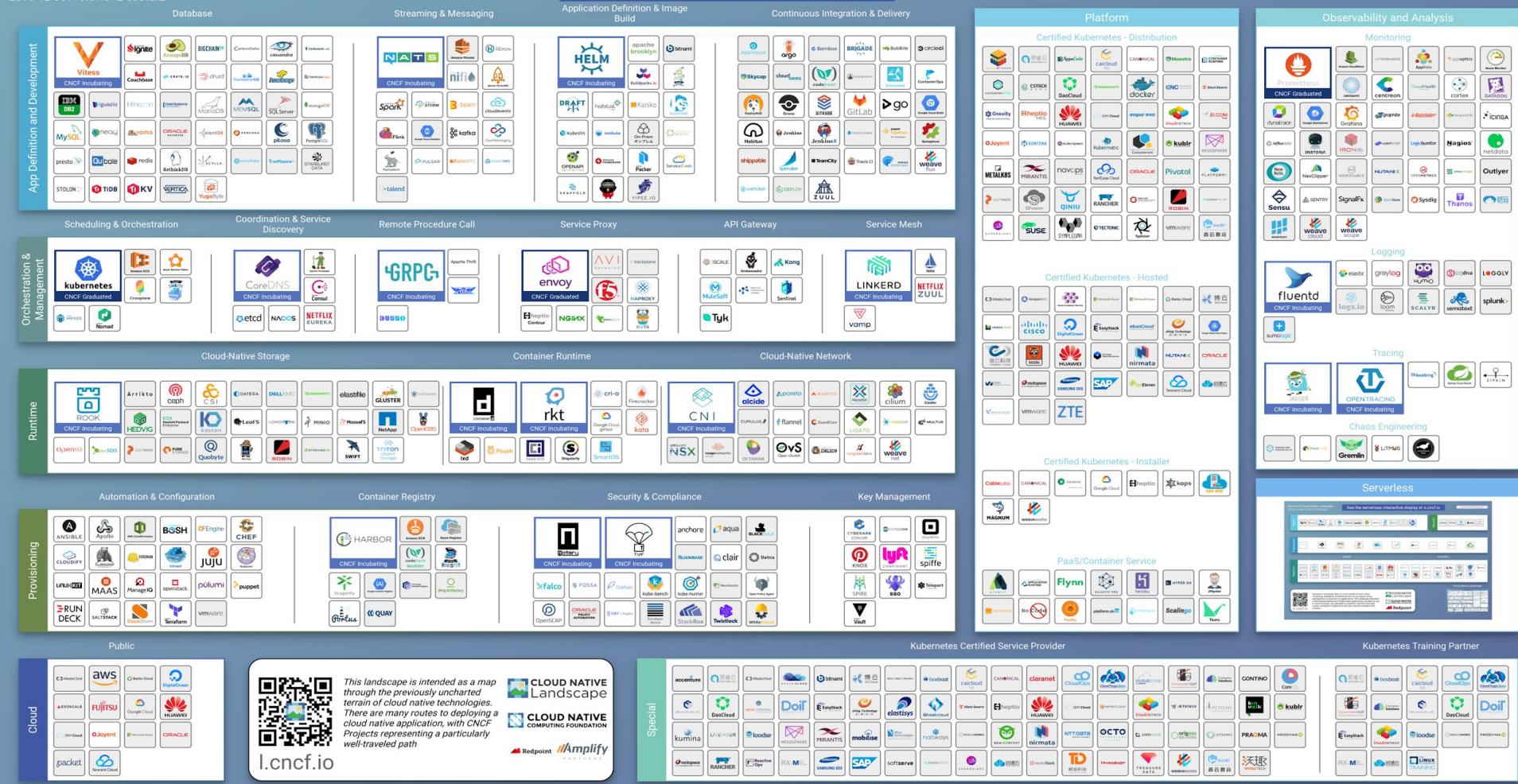
When you need higher performance than JSON+REST, consider using gRPC or NATS. gRPC is a universal RPC framework. NATS is a multi-modal messaging system that includes request/reply, pub/sub and load balanced queues.



10. SOFTWARE DISTRIBUTION

If you need to do secure software distribution, evaluate Notary, an implementation of The Update Framework.



See the interactive landscape at l.cncf.io

Tools



Security



Framework



Hosted

Installable

Platform

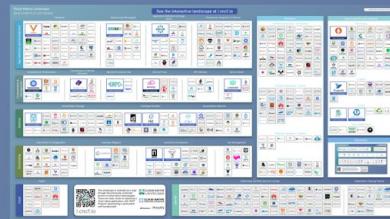


Cloud Native Landscape



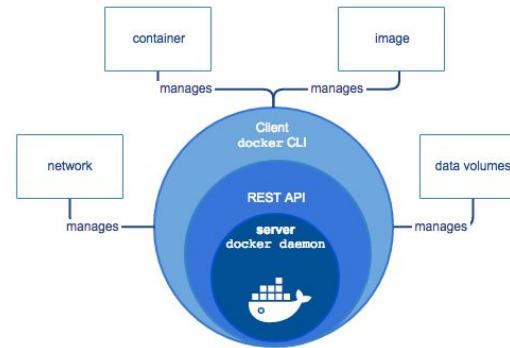
s.cncf.io

Serverless computing refers to a new model of cloud native computing, enabled by architectures that do not require server management to build and run applications. This landscape illustrates a finer-grained deployment model where applications, bundled as one or more functions, are uploaded to a platform and then executed, scaled, and billed in response to the exact demand needed at the moment.



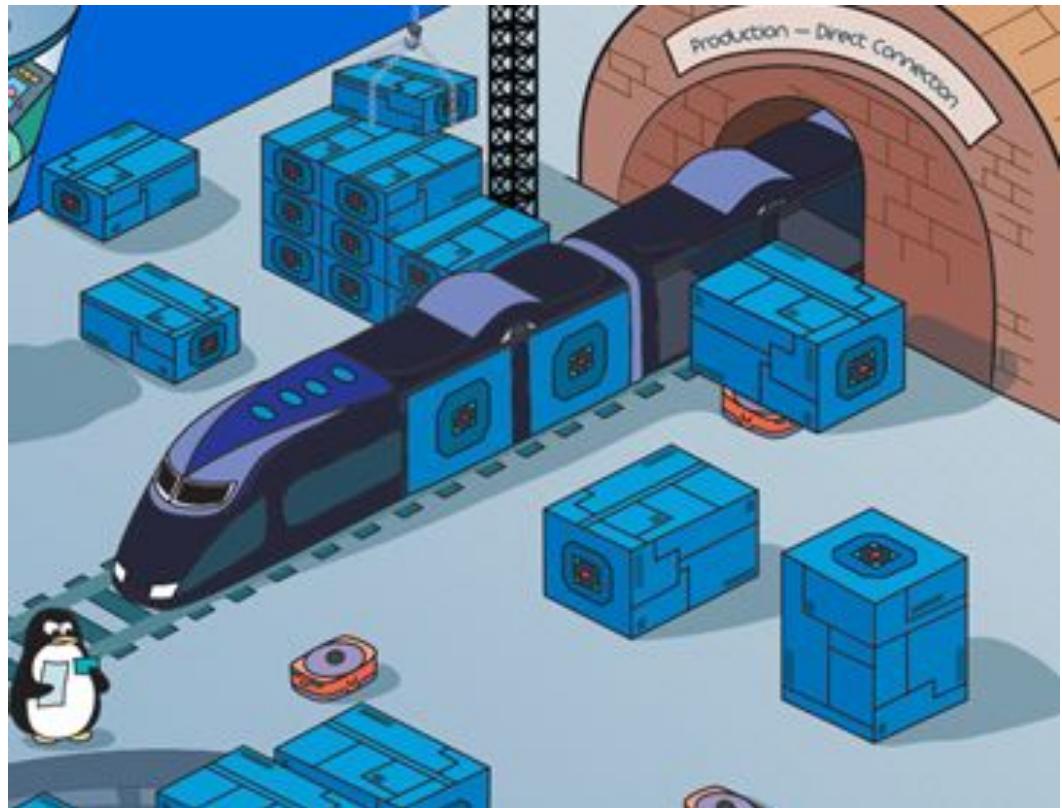
Containerization to the rescue

- It's been around for over 10 years, but popular since 2014 thanks to Docker
- Many other alternatives (rkt, kata, shifter, singularity, etc...)
- Lightweight, stand-alone, executable package of a piece of software that includes everything to run it
- Not just applications
- Software designed storage
- Software designed network



Container organization and orchestration

- We can create a container with an application inside, now what?
- Need to consider:
 - Resource needs
 - Fault tolerant
 - Load balancing
 - Storage management
 - Lifecycle
 - Service Discovery
 - Scalability



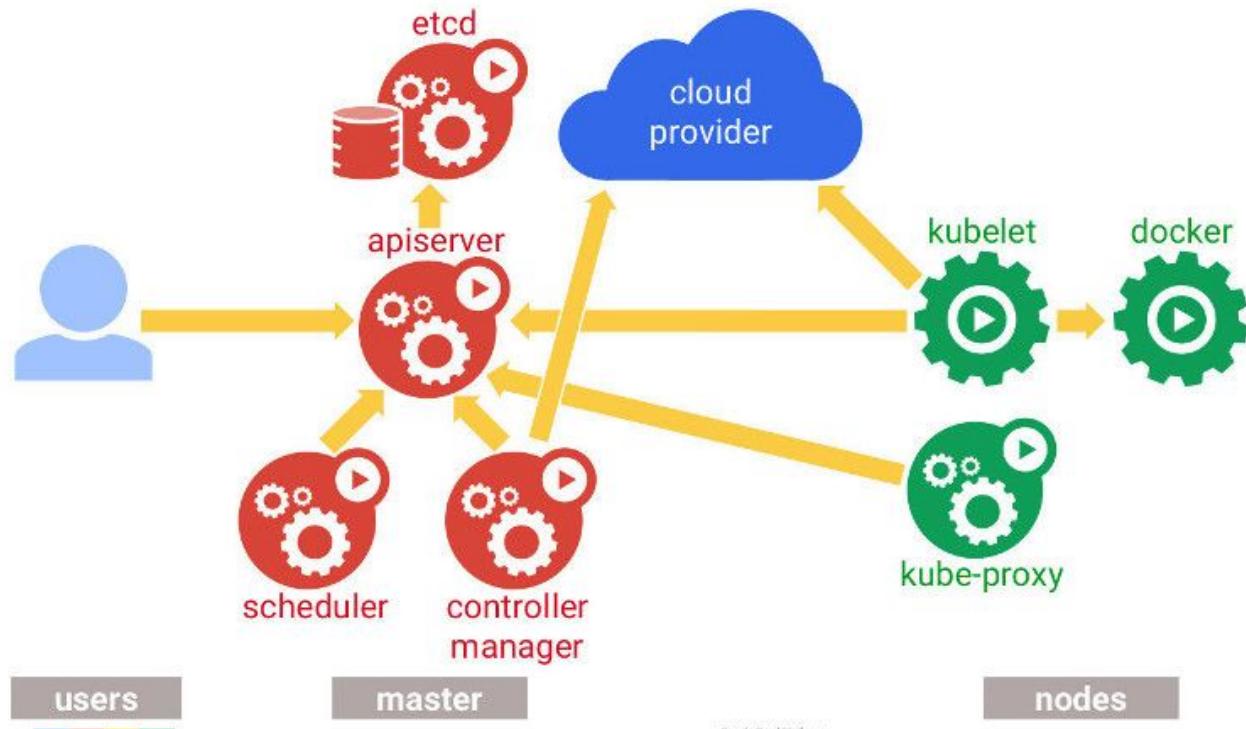
The Kubernetes Factor

- It solves all previous issues and more (not the only one but most popular)
- Open source container management and orchestration platform
- Developed by Google, made open sourced
- One of top 5 most commented open source repositories and #2 in number of pull request
- Standard within all cloud platforms
- Flexible and extensible, customize schedulers
- Is changing the cloud computing paradigm



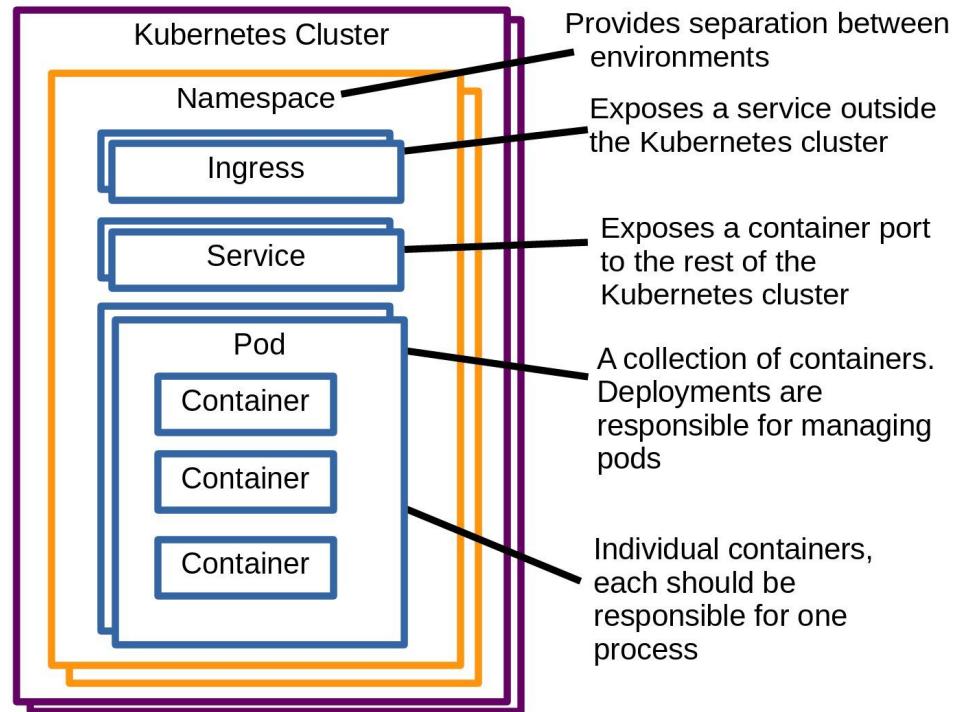
Kubernetes Overview

- Cloud democratization
- Easy deployment
- Controls most of the aspects
- Adopted at NCSA, CERN, LSST, NASA
- Edge Computing
- Scalability
- Federation
- Resource Manager

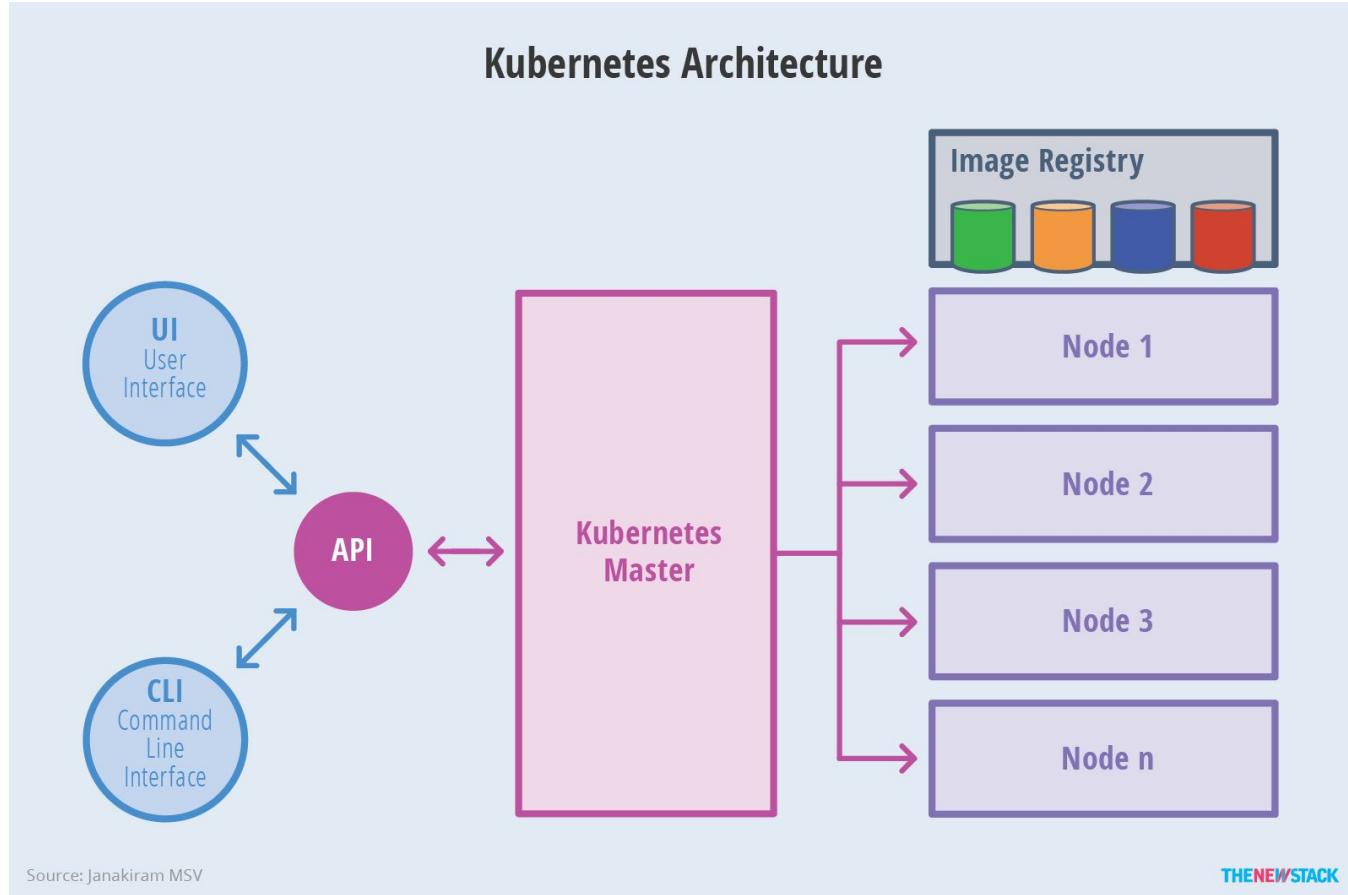


Kubernetes Key Concepts

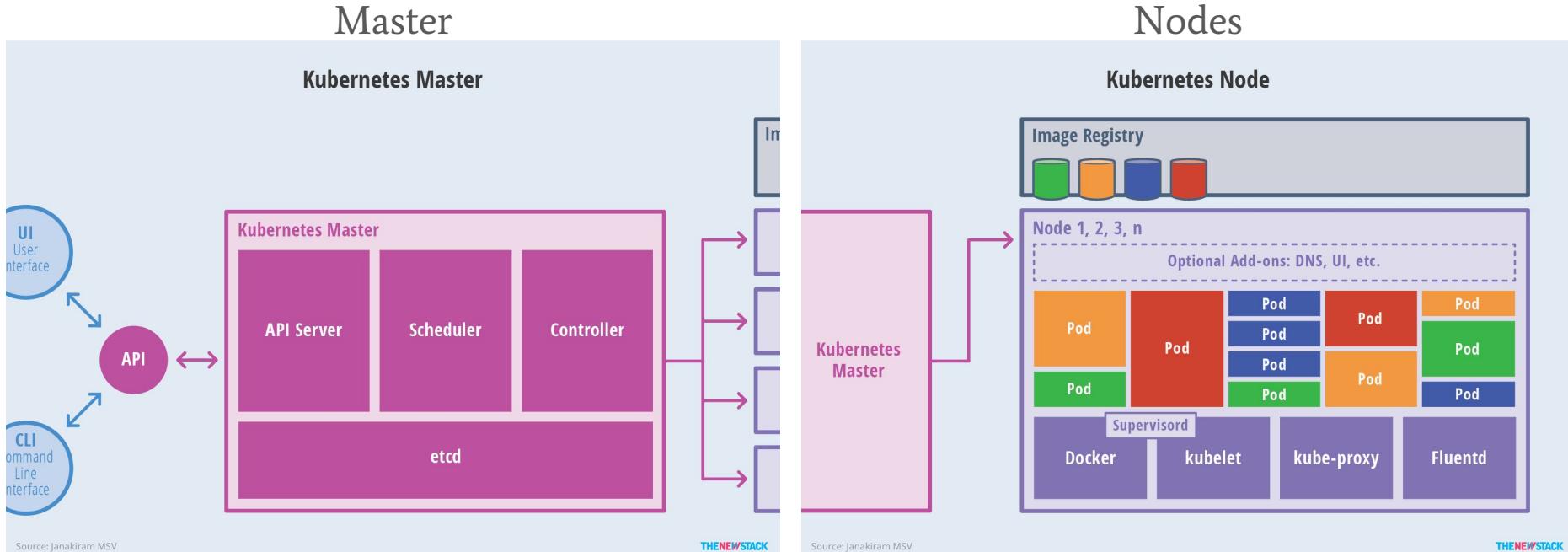
- **Pod** - A group of Containers
- **Labels** - Labels for identifying pods
- **Kubelet** - Container Agent
- **Proxy** - A load balancer for Pods
- **etcd** - A metadata service
- **cAdvisor** - Container Advisor provides resource usage/performance statistics
- **Replication Controller** - Manages replication of pods
- **Scheduler** - Schedules pods in worker nodes
- **API Server** - Kubernetes API server



The Kubernetes Architecture



The Kubernetes Architecture



Applications



The Dark Energy Survey



- 4 meters telescope, 520 Mpx camera
- 5 year survey, $\frac{1}{8}$ of the sky, Telescope in Chile, data @ NCSA, about to start 6th season
- Main Goal: To constrain the models of the Universe regarding Dark Energy and Dark Matter.
- Many other Science Cases! (New dwarf planet, New galaxy satellites, Supernovae, etc)
- 1 - 3 TB of data per night, 1 PB of data
- Processing done at FermiGrid, Campus Cluster and Blue Waters
- Thousands of images and billions of rows, ~500 millions objects
- 1st Public Data Release in January 2018
- NCSA provide means to access and interact with data → Containers

The DES Data Access

Challenges:

- Data access wasn't very clear in original proposal
- People
- Time
- Collaborations Needs
- All the rest of technical challenges



- DES Survey: Gold (Data) Mine
- DESDM: Excellent job at mining the data
- Consumers outside the mine
- Need to bring/expose gold (data) outside
- Tools and interfaces
- DES DR1 is out!

easyaccess: DES command line tool



DARK ENERGY SURVEY
DATA MANAGEMENT

```
easyaccess 1.4.0. The DESDM Database shell.  
Connected as mcarras2 to dessci.  
** Type 'help' or '?' to list commands. **  
  
*General Commands* (type help <command>):  
=====  
clear edit help history prefetch version  
config exit help_function import shell  
  
*DB Commands* (type help <command>):  
=====  
add_comment find_tables myquota show_index  
append_table find_tables_with_column mytables user_tables  
change_db find_user refresh_metadata_cache whoami  
describe_table load_table set_password  
execproc loadsqL show_db  
  
*Default Input*  
=====  
* To run SQL queries just add ; at the end of query  
* To write to a file : select ... from ... where ... ; > filename  
* Supported file formats (.csv, .tab, .fits, .h5)  
* To check SQL syntax : select ... from ... where ... ; < check  
* To see the Oracle execution plan : select ... from ... where ... ; < explain  
  
* To access an online tutorial type: online_tutorial  
  
DESDB ~>
```

- DES DB in Oracle
- Specifically designed for DES (internal and public)
- Enhanced SQL command line interpreter in Python
- Astronomer friendly
- Python API, web interface
- There are many other CLI and GUI clients.
- Needed a simple tool, easy to use and install
- Autocompletion
- Load/Save to hdf5, fits, csv

easyaccess: DES command line tool

```
matias@XPS:~$ e
```

- DES DB in Oracle
- Specifically designed for DES (internal and public)
- Enhanced SQL command line interpreter in Python
- Astronomer friendly
- Python API, web interface
- There are many other CLI and GUI clients.
- Needed a simple tool, easy to use and install
- Autocompletion
- Load/Save to hdf5, fits, csv

DES Labs: Collection of containerized tools for DES



DES Labs

- Launched March 2015
- Used by the Collaboration
- Running using Kubernetes at NCSA cloud
- Currently being migrated to match DRL Infrastructure

Easyaccess web



Jupyterhub + easyacces



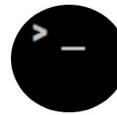
DES cutouts



Footprint



Easyaccess online



DESDM Services status



External Links

Science Server



NOAO Data Lab



CosmoHub



NCSA DESaccess: Services



DARK ENERGY SURVEY desaccess



mcarras2
mcarras2@ncsa.illinois.edu

Home

DB access

DES Table Schema

Example Queries

Cutouts Service

DES JupyterLab

Finding Chart

DES Footprint

Data Analysis

My Jobs

Help



```
SELECT dr1.RA,dr1.DEC,dr1.COADD_OBJECT_ID
FROM dr1_main_sample(0.01) dr1
WHERE
dr1.MAG_AUTO_G < 18 and
dr1.WAVG_SPREAD_MODEL_I + 2.0*dr1.WAVG_SPREADERR_M
dr1.WAVG_SPREAD_MODEL_I - 1.0*dr1.WAVG_SPREADERR_M
dr1.WAVG_SPREAD_MODEL_I - 1.0*dr1.WAVG_SPREADERR_M
dr1.IMAFLAGS_ISO_G = 0 and
dr1.IMAFLAGS_ISO_R = 0 and
dr1.IMAFLAGS_ISO_I = 0 and
```



DB ACCESS

Oracle SQL web-client

[More...](#)

DES TABLE SCHEMA

Browse all tables

[More...](#)

EXAMPLE QUERIES

See some example queries as a start

[More...](#)

CUTOUTS SERVICE

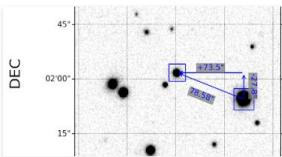
Generate cutouts for positions or ids

[More...](#)

DES JupyterLabs

(Beta) Jupyter Labs

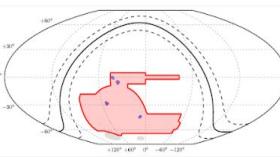
[More...](#)



FINDING CHART

Find your object

[More...](#)



DES FOOTPRINT

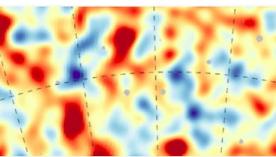
Interactive globe



DATA ANALYSIS

SEDs and color-color diagrams

[More...](#)



MY JOBS

List of submitted jobs

[More...](#)



HELP

Help form

[More...](#)

NCSA DESaccess: DB access



DARK ENERGY SURVEY desaccess



mck
mcarras2@illinois.edu

Home

DB access

DR1 Table Schema

Example Queries

Cutout Service

DR1 Footprint

My Jobs

DES JupyterLab

Help

Query box

Insert your query in the box below. Data results for "Quick" Jobs (30 sec.) will be displayed at the bottom.

```

1 --
2 -- Example Query --
3 -- This query selects stars around the center of globular cluster M2
4 SELECT
5 COADD_OBJECT_ID,RA,DEC,
6 MAG_AUTO_G,G,
7 MAG_AUTO_R,R,
8 WAVG_MAG_PSF_G,G_PSF,
9 WAVG_MAG_PSF_R,R_PSF
10 FROM DR1_MAIN
11 WHERE
12 RA between 323.36-0.12 and 323.36+0.12 and
13 DEC between -0.82-0.12 and -0.82+0.12 and
14 WAVG_SPREAD_MODEL_I + 3.0*WAVG_SPREADERR_MODEL_I < 0.005 and
15 WAVG_SPREAD_MODEL_I > -1 and
16 IMAFLAGS_ISO_G = 0 and
17 IMAFLAGS_ISO_R = 0 and
18 FLAGS_G < 4 and
19 FLAGS_R < 4
20

```

Submit Job

Clear

Check

Quick

See Examples

Output file (.csv, .fits or .h5). Enable in order to submit.

Output file



Options:

Compressed files (csv and h5 files). Slightly longer jobs but smaller files

Job Name (optional)

Send email after completion

Email

NCSA DESaccess: Cutouts Service

DARK ENERGY SURVEY desaccess

Coadds Images Cutout Form

Upload the file with the positions or enter the positions by hand and run the desthumb generator

 Upload File (csv, with RA,DEC as uncommented header)

 Enter Values

Xsize (in arcminutes): 1.0

Ysize (in arcminutes): 1.0

Job Name

Send email on completion Email

Return just list of files (do not produce and display pngs, i.e. faster)

 Clear Form

 Submit Job

mck mcarras2@illinois.edu

Home

DB access

DR1 Table Schema

Example Queries

Cutout Service

DR1 Footprint

My Jobs

DES JupyterLab

Help

NCSA DESaccess: Cutouts Service

 DARK ENERGY SURVEY desaccess

 mck
mcarras2@illinois.edu

Coadds Images Cutout Form

Upload the file with the positions or enter the positions by hand and run the desthumt command.

 Upload File (csv, with RA,DEC as uncommented header)

 Enter Values

 Xsize (in arcminutes): 1.0

 Ysize (in arcminutes): 1.0

 Job Name

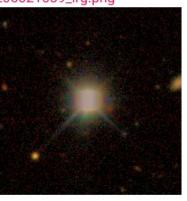
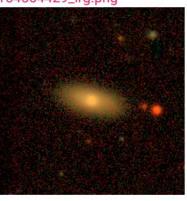
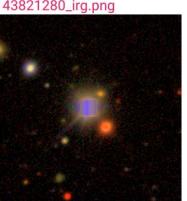
 Email Options

 Return Type



Job1 : d927a264_746c_4f7a_82cd_f46ebce496c7 (19 objects)



NCSA DESaccess: Asynchronous Jobs



DARK ENERGY SURVEY desaccess



mck
mcarras@illinois.edu

Home

DB access

DR1 Table Schema

Example Queries

Cutout Service

DR1 Footprint

My Jobs

DES JupyterLab

Help

My Jobs

#	Status	Job Name	Job type	Execution time (s)	Cancel Job	Queries	Results	Files
0	Green	Name: Job id: 6b4cac2b-b544-4461-96fb-58cd4968a338 6 days and 0 hours ago (Expired)	query	0	X	Query	Cutouts	Files
1	Green	Name: Job id: daf5ee3c-461e-42ed-8efb-5fcfbf684047 6 days and 0 hours ago (Expired)	cutout	1	X	Query	Cutouts	Files
2	Green	Name: testapi Job id: 0d6c5a58-b00a-4798-834f-9816c6fa98e5 7 days and 4 hours ago (Expired)	cutout	3	X	Query	Cutouts	Files
3	Green	Name: testapi Job id: 12861656-8075-4629-8e4f-fd4378013634 7 days and 4 hours ago (Expired)	cutout	3	X	Query	Cutouts	Files
4	Green	Name: testapi Job id: d9a37fe9-209b-4296-b87d-c6567cde0649 7 days and 4 hours ago (Expired)	cutout	1	X	Query	Cutouts	Files
5	Green	Name: Job id: 6d10cf32-3cd6-4050-bb90-344268dd615f 7 days and 5 hours ago (Expired)	cutout	1	X	Query	Cutouts	Files
6	Red	Name: testapi Job id: b85e747-5201-4e49-a0eb-f2bb7f26de 7 days and 5 hours ago (Expired)	cutout	-1	X	Query	Cutouts	Files
7	Green	Name: Job id: f8fees56a-4685-49ff-b7be-603310ccdeb 8 days and 16 hours ago (Expired)	query	577	X	Query	Cutouts	Files
8	Green	Name: Job id: df8a57c4-b1d5-4332-80d5-a08a27b537d9 8 days and 16 hours ago (Expired)	query	1042	X	Query	Cutouts	Files
9	Red	Name: Job id: 7ffdb550-4d38-441f-a037-ed659b3b79c9 8 days and 16 hours ago (Expired)	query	-1	X	Query	Cutouts	Files
10	Green	Name: Job id: fcacaaec-9d63-45a4-92f2-4f847b9b415c 8 days and 16 hours ago (Expired)	query	9	X	Query	Cutouts	Files
11	Green	Name: Job id: a88b79cc-fd71-4e00-a33d-92b5be98106f 8 days and 17 hours ago (Expired)	query	9	X	Query	Cutouts	Files
		Name: demo1						

REFRESH

DELETE

NCSA DESaccess: Footprint and Jupyter Labs

DARK ENERGY SURVEY desaccess

mck
mcarras2@illinois.edu

DES DR1 Footprint

Use the footprint tool to search a tile by position or name. Double click to select a tile.

Position (ra,dec) Tilename

Coordinates DR1 TILES HPIX nside=32

Tile properties

Name :
Tile Center :
No Objects :
RA Corners :
DEC Corners :

Get Tile Files

Click [here](#) to get access to all the tiles

Home DB access DR1 Table Schema Example Queries Cutout Service **DR1 Footprint** My Jobs DES JupyterLab Help

DARK ENERGY SURVEY desaccess

mck
mcarras2@illinois.edu

DES Jupyter Labs (Beta)

This feature is experimental only. Please use with caution. You can launch, access and delete your Jupyter Notebook. This Notebook will run with 1 CPU and 2GB of RAM.

Deploy Lab Delete Lab

Status

● Ready
Status: Running
 Go To Lab

REFRESH

Home DB access DR1 Table Schema Example Queries Cutout Service DR1 Footprint My Jobs **DES JupyterLab** Help

NCSA DESaccess: Labs with access to Jobs and easyaccess

File Edit View Run Kernel Tabs Settings Help

Running

Commands

Cell Tools

Files

+

basics_plotting x

Markdown v

Python 3

Name Last Modified

- 0d6c5a58-b00a-4798... 7 days ago
- 0fa487cd-75a5-4015... 7 days ago
- 12861656-8075-4629... 7 days ago
- 1aae7465-aef6-44fc... 7 days ago
- 507683dc-53d6-4033... 7 days ago
- 6b4cac2b-b544-44e1... 6 days ago
- 6d10cf32-3cd6-4090... 7 days ago
- 7ffd5b50-4d38-441f-a... 10 hours ago
- 810a2aee-a8d7-4356... 16 days ago
- 8b7290af-8a2e-4ace-... 7 days ago
- b8fea56a-4685-49f9-... 10 hours ago
- a88b79cc-fd71-4ed0-... 10 hours ago
- b85ea747-5201-4e49... 7 days ago
- d3822272c23d4a3b6c... 7 days ago
- d9a37fe9-209b-4296-... 7 days ago
- da5ee3c-461e-42ed-... 6 days ago
- df8a57c4-b1d5-4332-... 10 hours ago
- fcaacdec-96d3-45a4-... 10 hours ago
- 0d6c5a58-b00a-4798... 7 days ago
- 0d6c5a58-b00a-4798... 7 days ago
- 12861656-8075-4629... 7 days ago
- 12861656-8075-4629... 7 days ago
- 6b4cac2b-b544-44e1... 6 days ago
- 6d10cf32-3cd6-4090... 7 days ago
- 6d10cf32-3cd6-4090... 7 days ago
- b85ea747-5201-4e49... 7 days ago
- d9a37fe9-209b-4296-... 7 days ago
- d9a37fe9-209b-4296-... 7 days ago
- da5ee3c-461e-42ed-... 6 days ago
- da5ee3c-461e-42ed-... 6 days ago
- quickResults.csv 6 days ago

<Figure size 720x720 with 0 Axes>

MAG AUTO I

MAG AUTO R

Some interactive plots using Bokeh and Holoviews

In [9]: `import holoviews as hv
hv.extension('bokeh')`

In [10]: `hextiles = hv.HexTiles(df, [('MAG_AUTO_R', 'R'), ('MAG_AUTO_I', 'I')], [], extents=(20,26,20,26))`

In [11]: `hextiles.options(width=500, height=500, min_count=0, tools=['hover'], colorbar=True) * hv.`

Out[11]:

Terminal 4 x

DARK ENERGY SURVEY DATA MANAGEMENT

easyaccess 1.4.4. The DESDM Database shell.
Connected as nck to desdr.
** Type 'help' or '?' to list commands. **

General Commands (type help <command>):

```
clear edit help history prefetch version
config exit help_function import shell
```

DB Commands (type help <command>):

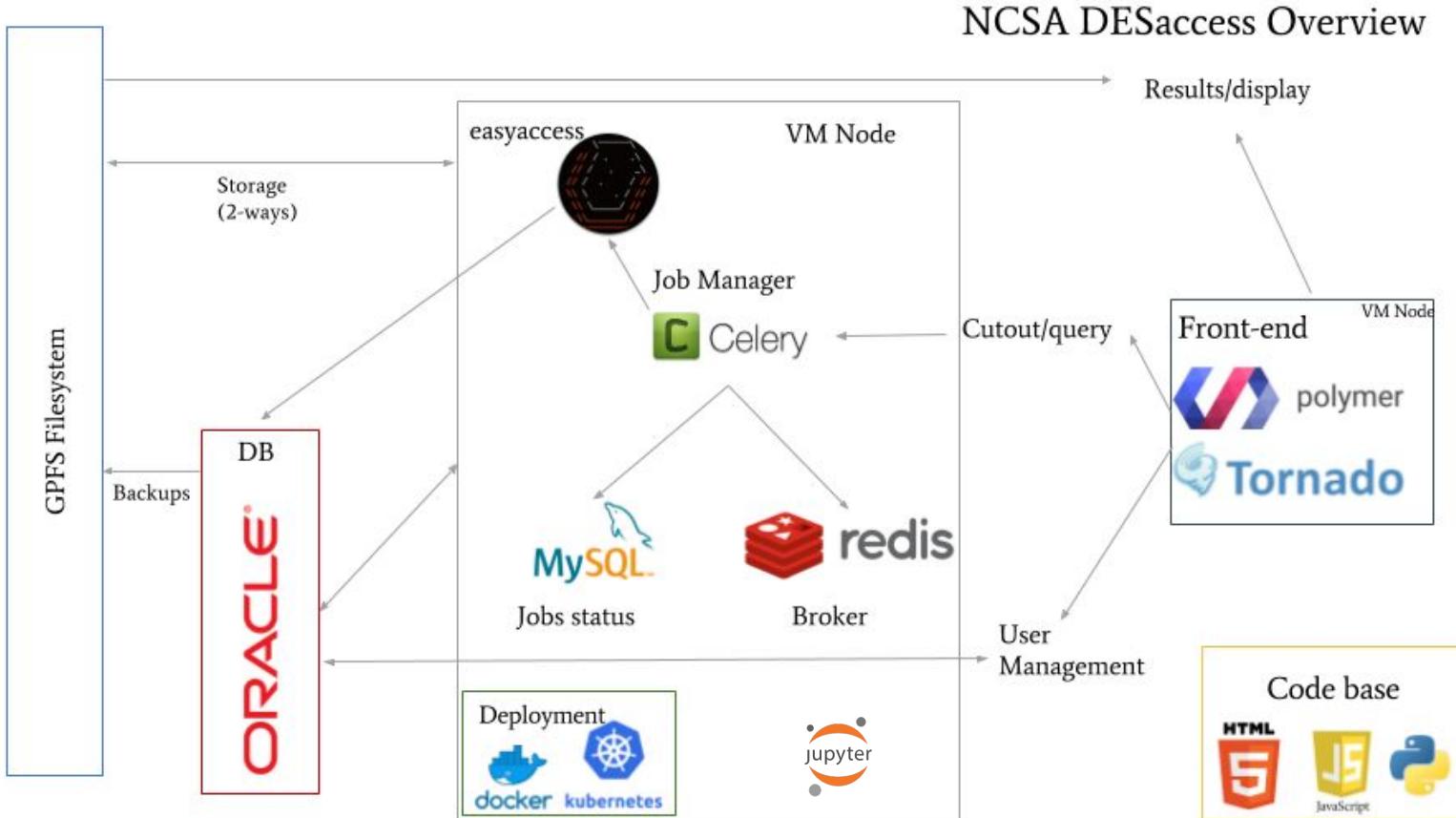
```
describe_table loadsql refresh_metadata_cache show_db
find_tables show_index
find_tables_with_column set_password whoami
```

Default Input

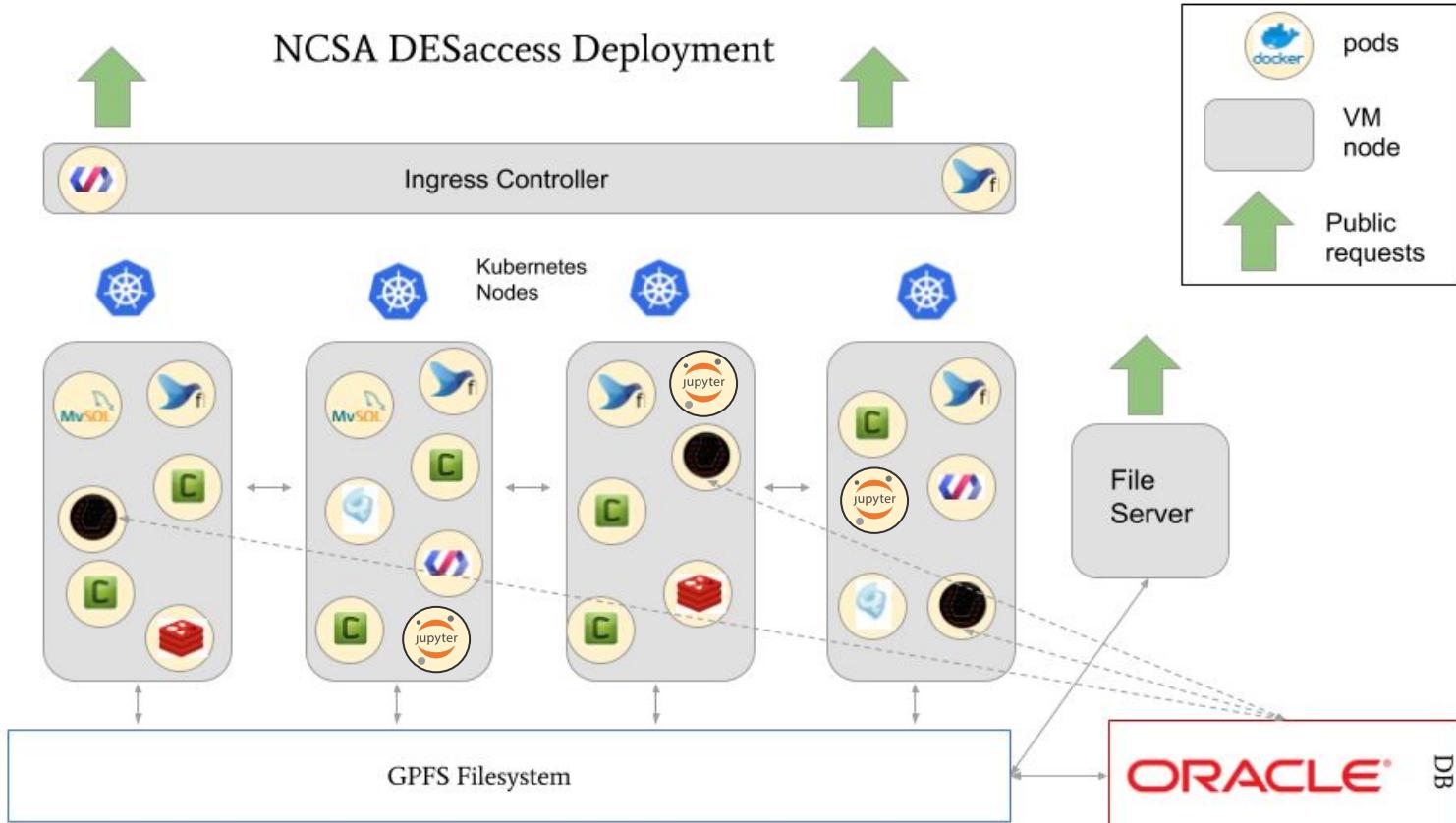
```
* To run SQL queries just add ; at the end of query
* To write to a file :select ... from ... where ... ; > filename
* Supported file formats .csv, .tab., .fits, .h5)
* To check SQL syntax: select ... from ... where ... ; < check
* To see the Oracle execution plan :select ... from ... where ... ; < explain
* To access an online tutorial type: online_tutorial
```

DESDB -> []

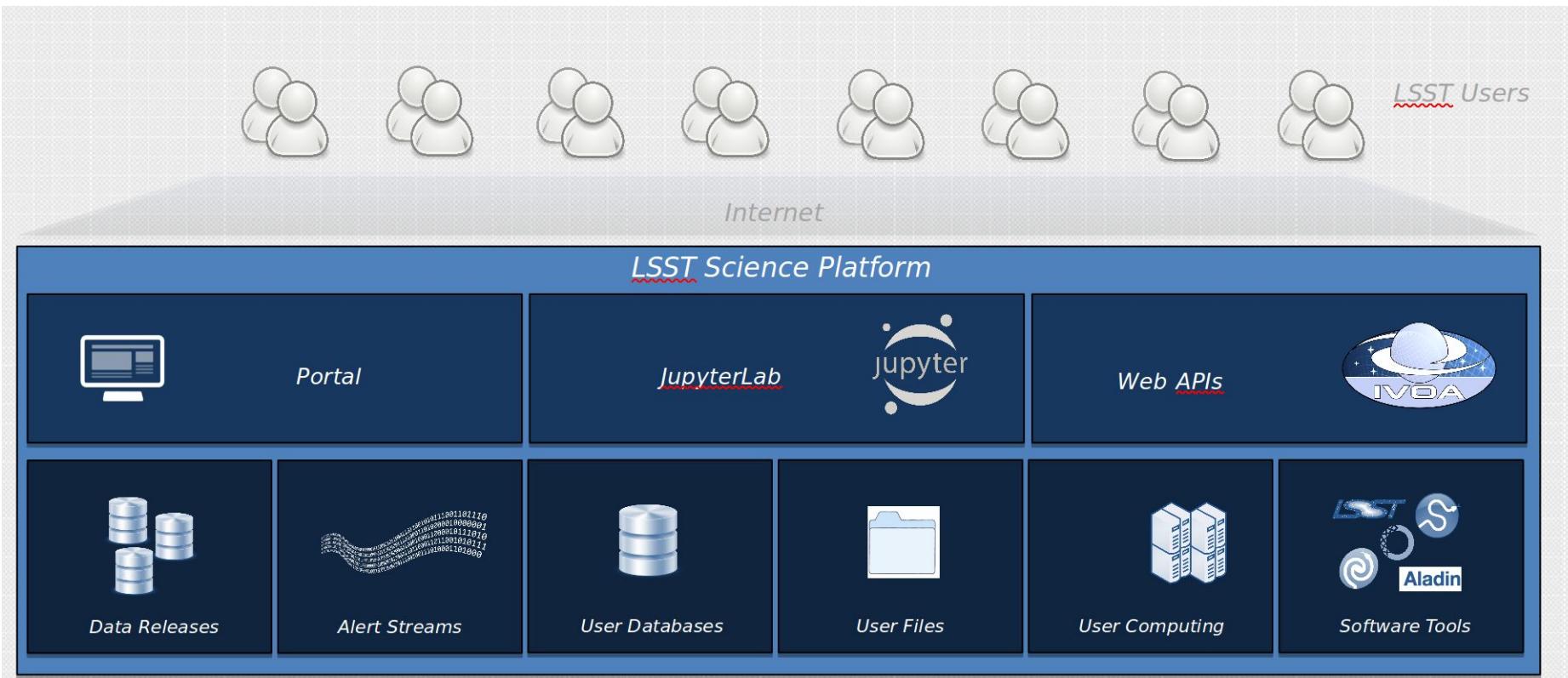
NCSA DESaccess: Technology Overview



NCSA DESacces: Deployment



LSST Science Platform



GAIA Archive

[EUROPEAN SPACE AGENCY](#) [ABOUT ESAC](#)

SIGN IN

gaia archive

[HOME](#) [SEARCH](#) [STATISTICS](#) [VISUALISATION](#) [DOCUMENTATION](#) [HELP](#)



Welcome to the Gaia Archive

Gaia is an ambitious mission to chart a three-dimensional map of our Galaxy, the Milky Way, in the process revealing the composition, formation and evolution of the Galaxy. Gaia will provide unprecedented positional and radial velocity measurements with the accuracies needed to produce a stereoscopic and kinematic census of about one billion stars in our Galaxy and throughout the Local Group. This amounts to about 1 per cent of the Galactic stellar population.



Top Features



Citation
How to cite and acknowledge Gaia.



Search
Query for Gaia sources using an ADQL (Astronomical Data Query Language) Interface in an asynchronous mode (UWS).



Download
Direct download of Gaia data files.



Help
For questions, suggestions or problem reports, contact the Helpdesk.



Documents
Links to Gaia Archive and related Gaia documentation.



Gaia Mission
News, information, and resources on the Gaia mission for the scientific community.



Statistics
Show statistics of Gaia tables.

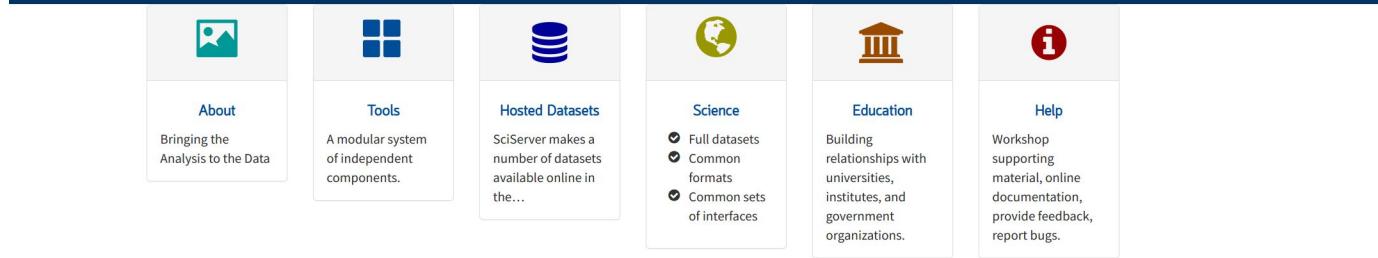


Partners
Partner data centres also serving Gaia data.

SciServer



The screenshot shows the SciServer homepage. At the top, there's a dark blue header with the SciServer logo on the left, a "Collaborative data-driven science" tagline in the center, and a "new message [JIRA] (LSST-1) cluster to 1.11" notification on the right. Below the header is a large banner with the text "A new vision for science" and "A collaborative research environment for large-scale data-driven science". The main content area is titled "SciServer Betelgeuse v2.0.3" and features six cards with icons and descriptions: "About", "Tools", "Hosted Datasets", "Science", "Education", and "Help". A "Login to SciServer" button is located at the top right of the main content area.



The content area contains six cards:

- About**: Bringing the Analysis to the Data
- Tools**: A modular system of independent components.
- Hosted Datasets**: SciServer makes a number of datasets available online in the...
- Science**:
 - Full datasets
 - Common formats
 - Common sets of interfaces
- Education**: Building relationships with universities, institutes, and government organizations.
- Help**: Workshop supporting material, online documentation, provide feedback, report bugs.



SciServer is administered by **idies** JOHNS HOPKINS UNIVERSITY

SciServer is funded by National Science Foundation award ACI-1261715



SCIaaS Example: Anomaly detection service

Goal: Build a resilient scalable anomaly detection service.

Motivation: Astronomical data (both literal and figurative)

Algorithm: Extended Isolation Forest

Infrastructure: Kubernetes cluster

MapReduce package: Spark

SCIaaS Example: Galaxy selection and similarity search



rec 1



rec 2



rec 3



rec 4



rec 5



rec 6



rec 7



rec 8



rec 1



rec 2



rec 3



rec 4



rec 5



rec 6



rec 7



rec 8



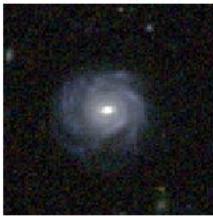
Compress images from 200x200 to 50 or less, for fast search

SCIaaS Example: Galaxy selection and similarity search

1 's similar galaxy



2 's similar galaxy



3 's similar galaxy



4 's similar galaxy



5 's similar galaxy



6 's similar galaxy



7 's similar galaxy



8 's similar galaxy



9 's similar galaxy



0 's similar galaxy



1 's similar galaxy



2 's similar galaxy



3 's similar galaxy



4 's similar galaxy



5 's similar galaxy



6 's similar galaxy



7 's similar galaxy



8 's similar galaxy



9 's similar galaxy



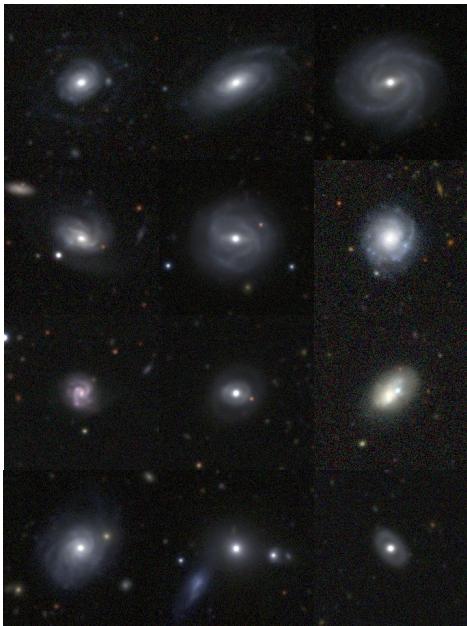
10 's similar galaxy



11 's similar galaxy

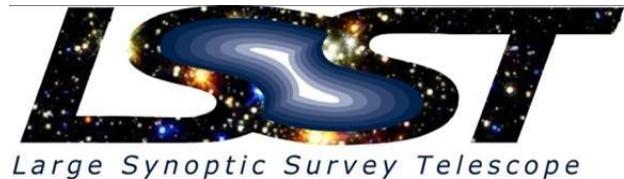


Part of the Motivation



Astronomy is just one example where data exploration needs to be automated.

Large catalogs, Large number of images, many unexpected objects/problems → Anomaly detection



- In operations 2020
- Every night for 10 years
- 18 billions objects (first year), ~40 billions by the end of survey
- ~1500 images per night
- Stream and static data
- Target to capture new physics (moving and variable objects)

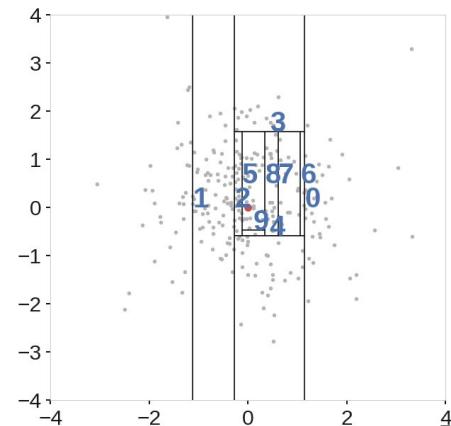
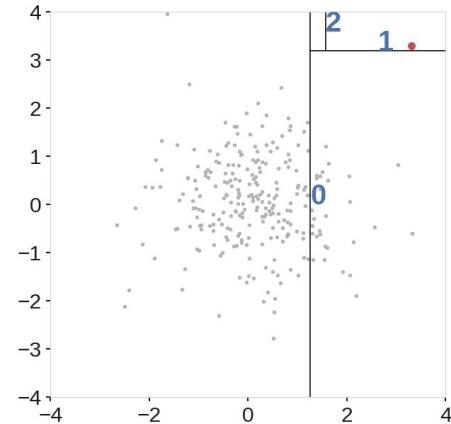


- More than 500 nights of observation over 5 years
- 500 millions cataloged galaxies and 100 millions stars
- Many open problems: Systematics, new objects, new physics, etc.
- Almost completed

Anomaly Detection with Isolation Forest

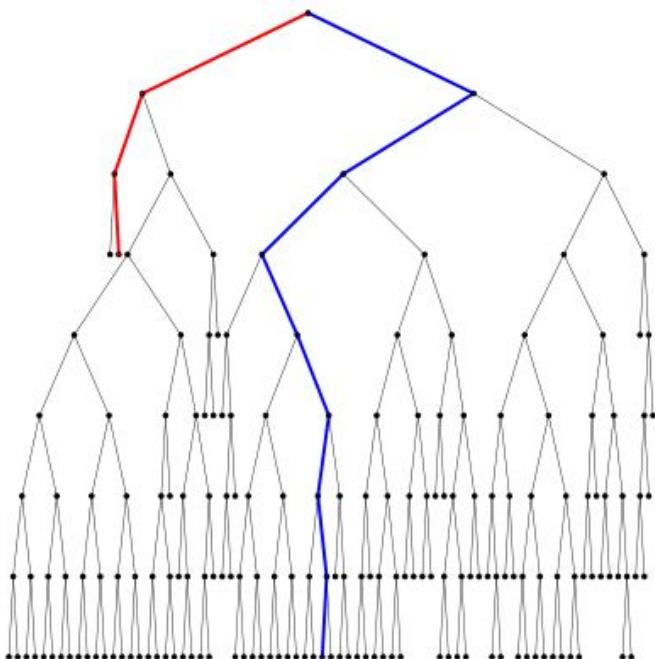
- Few and different to be isolated quicker
- For each tree:
 - Get a sample of the data
 - Randomly select a dimension
 - Randomly pick a value in that dimension
 - Draw a straight line through the data at that value and split data
 - Repeat until tree is complete
- Generate multiple trees → forest
- Anomalies will be isolated in only a few steps
- Nominal points in more
- To score points:
 - Run point down tree, record path
 - Repeat for each tree, aggregate scores
 - Score distribution

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

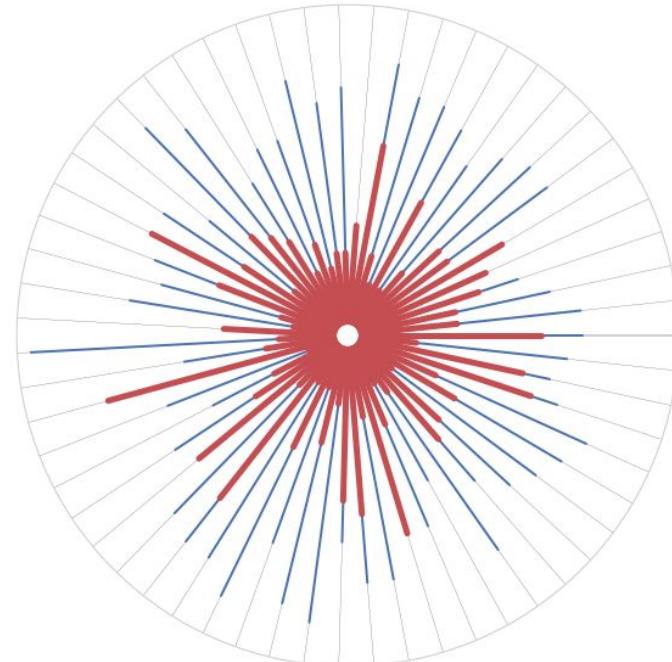


Anomaly Detection with Isolation Forest

Single Tree scores for
anomaly and **nominal** points



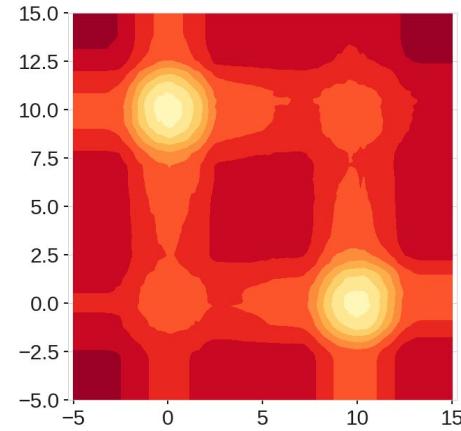
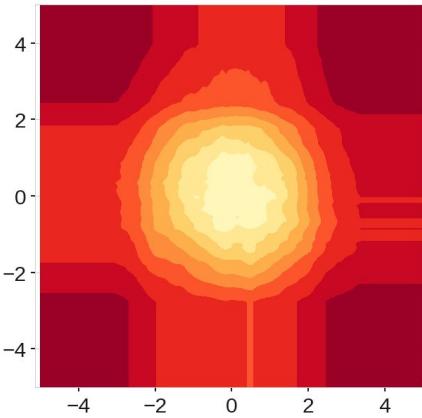
Forest plotted radially.
Scores for **anomaly** and
nominal shown as lines



Anomaly Detection with Extended Isolation Forest

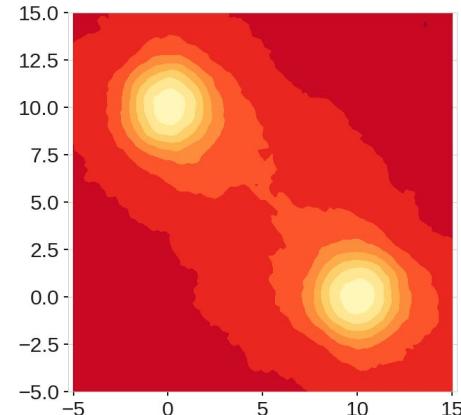
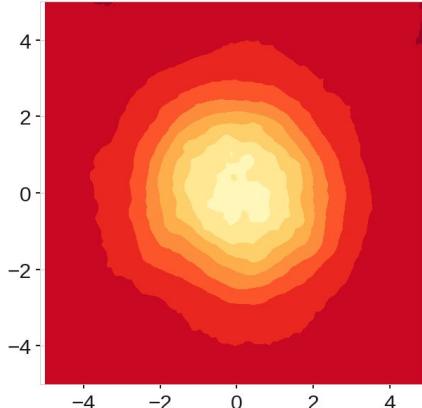
Isolation Forest:

- ✓ Model free
- ✓ Computationally efficient
- ✓ Readily applicable to parallelization
- ✓ Readily application to high dimensional data
- ✗ Inconsistent scoring seen in score maps



Extended Isolation Forest:

- ✓ Model free
- ✓ Computationally efficient
- ✓ Readily applicable to parallelization
- ✓ Readily application to high dimensional data
- ✓ Consistent scoring



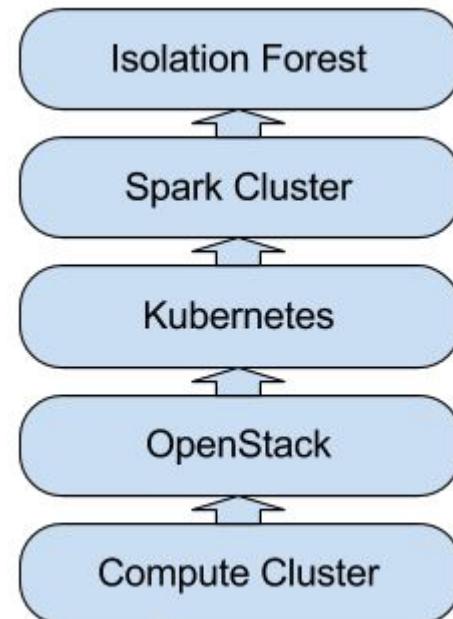
Technology Stack For Anomaly Service

Batch and online anomaly detection for scientific applications in
a Kubernetes environment

Sahand Hariri*
University of Illinois at Urbana-Champaign
sahandha@gmail.com

Matias Carrasco Kind†
National Center for Supercomputing Applications
mcarras2@illinois.edu

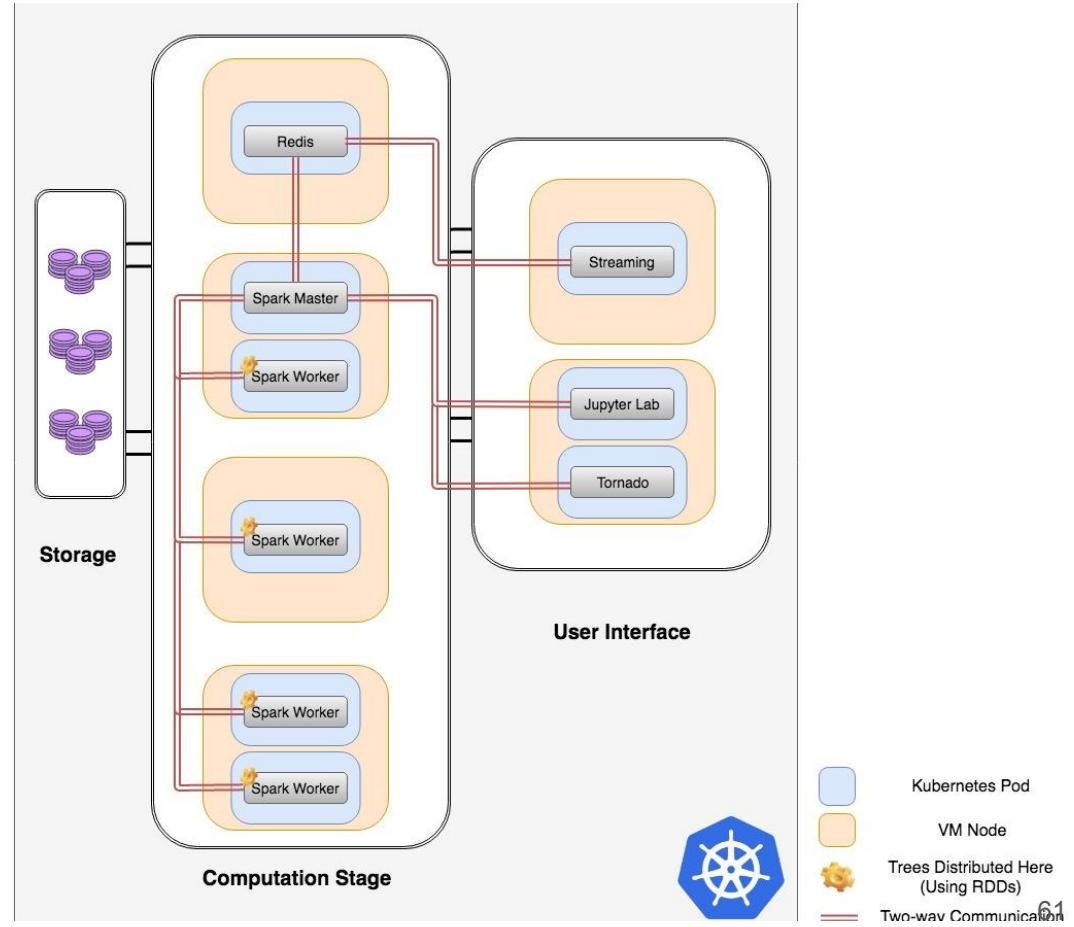
- Use Extended Isolation Forest as core algorithm
- Use Spark to parallelize trees and scoring
- Use Redis as a broker communicator
- To easily deploy in any environment, use Docker
- For orchestration of Docker containers, use Kubernetes
- Kubernetes cluster built on top of OpenStack, but it can be deployed also in AWS, GKE, etc.



Framework Architecture

There are three main components:

1. Storage
2. Computation Stage
3. User Interface / Streaming



Framework Architecture

Storage:

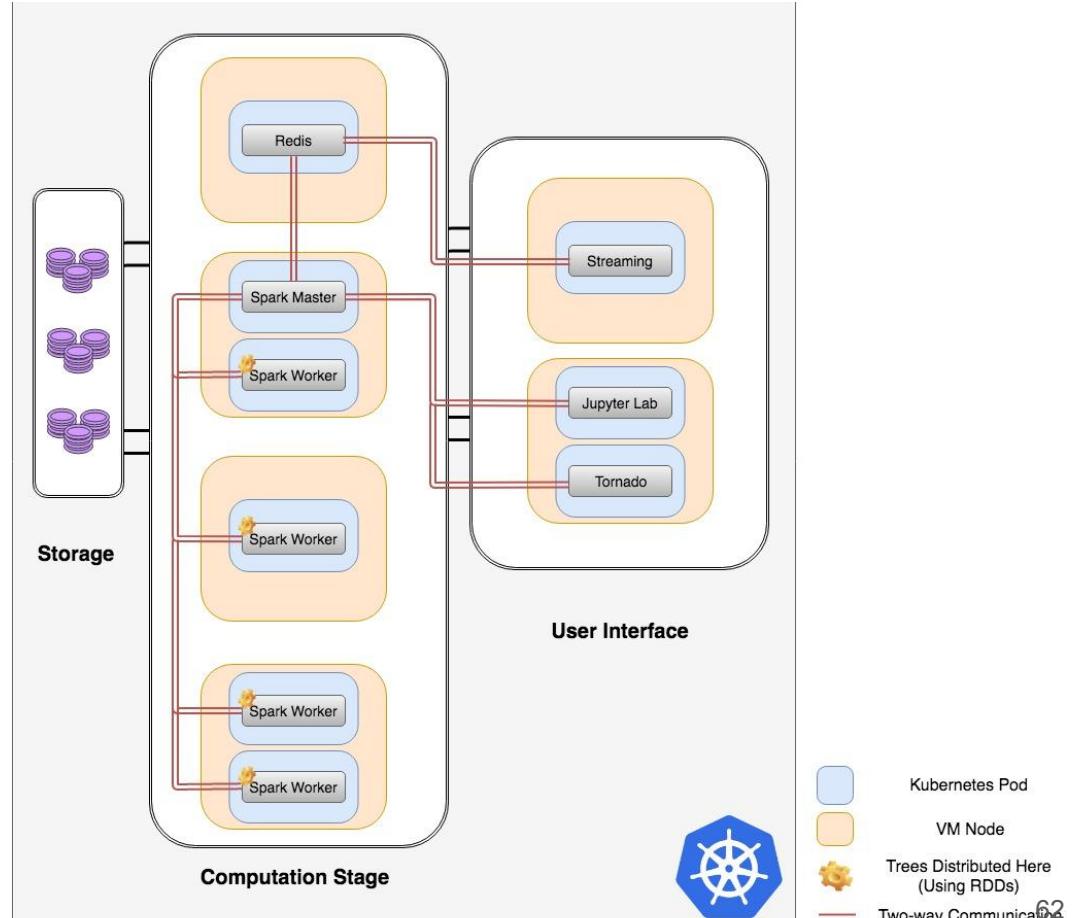
- NFS (Kubernetes PV/PVC)
- Redis
- RDD for Trees and Spark

User Interface:

- Jupyter notebooks
- Interactive web app for submitting jobs
- Streaming service

Computation Stage:

- Spark Master and Workers
- Communicator with Spark Master
- Subscription



Deployment

- Kubernetes allows very easy deployment, orchestration, scalability, resilience, replication, workloads and more
- Federation of services and Jobs
- From 0 to anomaly service → in minutes and config files
- Scale up/down (spark cluster and front-end) → Auto-scaling as an option
- Prototype support multiple users/projects, batch and streaming process
- Fault tolerant, disaster recovery



Example: Jupyter Notebooks

jupyter IFFParallelExample Last Checkpoint: 4 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Trusted Python 3 O

Create Spark Context

```
In [123]: from pyspark import SparkContext, SparkConf
In [124]: conf = SparkConf().setAppName("JupyterExamples").setMaster("spark://spark-master:7077")
conf.set("spark.cores.max", 4)
Out[124]: <spark.conf.SparkConf at 0x7f7419428470>
In [134]: if sc:
    sc.stop()
sc = SparkContext(conf=conf)
```

Imports

```
In [135]: import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import multivariate_normal
import random as rn
import iso.forest as iso
import seaborn as sb
import time
sb.set_style(style="whitegrid")
sb.set_color_codes()
```

Helper Functions

```
In [136]: def getBlobData(N=2000):
    mean = [10, 1]
    cov1 = [[10, 0], [0, 1]] # diagonal covariance
    Nobj = 4000
    x, y = np.random.multivariate_normal(mean, cov, Nobj).T
    #Add manual outlier
    x[0]=3.3
    y[0]=3.3
    x=np.array([x,y]).T
    plt.figure(figsize=(7,7))
    plt.scatter(x,y,s=45,c=[0.5,0.5,0.5],alpha=0.3)
    plt.show()

    return (x,y)
In [137]: def getMultiblobData(N=2000):
    mean1 = [10, 0]
    cov1 = [[1, 0], [0, 1]] # diagonal covariance
    mean2 = [0, 10]
    cov2 = [[1, 0], [0, 1]] # diagonal covariance
```

jupyter IFFParallelExample Last Checkpoint: 5 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Trusted Python 3 O

```
plt.plot(X[:,0],X[:,1],'o', markersize=10, color=[0.5,0.5,0.5], alpha=0.3)
plt.axis('equal')
plt.show()
return (x,y,X)

In [138]: def getSinusoidData(N=4000):
    x = np.random.rand(N)*8*np.pi
    y = np.sin(x) + np.random.randn(N)/4.

    #Add manual outlier
    x[0]=3.3
    y[0]=3.3
    X=np.array([x,y]).T

    fig=plt.figure(figsize=(7,7))
    fig.add_subplot(111)
    plt.plot(X[:,0],X[:,1],'o', markersize=10, color=[0.5,0.5,0.5], alpha=0.3)
    plt.show()
    return (x,y,X)

In [139]: def partition(l,n):
    return l[i:i+n] for i in range(0,len(l),n)

In [140]: def runIF(X):
    data = sc.parallelize(partition(X,int(len(X)/8)))
    forest = data.map(lambda x: iso.IForest(x,ntrees=100, sample_size=256))
    S_t = forest.map(lambda F: F.compute_paths(X))
    S = S_t.reduce(lambda a,b: a+b)
    return S

In [141]: def plotResults(x,y,scores):
    plt.rcParams['figure.figsize'] = (15, 5)
    plt.figure()
    plt.subplot(1,2,1)
    plt.densityplot(score, kde=True, color=[0.5,0.5,0.5])
    plt.xlabel('Anomaly Score', fontsize=20)
    plt.subplot(1,2,2)
    sns.prcplot(score)
    plt.scatter(x,y,s=45,c=[0.5,0.5,0.5],alpha=0.3)
    plt.scatter(x[ss[-10:]],y[ss[-10:]],s=55,c='r')
    plt.scatter(x[ss[:10]],y[ss[:10]],s=55,c='g')
    plt.show()
```

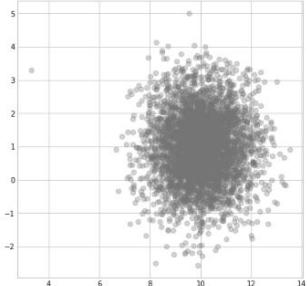
Examples

Example: Jupyter Notebooks

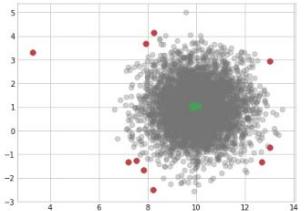


Blob

```
In [148]: x,y,X = getBlobData()
```

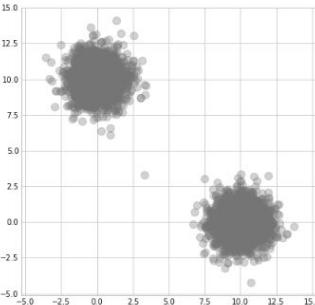


```
In [149]: S = runIF(X)  
plotresults(x,y,S)
```

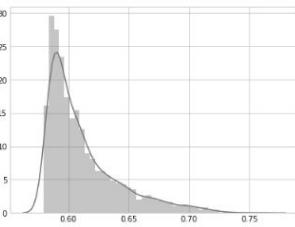
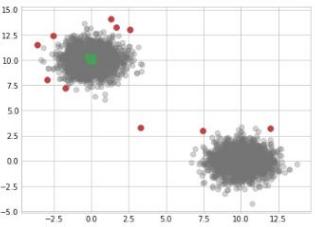


MultiBlob

```
In [150]: x,y,X = getMultiBlobData()
```



```
In [151]: S = runIF(X)  
plotresults(x,y,S)
```



Final Remarks

Matias Carrasco Kind -- NCSA
mcarras2@illinois.edu
github.com/mgckind
matias-ck.com

- It's all about the user
- Jupyter as Scientific tool but not only solution
- Science on the cloud is happening in many scientific fields including Astronomy
- Containerized solutions to ease management of the applications
- HPC is adopting cloud technologies to leverage the benefits of both worlds
- Kubernetes provide means to have 'the cloud' outside the commercial world
- Production services for large datasets
- YOU are not alone

... this is changing the way we do astronomy

Thank you!

Questions?

Matias Carrasco Kind -- NCSA

mcarras2@illinois.edu

github.com/mgkind

matias-ck.com