



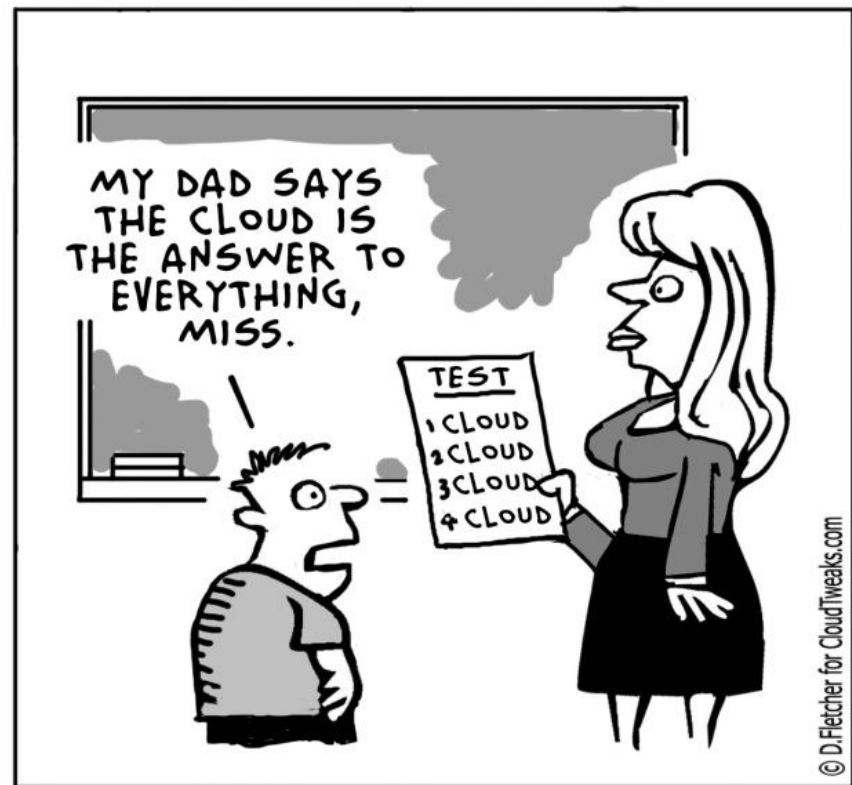
Leveraging current and future technologies for scientific discovery

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

AstroData2020s Science Workshop, Caltech
Dec 4th, 2018

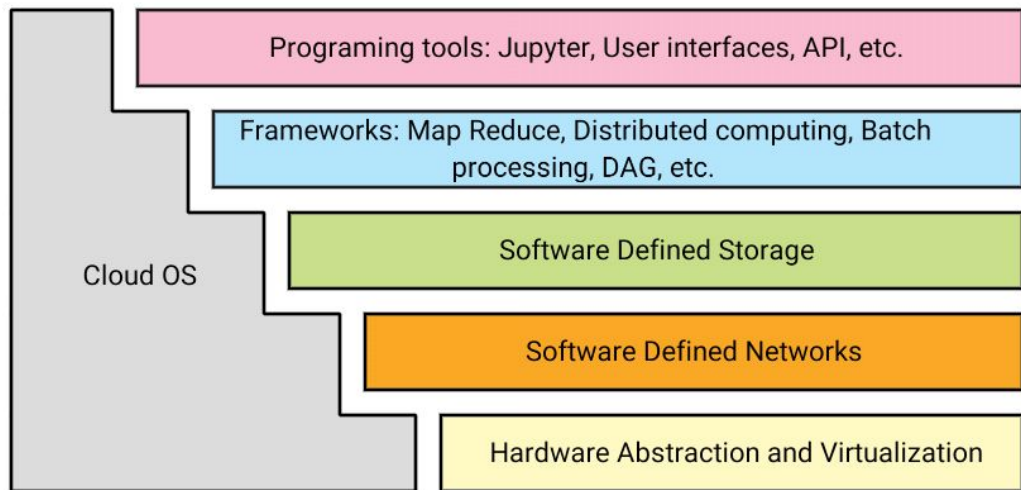
Scientific Cloud Computing

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



Why we should be doing science on the cloud

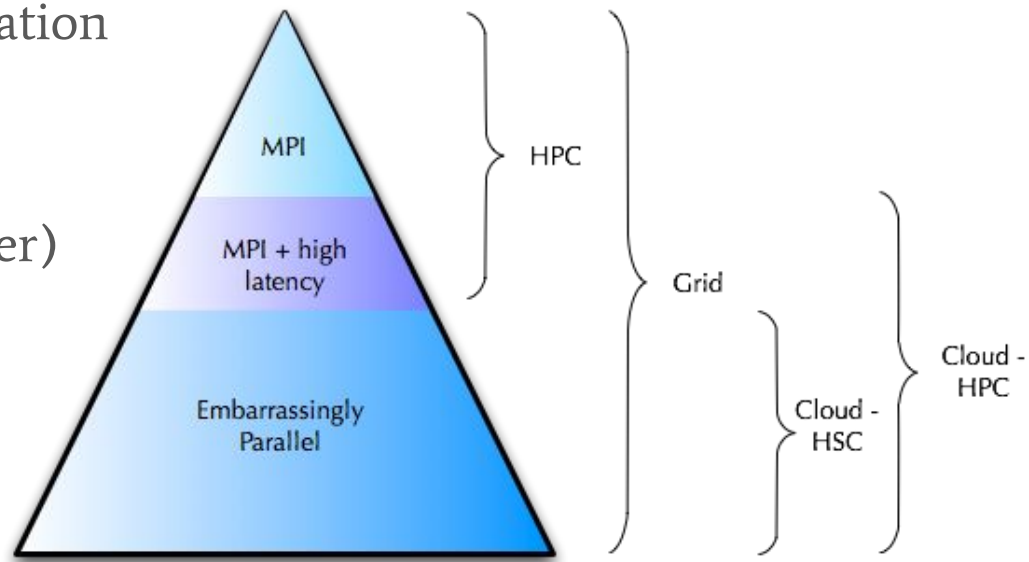
- Remote and dynamic data (!= Big data)
- Big data \Rightarrow Data Gravity, Data Lake, etc.
- Remote software/server
- Easy to deploy*
- Asynchronous
- Web applications / Shareable
- Serverless applications
- Federation of Services
- Tablets/ChromeOS
- more...



*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 ...

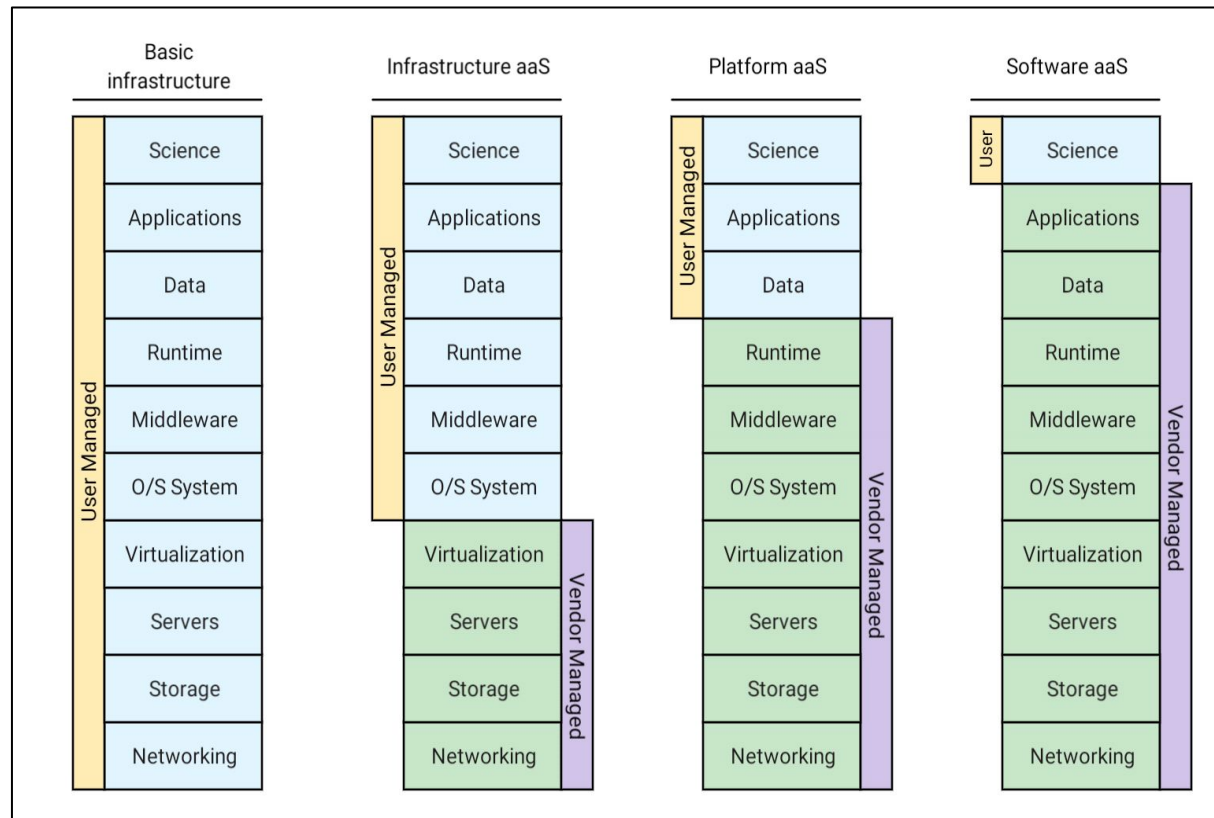


[^]arguable

*arguable (CI, CD)

What kind of science/projects? → Which model

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



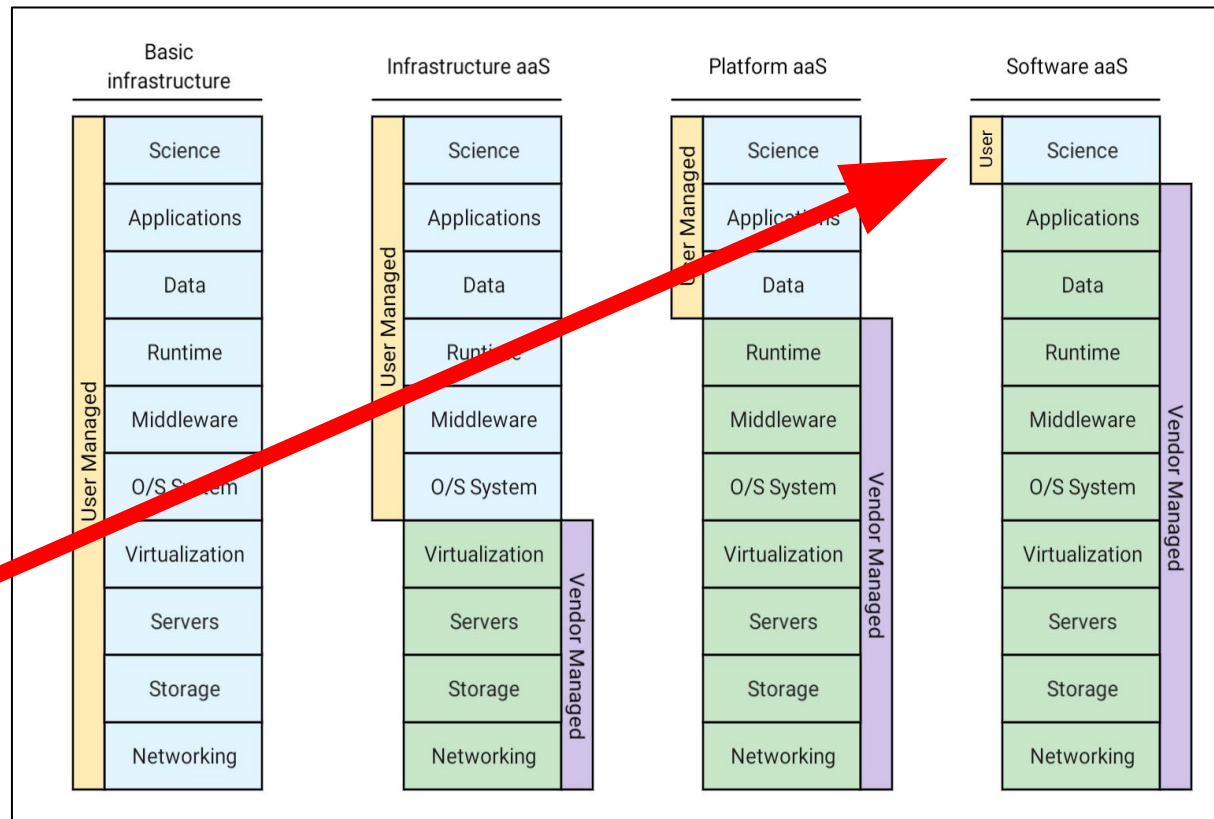
*arguable

What kind of science/projects? → Which model

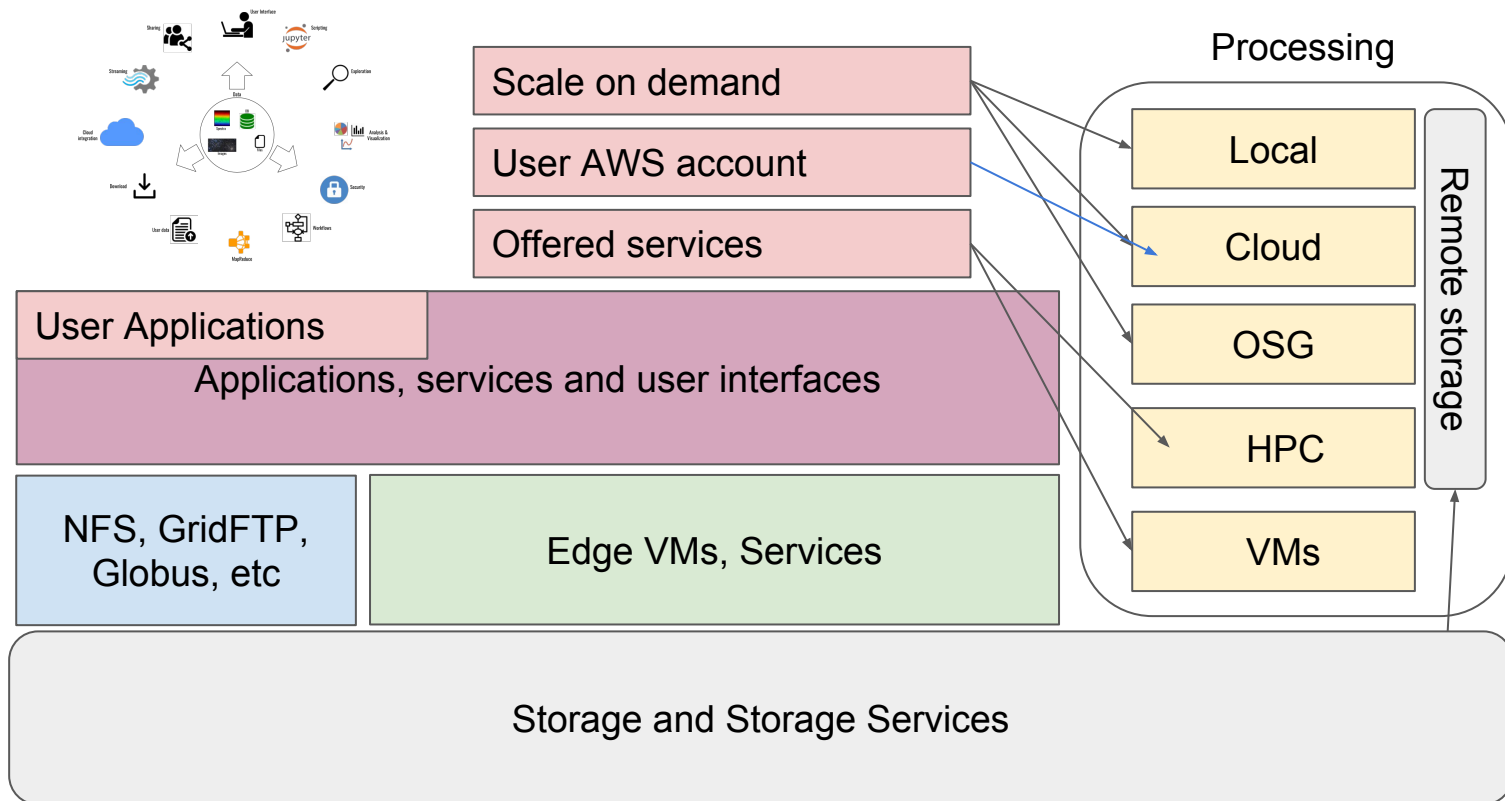
- HTC vs HPC vs HSC
- Interactive
- Small projects
- Visualizations
- Short term projects*

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

*arguable



Why not both?



I What kind of users/groups?

- Runaway users (sleep infinity)
- Super user 1 → Data In Intensive Jobs (ML)
- Super user 2 → Data Out Intensive Jobs (Sims)
- Super user 3 → Resource Intensive Jobs (Modeling/Fitting)
- User with no exposure/experience
- Users with computers to provide
- Users with money to provide
- Ephemeral users (workshops)

CLOUD NATIVE TRAIL MAP

The Cloud Native Landscape [Landscape](https://landscape.cncf.io) 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/ksps

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 this 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/ck
- Helm Charts help you define, install, and upgrade even the most complex Kubernetes application



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.

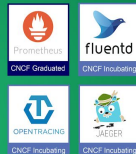


2. CI/CD

- 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
- Setup 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-RPC, 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.



App Definition and Development

Database Vitess, CockroachDB, ScyllaDB, Cassandra, Redis, Aerospike, MongoDB, PostgreSQL, MySQL, Oracle Database, InfluxDB, Kudu, HBase, Vertica, YugabyteDB, etc.	Streaming & Messaging NATS, Apache Kafka, RocketMQ, Pulsar, etc.	Application Definition & Image Build Helm, Apache Maven, Jenkins, etc.	Continuous Integration & Delivery Argo, Jenkins, GitLab CI, GitHub Actions, etc.
--	--	--	--

Orchestration & Management

Scheduling & Orchestration Kubernetes, Nomad, etc.	Coordination & Service Discovery CoreDNS, Consul, etcd, NACOS, etc.	Remote Procedure Call gRPC, Apache Thrift, etc.	Service Proxy Envoy, NGINX, HAProxy, etc.	API Gateway Kong, MuleSoft, Tyk, etc.	Service Mesh Linkerd, Netflix Zuul, etc.
--	---	---	---	---	--

Runtime

Cloud-Native Storage MinIO, Ceph, Gluster, etc.	Container Runtime Docker, rkt, CRI-O, etc.	Cloud-Native Network CNV, Calico, Flannel, etc.
---	--	---

Provisioning

Automation & Configuration Ansible, Puppet, etc.	Container Registry Harbor, Quay, etc.	Security & Compliance Falco, Clair, etc.	Key Management Vault, etc.
--	---	--	--------------------------------------

Public

AWS, Google Cloud, Huawei, etc.

l.cncf.io

This landscape is intended as a map through the previously uncharted terrain of cloud native technologies. There are many routes to deploying a cloud native application, with CNCF Projects representing a particularly well-traveled path.

CLOUD NATIVE Landscape

CLOUD NATIVE COMPUTING FOUNDATION

Redpoint Amplify

Platform

Certified Kubernetes - Distribution

Certified Kubernetes - Hosted

Certified Kubernetes - Installer

PaaS/Container Service

Observability and Analysis

Monitoring

Logging

Tracing

Chaos Engineering

Serverless

See the serverless interactive display at l.cncf.io

Cloud

Kubernetes Certified Service Provider

Kubernetes Training Partner

Special

Tools



Security



Framework



Hosted

Installable

Platform

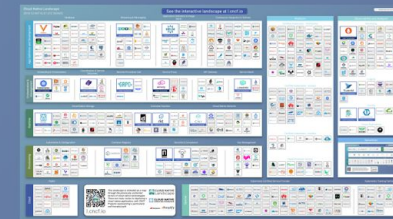


Cloud Native Landscape



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

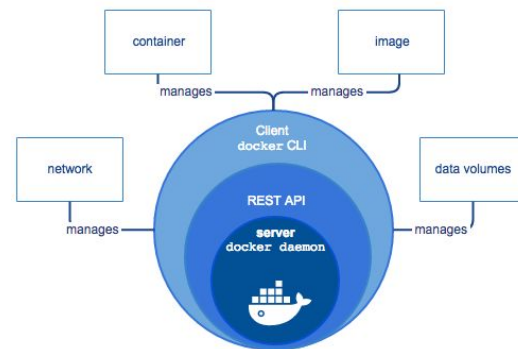


And there are many more...



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




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

easyaccess: DES command line tool



```

DARK ENERGY SURVEY
DATA MANAGEMENT

easyaccess 1.4.0. The DESDM Database shell.
Connected as mcarras2 to desdct.
** 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

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 DR1 Infrastructure

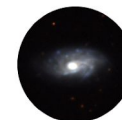
Easyaccess web



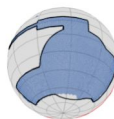
Jupyterhub + easyaccess



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

Welcome to dessci, Matias!



DB ACCESS

Oracle SQL web-client

[More...](#)



DES TABLE SCHEMA

Browse all tables

[More...](#)

```
SELECT dr1.RA,dr1.DEC,dr1.COADD_OBJECT_ID
FROM dr1_main sample(8.01) dr1
WHERE
dr1.MAG_AUTO_G < 18 and
dr1.WAVG_SPREAD_MODEL_I + 3.0*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.WAVG_SPREAD_MODEL_I > -1 and
dr1.IMAFLAGS_ISO_G = 0 and
dr1.IMAFLAGS_ISO_R = 0 and
dr1.IMAFLAGS_ISO_I = 0 and
```

EXAMPLE QUERIES

See some example queries as a start

[More...](#)



CUTOUTS SERVICE

Generate cutouts for positions or ids

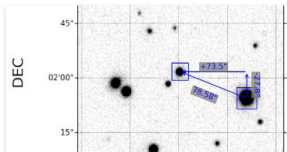
[More...](#)



DES JupyterLabs

(Beta) Jupyter Labs

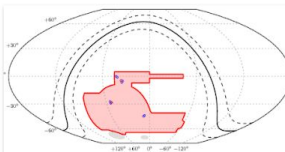
[More...](#)



FINDING CHART

Find your object

[More...](#)



DES FOOTPRINT

Interactive globe

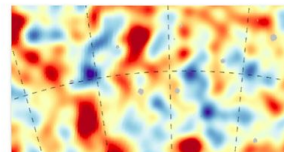
[More...](#)



DATA ANALYSIS

SEDs and color-color diagrams

[More...](#)



MY JOBS

List of submitted jobs

[More...](#)



HELP

Help form

[More...](#)

Home

DB access

DES Table Schema

Example Queries

Cutouts Service

DES JupyterLab

Finding Chart

DES Footprint

Data Analysis

My Jobs

Help

NCSA DESaccess: DB access

DARK ENERGY SURVEY desaccess



mck

mcarras2@illinois.edu

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 glubular 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_PSF,
9 WAVG_MAG_PSF_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

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

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

Coads 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)
 Upload File
- 📄 Enter Values
 Enter Values
- 📏 Xsize (in arcminutes): 1.0
- 📏 Ysize (in arcminutes): 1.0
- ✍️ Job Name
- ✉️ Email Options

Send email on completion Email
- 📁 Return Type

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

Clear Form
Submit Job


NCSA DESaccess: Cutouts Service



DARK ENERGY SURVEY desaccess

Job1 : d927a264_746c_4f7a_82cd_f46ebce496c7 (19 objects)

See Log
Download All
Close



mck
mcarras2@illinois.edu

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

Coadds Images Cutout Form

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

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

📄 Enter Values

📏 Xsize (in arcminutes): 1.0

📏 Ysize (in arcminutes): 1.0

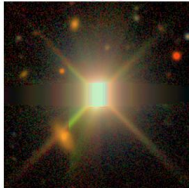
✍️ Job Name

✉️ Email Options

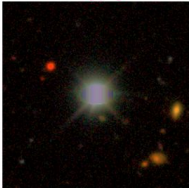
📄 Return Type

⚙️ Clear Form

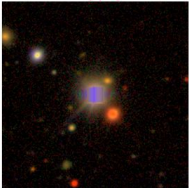
143782572_irg.png



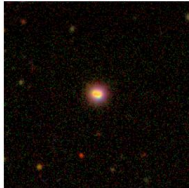
143830746_irg.png



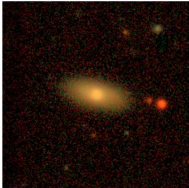
143821280_irg.png



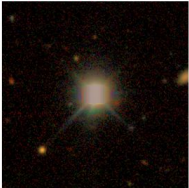
197515917_irg.png



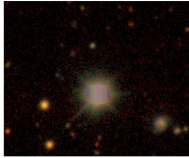
154564429_irg.png



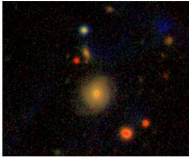
206021039_irg.png



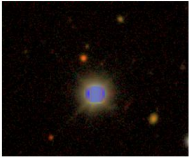
179484171_irg.png



149989928_irg.png



143470014_irg.png



👤

?

1

1

NCSA DESaccess: Asynchronous Jobs

DARK ENERGY SURVEY desaccess



mck
mcarras2@illinois.edu

My Jobs

#	Status	Job Name	Job type	Execution time (s)	Cancel Job	Queries	Results	Files
0	●	Name: Job id: 6b4cac2b-b544-44e1-96bf-58cd4968a338 6 days and 0 hours ago (Expired)	query	0	⊗	Query	Cutouts	Files
1	●	Name: Job id: daf5e3c-461e-42ed-8efb-5fcbf684047 6 days and 0 hours ago (Expired)	cutout	1	⊗	Query	Cutouts	Files
2	●	Name: testapi Job id: 0dfc5a58-b00a-4798-834f-9816c6f9a985 7 days and 4 hours ago (Expired)	cutout	3	⊗	Query	Cutouts	Files
3	●	Name: testapi Job id: 12961656-8075-4629-8e4f-fd4378013634 7 days and 4 hours ago (Expired)	cutout	3	⊗	Query	Cutouts	Files
4	●	Name: testapi Job id: 09a37f69-2096-4296-b87d-c6567cde0649 7 days and 4 hours ago (Expired)	cutout	1	⊗	Query	Cutouts	Files
5	●	Name: Job id: fd10cf32-3cc6-4090-bb90-344268dd615e 7 days and 5 hours ago (Expired)	cutout	1	⊗	Query	Cutouts	Files
6	●	Name: testapi Job id: b85ea747-5201-4e49-a0eb-f2b667f266de 7 days and 5 hours ago (Expired)	cutout	-1	⊗	Query	Cutouts	Files
7	●	Name: Job id: 8f8fa56a-4685-49f9-b7be-603310ccdddb 8 days and 16 hours ago (Expired)	query	577	⊗	Query	Cutouts	Files
8	●	Name: Job id: df8a57c4-b1d5-4332-80d5-a08a27b537d9 8 days and 16 hours ago (Expired)	query	1042	⊗	Query	Cutouts	Files
9	●	Name: Job id: 7f1db550-4d38-441f-a037-ed659b3b79c9 8 days and 16 hours ago (Expired)	query	-1	⊗	Query	Cutouts	Files
10	●	Name: Job id: f0aacde0-9d63-45d4-92f2-4f847b9b415c 8 days and 16 hours ago (Expired)	query	9	⊗	Query	Cutouts	Files
11	●	Name: Job id: a88b79cc-fd71-4ee0-a33d-92b5be98106f 8 days and 17 hours ago (Expired)	query	9	⊗	Query	Cutouts	Files
	●	Name: demo1						

REFRESH ↻

DELETE 🗑

Home

DB access

DR1 Table Schema

Example Queries

Cutout Service

DR1 Footprint


My Jobs

DES JupyterLab

Help

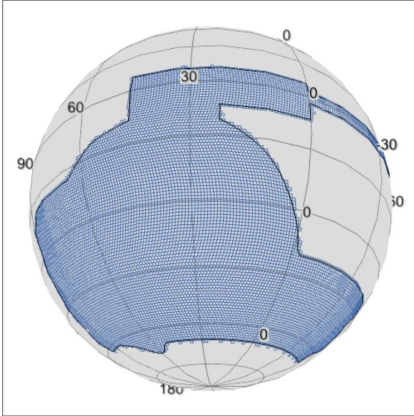
NCSA DESaccess: Footprint and Jupyter Labs

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 :

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.

Home

DB access

DR1 Table Schema

Example Queries

Cutout Service

DR1 Footprint

My Jobs

DES JupyterLab

Help

Status

● Ready
 Status: Running

REFRESH ↻

NCSA DESaccess: Labs with access to Jobs and easyaccess

The screenshot displays a JupyterLab environment with three main components:

- File Browser (Left):** Shows a directory structure under 'jobs' with various folders and files, including their last modified dates.
- Plotting Window (Center):** A scatter plot titled 'basics_plotting' showing a positive correlation between 'MAG_AUTO_R' (x-axis) and 'MAG_AUTO_I' (y-axis). The plot includes a density heatmap and marginal distributions on the top and right axes.
- Terminal (Right):** A terminal window titled 'Terminal 4' showing the output of the 'easyaccess' command. It displays the 'DARK ENERGY SURVEY DATA MANAGEMENT' logo and lists available commands for the DESDB database shell.

Below the plot, the following code is shown in the JupyterLab cell:

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

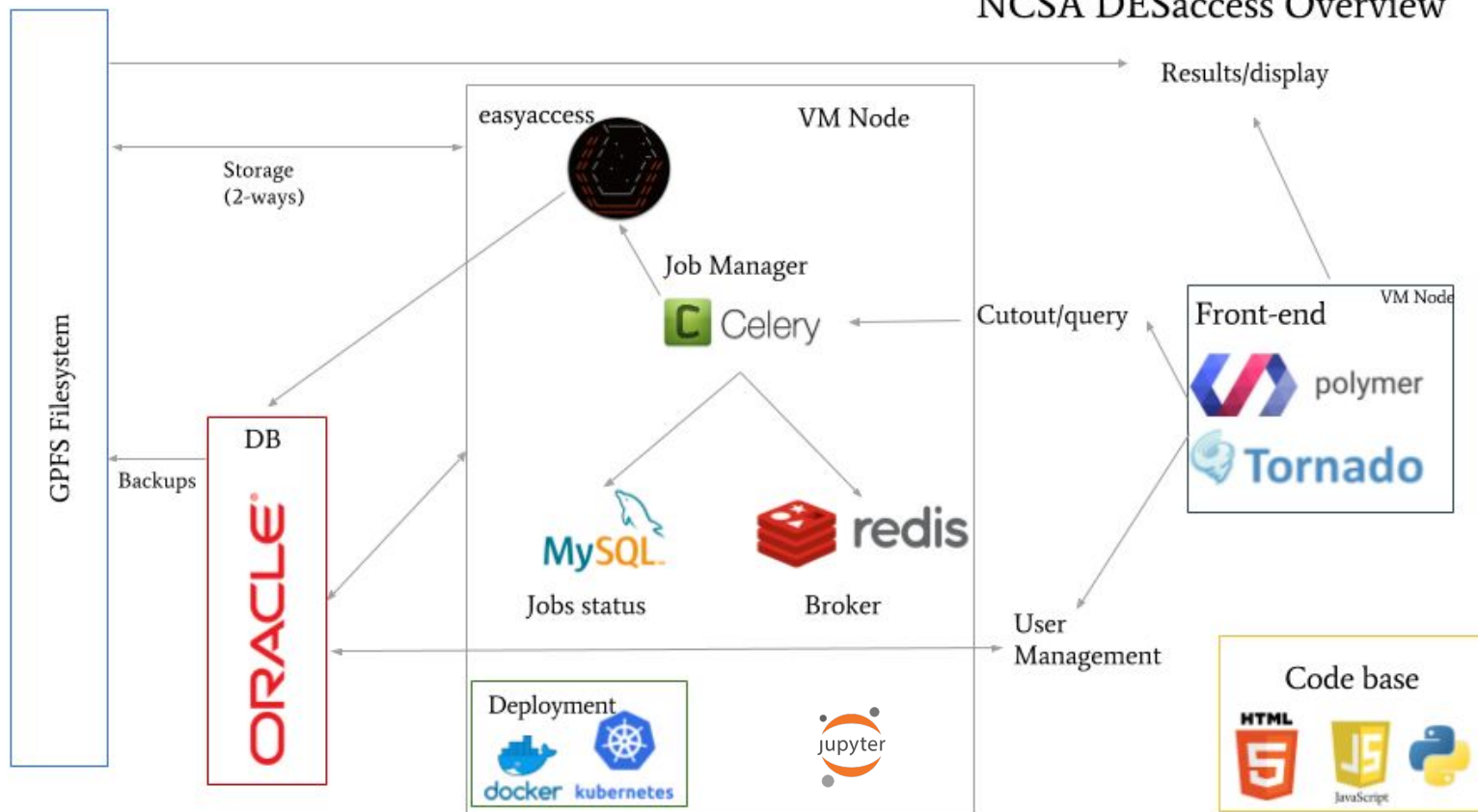
In [10]: hextiles = hv.HexTiles(df, [('MAG_AUTO_R', 'R'), ('MAG_AUTO_I', 'I')], [], extents=(20,26,20,26)

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

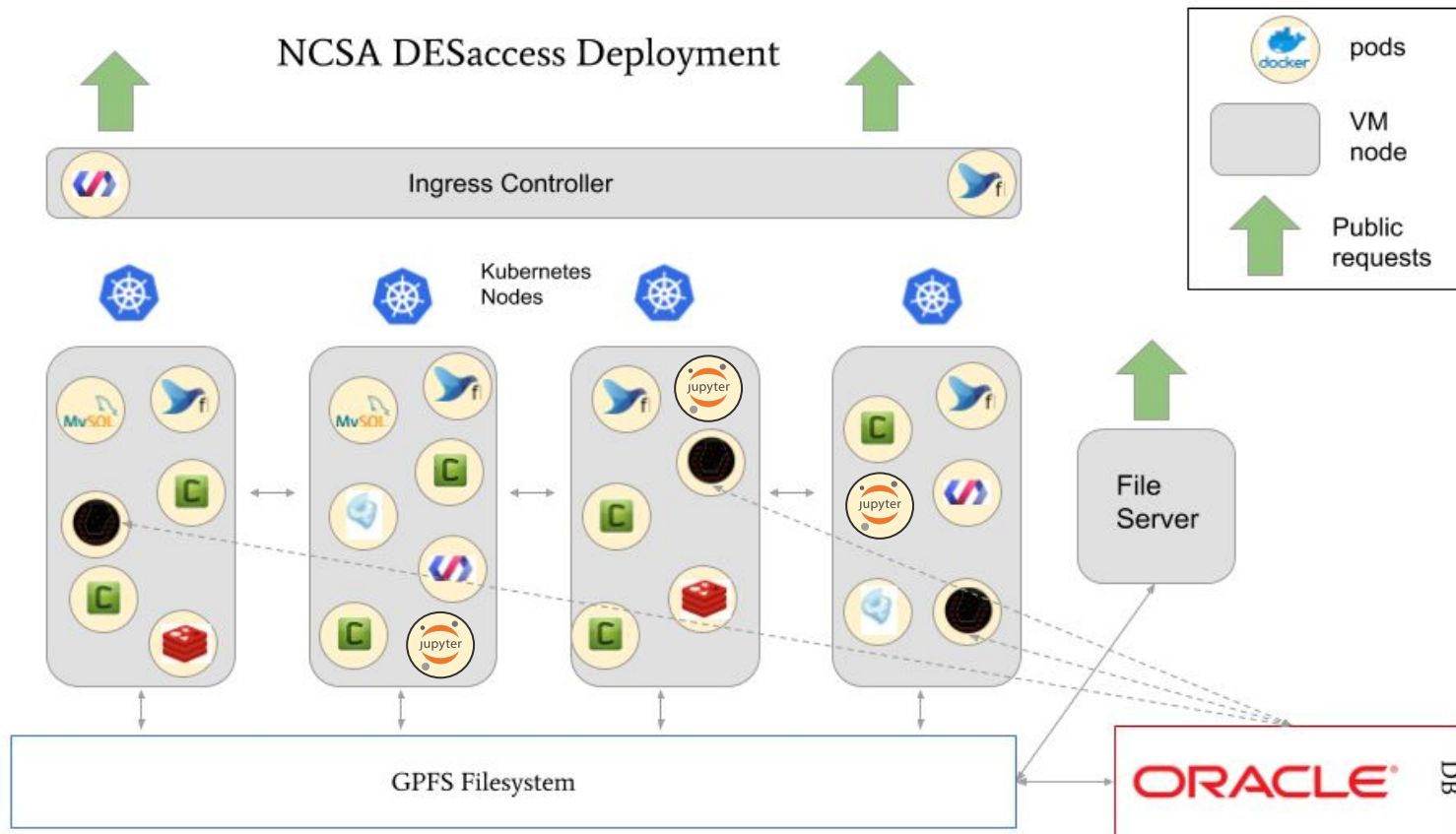
The output of the code is a hexagonal heatmap visualization of the data, with a colorbar on the right indicating values from 0 to 800.

NCSA DESaccess: Technology Overview

NCSA DESaccess Overview



NCSA DESaccess: Deployment



Thank you!

Questions?

Matias Carrasco Kind -- NCSA

mcarras2@illinois.edu

github.com/mgkind

matias-ck.com