Clouds, Clusters, and Containers: Tools for responsible, collaborative computing

Matt Vaughn @mattdotvaughn, John Fonner

#cyverse #agaveapi #usetacc

Part One: Overview and Introductions

You should have a Cyverse user account https://user.cyverse.org/ ready to go in order to be productive in the next sessions

What is Cloud?

We generally care about **reliably expanding our capacity and capability**

We generally don't want to care about **monitoring, business models, developments in systems architecture, hardware**

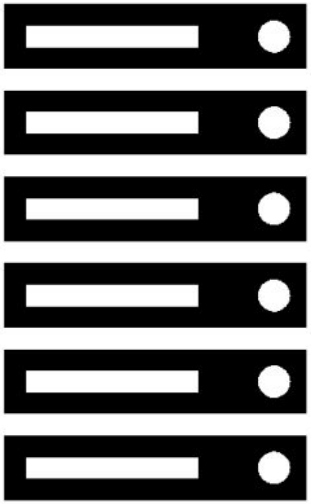
Cloud is a useful abstraction that means that the things we don't want to mess with are someone else's problem

But… it can bring its own challenges

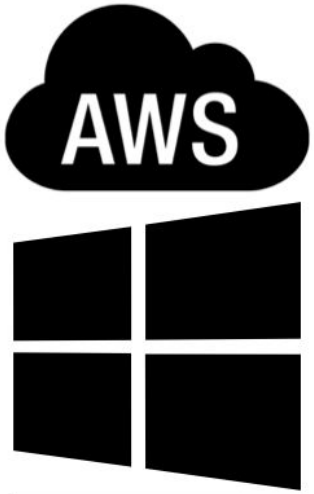
- Reproducibility
- Need for high-level IT skills to use it
- Paying for it

~4 CPU Cores

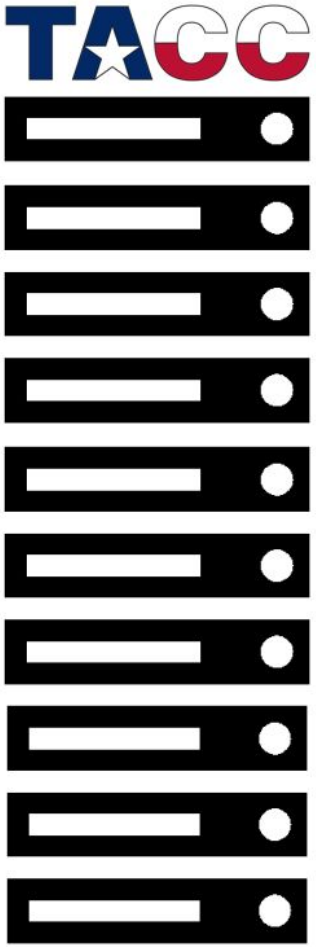
~100 CPU Cores



1,000+ CPU Cores



500,000+ CPU Cores



Hammers, scalpels, and scopes

Hammers

- Leadership systems: Stampede. Comet
- Big clusters: Lonestar, Hikari, Bridges

Scalpels

- Data intensive systems: Wrangler, Rustler
- Architecture Experiments: Catapult, Fabric
- Viz and GPU compute: Maverick, Stallion, Lasso

Scopes

- User-provisioned cloud: Chameleon, Jetstream
- Global FS: Stockyard
- Specialized interfaces: APIs, SaaS

What does Big Data feel like?

What kind of characteristics are commonly associated with Big Data?

1. Physical constraints
2. Big (meta)data volume
3. Big compute
4. Big memory
5. Slow networks
6. Bad algorithms

How are people handling Big Data?

- MapReduce: Hadoop, Storm
- Event & Streaming processing: Kinesis, Azure Stream Analytics, Camel, Streambase
- Machine Learning: Watson, Azure BI, SAS
- In-memory processing: Kognito, Apache Spark
- New data warehouse: Snowflake,
- FauxSQL

Today's **Big Data** solutions strangely resemble **distributed execution** frameworks with slightly **different schedulers**.

Scientific Big Data is a cultural problem

Mental challenges

- (Enterprise) Integration scenarios
- Software portability
- IT administration
- Performance tuning
- Security
- Provenance
- Reproducibility
- Technology changes

Scientific Big Data is a cultural problem

Social challenges

- Collaboration
- Publishing
- Ownership
- Attribution
- Team dynamics

Scientific Big Data is a cultural problem

Economic challenges

- Infrastructure operations
- Data preservation
- Software maintenance
- Copyright

Scientific Big Data is a cultural problem

Legal challenges

- Copyright
- Purchasing
- HIPAA (and other privacy frameworks)
- Export control

Impactful “Big Data” **solutions** won’t be found along a single axis. The next silver bullet will **look like a shotgun**.

THE AGAVE PLATFORM

DELIVERING SCIENCE-AS-A-SERVICE IN TODAY'S HYBRID CLOUD ENVIRONMENT

What is Agave?

Agave is a multi-tenant PaaS solution delivering

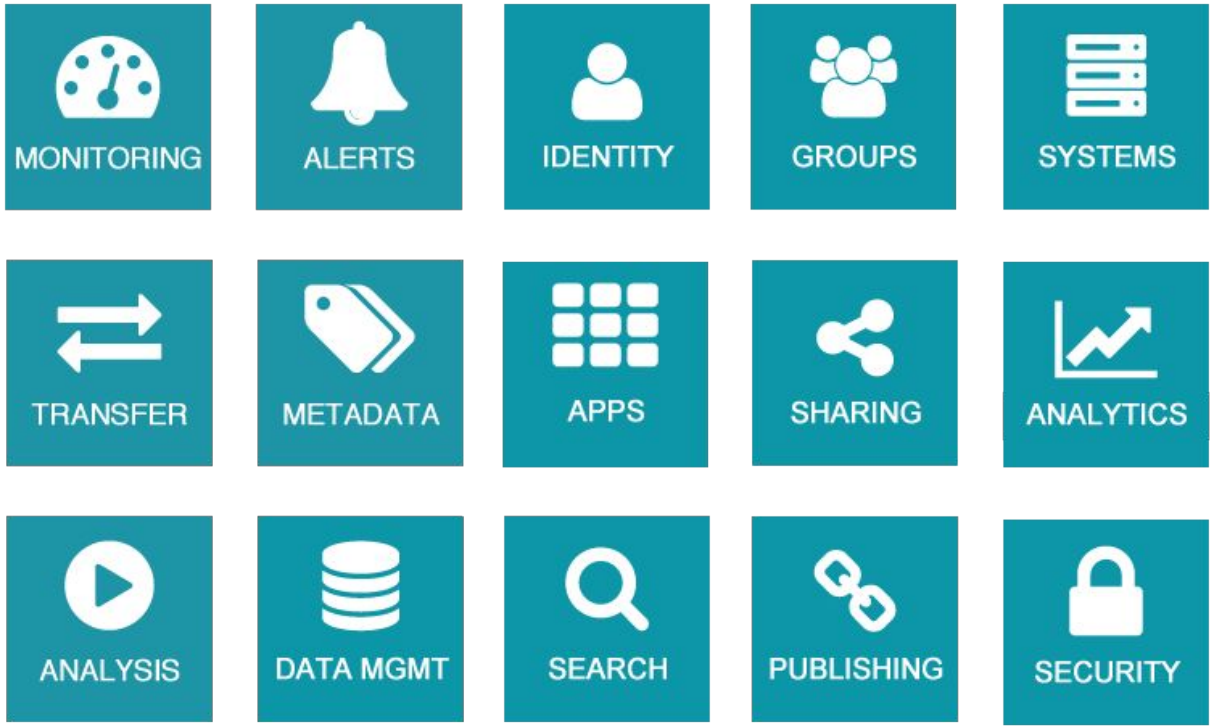
Science-as-a-Service

capabilities across hybrid cloud environments.

What does it do?

- **Run application codes**
  your own or community provided codes
- **...on HPC, HTC, and cloud resources**
  your own, shared, or commercial systems
- **...and manage your data**
  reliable, multi-protocol, async data movement
- **…in a collaborative way**
  fine grain ACL for working securely with others
- **...from the web**
  webhooks, rest, json, cors, oauth2
- **...and remember how you did it**
  deep provenance, history, and reproducibility built in

No, seriously, what does it do?



White Label PaaS

- Build and brand for your organization
- Customize with your own services and features.
- Let us operate it or host it yourself

Zero Install Deployment

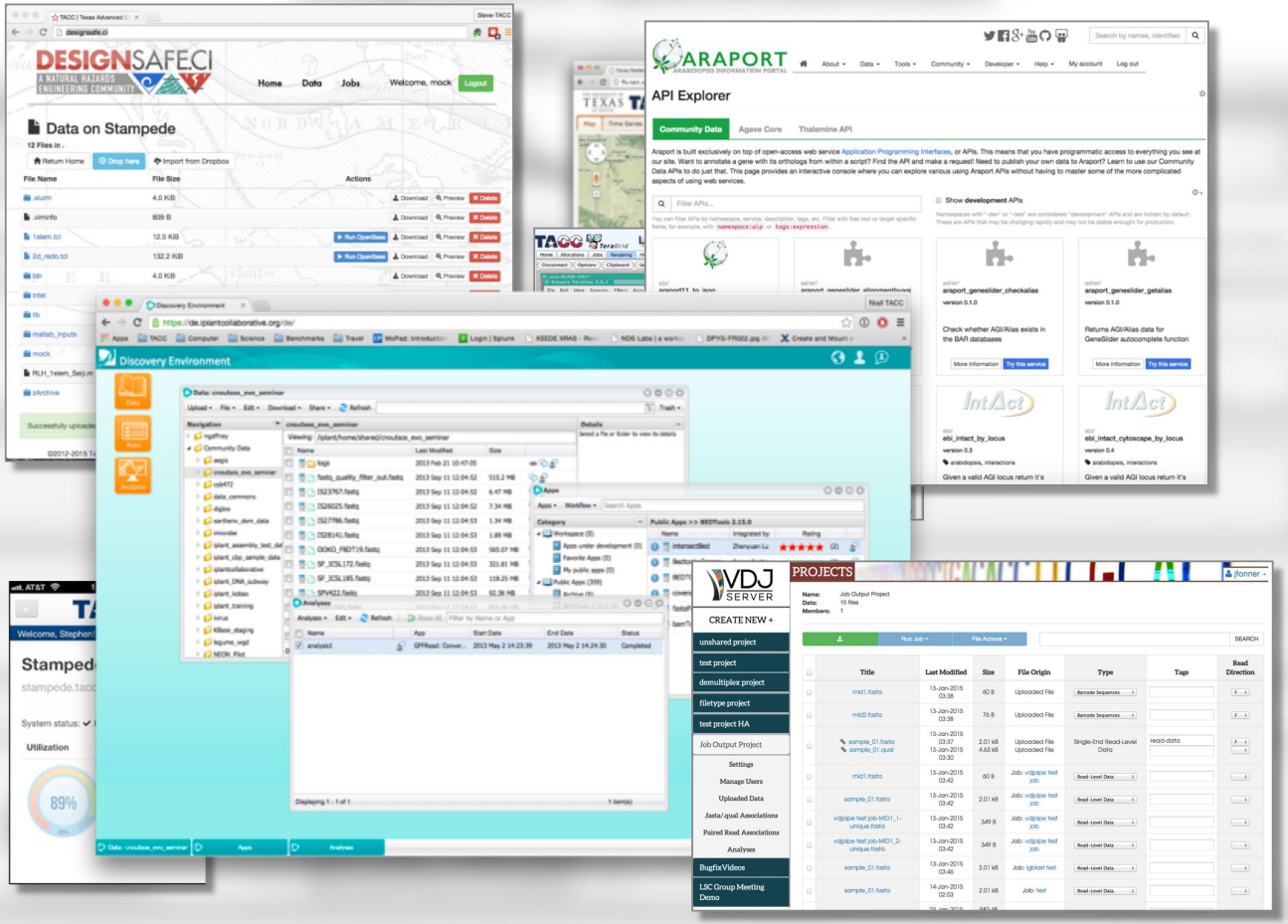
- Interacts with existing compute & storage
- Leverages your existing workload manager(s)
- Delegates to your existing IdP & security
- Uses your existing apps
- Creates a cohesive platform for your dev and user communities

Web friendly

- JSON in | JSON out
- Global ACLs on every resource
- Role-based management
- Public and private scopes for web publishing
- Sync and async interfaces
- Email & webhook notifications
- Event-driven design

Reproducibility As A Feature

- Deep provenance on everything
- Auto-capture contextual metadata
- Ability to re-run pipelines, processes, and data transfers baked in

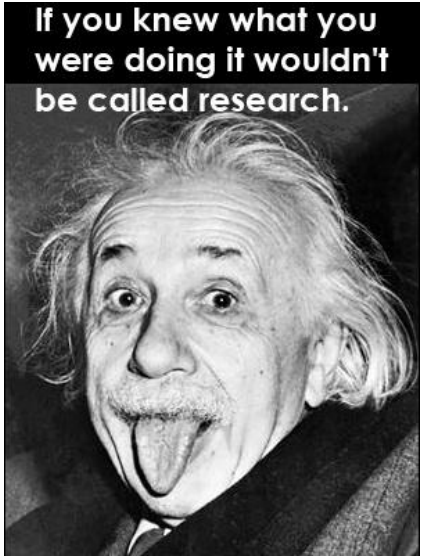


THE UNIVERSITY OF TEXAS AT AUSTIN

TEXAS ADVANCED COMPUTING CENTER

Containers for science

- Research is hard
- Coding is hard
- Research code is
  - ~~well designed~~,
  - ~~documented~~,
  - ~~leverages design patterns~~,
  - ~~highly reusable~~,
  - ~~portable~~,
  - and *usually* open source.



Scientists, with few exceptions, are ***not trained programmers***

Containers for science

- Truth be told, they don't actually even care.
- The ROI of better higher code quality ≈ 0
- No funding available for cleaning up code.

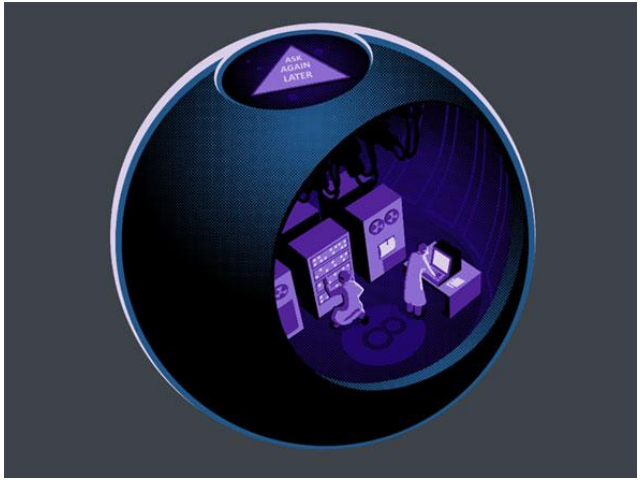
Despite the ***quality*** of the code, the ***science*** represented by the code is valuable and ***necessary*** for future discovery.

Containers for science

Compute containers are the Magic 8 Ball of science...

- Compartmentalize code
- Eliminate build and run complexities
- Introduce portability, reuse, & versioning
- Widgetize the creation of a scientific pipeline

...but better because results are reproducible.



Compute containers enable ***reproducible science*** via *composition.*

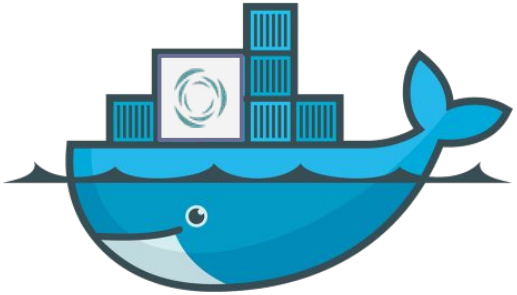
Containers for science

Data containers can serve as universal adapters between compute containers

- Transform data
- Bridge file systems
- Enable distributed data access
- Virtualize interfaces

Data containers enable clean integration between containers and ***standardize*** how we interact with ***distributed data***.

Containers are changing the landscape



- Cyverse has been an early adopter of container tech
  - Magic wand to make scientific software **deployed and usable**
    - Pushbutton Interfaces
    - Language-specific libraries
    - Scriptable CLI tools
- Galaxy, NIH Cancer Cloud Pilots, and lots of other folks are using them too

But they have their perils too...

- Managing and orchestrating containers + data + networking can be complicated
- There are a lot of emergent solutions
- We won’t touch on this today, but be careful in your technology selections