WORKSHOP MATERIAL

github.com/johnfonner/XSEDE16

SCIENTIFIC COMPUTING WITHOUT THE COMMAND-LINE: AN INTRO TO THE AGAVE API

Rion Dooley *@deardooley*

John Fonner *@johnfonner*

#agaveapi #usetacc

WHAT IS CLOUD?

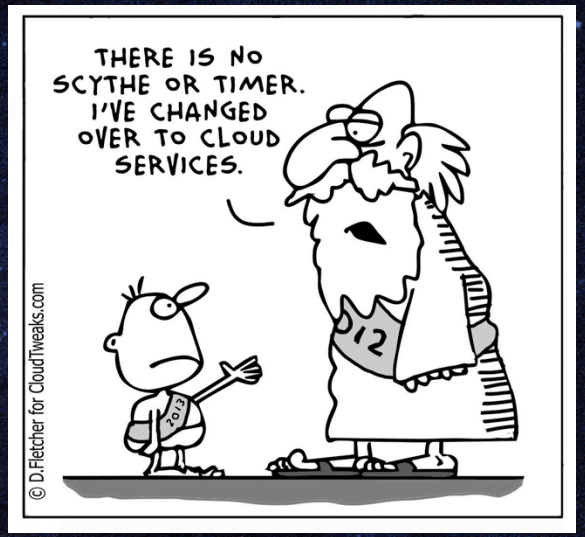
We generally care about **reliably expanding our capacity and capability**

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

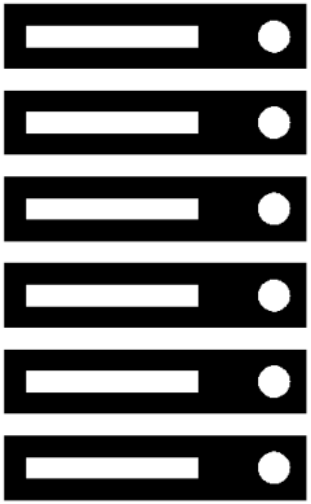
But… it can bring its own challenges

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

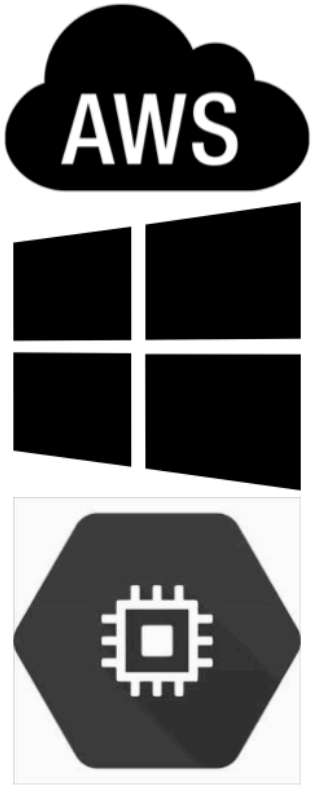


~4 CPU Cores

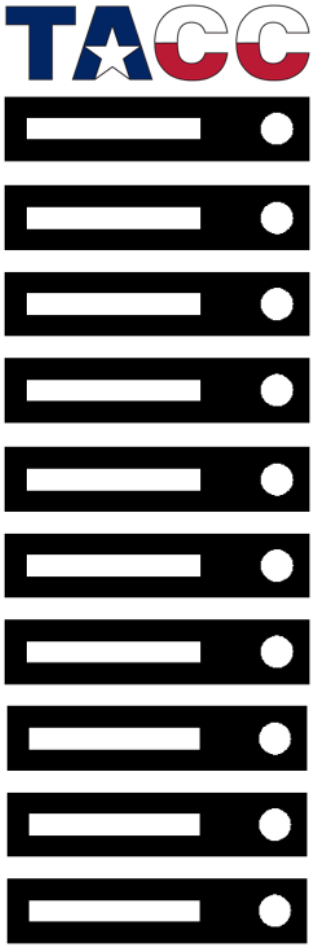
~100 CPU Cores



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

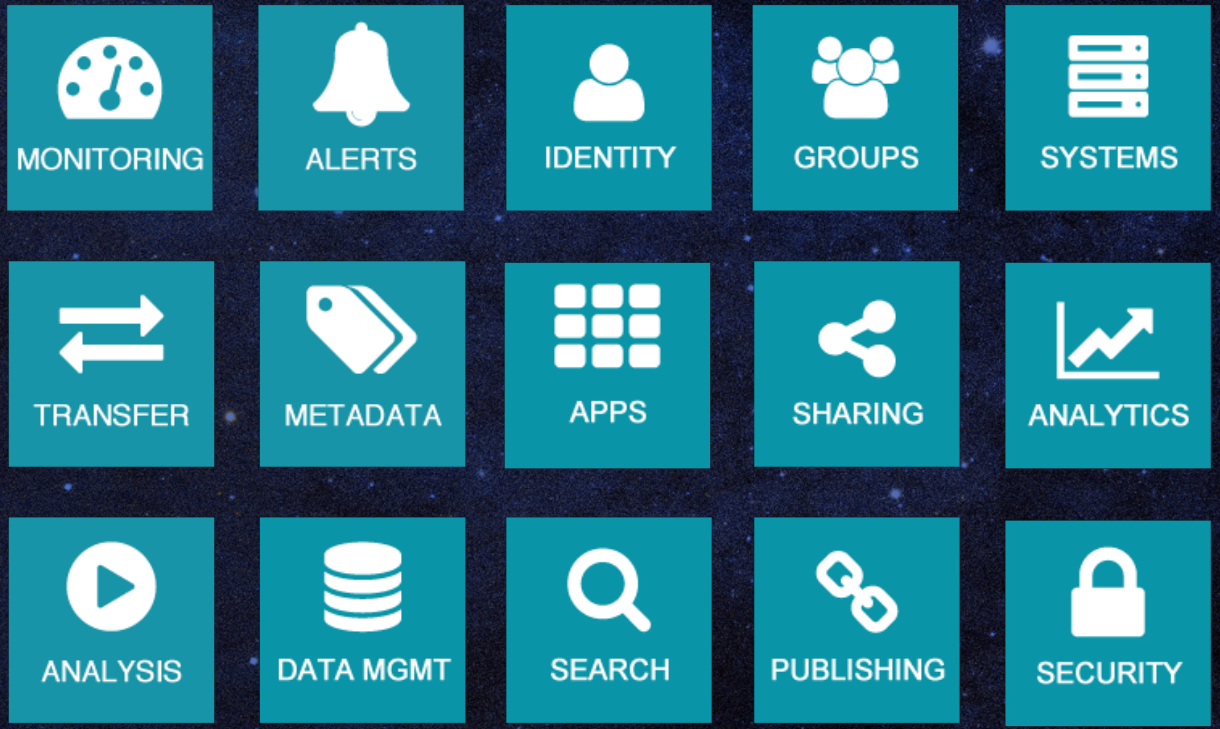
Science-as-a-Service

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

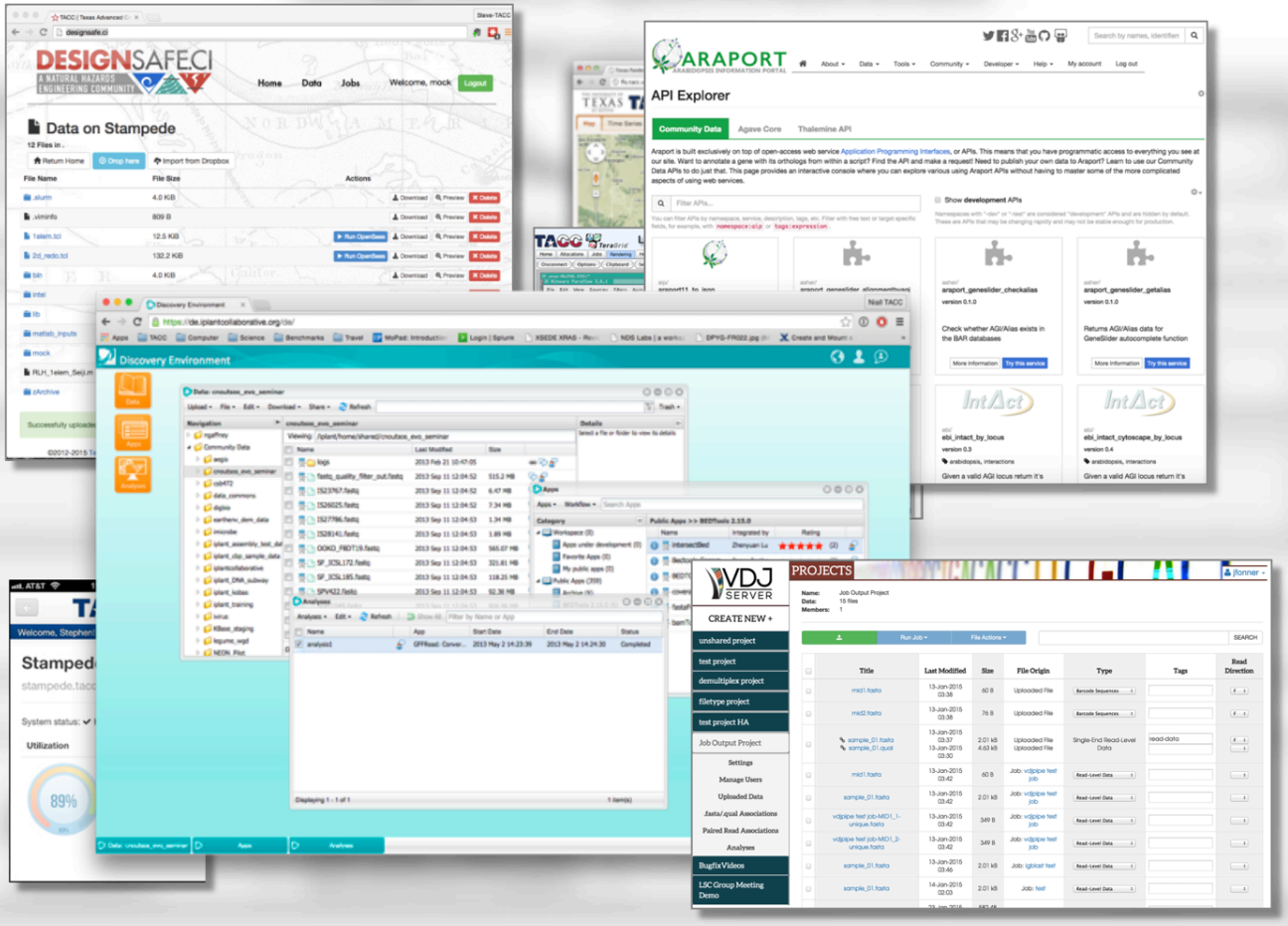
- Interacts with existing compute & storage
- Leverages your existing workload manager(s)
- Delegates to your existing IdP & security
- Uses your existing apps
- Creates a cohesive platform for your dev and user communities

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


DESIGNSAFECI
A NATURAL HAZARDS ENGINEERING COMMUNITY
Home
Data
Jobs
Welcome, mock
Logout
Data on Stampede
12 Files in .
Return Home
Drop here
Import from Dropbox
File Name
File Size
Actions
ARAPORT
ARABIDOPSIS INFORMATION PORTAL
API Explorer
Community Data
Agave Core
Thalemine API
araport_geneslider_checkalias
araport_geneslider_getalias
ebi_intact_by_locus
ebi_intact_cytoscape_by_locus
Discovery Environment
VDJ SERVER
CREATE NEW +
unshared project
test project
demultiplex project
filetype project
test project HA
Job Output Project
Settings
Manage Users
Uploaded Data
.fasta/.qual Associations
Paired Read Associations
Analyses
BugfixVideos
LSC Group Meeting Demo
PROJECTS
Stamped
Welcome, Stephen
System status:
Utilization
89%