



Cloud Sensibility

Hype aside, what can “the cloud” do for life sciences today?

Chris Dagdigan

2010 Bio-IT-World Cloud Workshop





Cloud Sensibility

Hype aside, what can “the cloud” do for life sciences today?

Chris Dagdigan

2010 Bio-IT-World Cloud Workshop



Who I am

- Part of the BioTeam
 - ▶ Bioinformatics → HPC & Research IT nerd
- Our business:
 - ▶ Bridging the gap between science & high performance IT
- **Warning:**
 - ▶ I'm known to speak fast & carry a large slide deck!
 - ▶ Slides will be available online



Why I'm here

- Doing “science” on the cloud since 2007
- Heavy IaaS user
 - ▶ Amazon AWS
- Can speak from multiple viewpoints:
 - ▶ Cloud User/consumer
 - ▶ Vendor/integrator



Defining our terms

- Gartner:

- ▶ *“Cloud computing is a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies.”*

- Jinesh Varia on AWS:

- ▶ *“... Amazon Web Services (AWS) cloud provides a highly reliable and scalable infrastructure for deploying web-scale solutions, **with minimal support and administration costs, and more flexibility than you’ve come to expect from your own infrastructure**, either on-premise or at a datacenter facility.”*

How we got here

- Enterprise applications have been trending away from tightly-coupled desktop or client-server apps
- Moving towards loosely-coupled web services and Service Oriented Architectures (“[SOA](#)”)
- And ...
 - ▶ High-profile Software-as-a-Service (“[SaaS](#)”) success stories in the market are well known
 - ▶ Hypervisor-based virtualization now extremely common in enterprise & clearly providing benefit

Even in the most conservative shops ...

- Virtualization no longer seen as risky or scary
- Hypervisor methods delivering clear benefit
- Google Apps, etc. showing that SaaS is more than hype
- Thus ...
 - ▶ Significant curiosity in efforts that virtualize and commoditize ***infrastructure***
 - ▶ More than just server virtualization
 - ▶ More than just a development platform
 - ▶ “Infrastructure-as-a-Service” (*IaaS*) is born

The Big Picture

Is it real? What does it look like?

Cloud Computing: Real or Hype?

- Yep. It's the real deal
 - ▶ It exists and it is usable/useful **today**
 - ▶ BioTeam started using it for real work in 2007
 - ▶ Needs to be on your radar
- I say this as:
 - ▶ A cynical industry type who has to deliver results efficiently & on a budget
 - ◆ Finely tuned BS detector
 - ◆ Adverse to empty hype & cynical marketing
 - ▶ Someone who sees the same marketers that co-opted and destroyed the term “grid computing” now looking greedily at the “C” word

What does it look like?

- We have acronyms!
- Three main cloud segments:
 - ▶ SaaS
 - ◆ “Software as a Service”
 - ▶ PaaS
 - ◆ “Platform as a Service”
 - ▶ IaaS
 - ◆ “Infrastructure as a Service”

Where is the worst hype?

- “Private Clouds”
 - ▶ Still mostly crap in 2010
 - ▶ 95% marketing, 5% usefulness
- I see two types of “private cloud”
 1. Marketers sticking the c-word onto the same boring VMware/ Xen methods that people have been using for years
 2. Vendors trying to convince you to rebuild your datacenter from scratch

Software As A Service

- Oldest & most mature model
- Software delivered to you over the net
- Typically on a subscription or pay-per-use model
- Think:
 - ▶ Google App Suite
 - ▶ 37Signals.com
 - ▶ FogBugz.com

Platform As A Service

- SaaS on steroids
 - ▶ Typically a hosted platform or environment that lets you manage the full lifecycle of application development, testing, deployment & management
 - ▶ You have much more responsibility & control than in a typical SaaS environment
- Think:
 - ▶ salesforce.com
 - ▶ Microsoft Azure
 - ▶ Google App Engine
 - ▶ CycleComputing

Infrastructure As A Service

- On-demand, outsourced datacenter
- Basic IT “foundational” building blocks for running, scaling, expanding or creating new applications
 - ▶ Pay-per-use model for:
 - ◆ Compute power & virtual servers
 - ◆ Storage & databases
 - ◆ Networks & bandwidth
- Think:
 - ▶ Amazon Web Services

Where the action is

- Platform As A Service (PaaS) is likely to have a large impact in our field
 - ▶ Think about:
 - ◆ Sample tracking & LIMS systems
 - ◆ Protocol/experiment management
 - ◆ Outsourced sequencing etc.
- ... many of these seem natural candidates for a cloud-resident platform service

Where the action is, cont.

- IaaS is hot right now
 - ▶ Delivering real value & new capabilities today
- Why?
 - ▶ Low entry cost
 - ▶ Easy learning curve
 - ▶ Instant feedback
 - ▶ Instant benefits

What to concentrate on

- Scientists & research IT staff should be looking most seriously at **laaS** opportunities
- Why?
 - ▶ Simply put, it's the cloud method that offers the most avenues for scientific, business or financial gain.
 - ◆ Rapid return on investment if done right
 - ◆ Easy learning curve, few complex barriers to entry

Infrastructure as a Service

“Your new scriptable datacenter ...”

Why are we having this discussion?

- In 2010
 - ▶ Chemistry, lab instruments & research protocols are changing faster than the underlying IT infrastructure
 - ▶ The old problem
 - ◆ [2004-today] Can't scale storage & CPU resources fast enough
 - ▶ The new problem
 - ◆ [2010-beyond] Scale-out just one of many problems; existing IT can't react fast enough to changes occurring in the labs
 - ◆ Finally being honest about true costs for operating & maintaining research IT facilities

“Scriptable Infrastructure”

- IaaS can mitigate some these problems:
 - **Scale-out & supporting peak demand**
 - ◆ Massive “internet-scale” applications are what IaaS folks designed their platforms to support
 - **Flexibility & Agility**
 - ◆ Virtualized and programmatically controllable IT building blocks (servers, storage, networks) generally far more flexible than anything you are doing in-house
 - ◆ Must faster to provision/deploy in many cases as well
 - ◆ Pay-as-you-go model reduces sunk costs

What makes IaaS work

- Simple economics of scale.
 - ▶ IaaS providers operate globe-spanning infrastructures of incredible size, scale & scope
 - ▶ Extreme scale allows for levels of efficiency, automation and optimization that none of us can match in-house
 - ▶ Result:
 - ◆ Facility & operational efficiency allows IaaS providers to sell services “cheaply” while still earning a profit

A blunt truth

- Amazon Web Services **owns** the IaaS space
 - ▶ No competitor is even close to achieving parity
 - ▶ Especially given the rate at which AWS rolls out new features, products & service enhancements
- The window for competitors to catch up is closing
 - ▶ Contenders:
 - ◆ Rackspace & Microsoft

AWS Rate of Change Examples

■ Dec 2009

- ▶ **Amazon VPC launch**
- ▶ **AWS Spot Instance launch**
- ▶ Windows Server 2008, SQL Server 2008 support
- ▶ **AWS Import/Export launch**
- ▶ US-West AWS region launch

■ Feb 2010

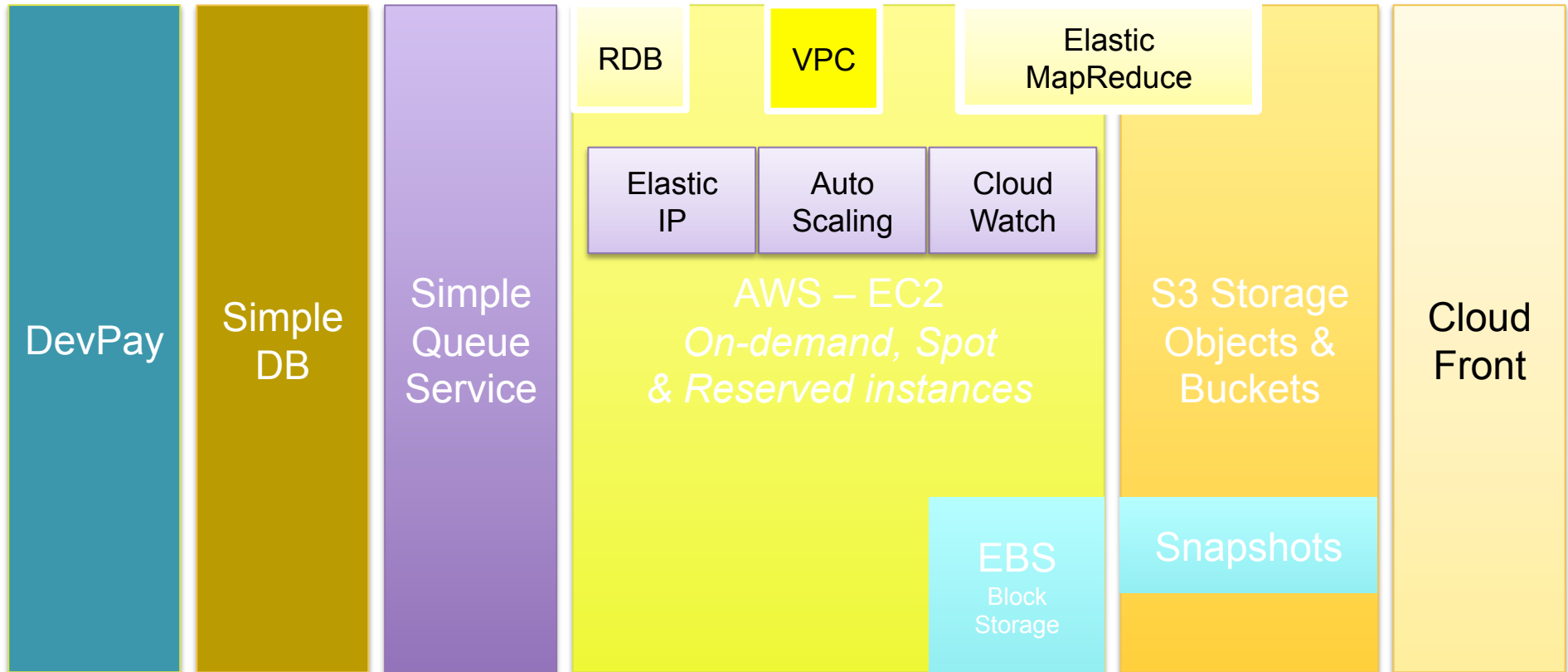
- ▶ SimpleDB consistency enhancements
- ▶ Reserved Instances (Windows)
- ▶ **m2.xlarge EC2 instance type**
- ▶ **AWS Consolidated Billing launch**
- ▶ S3 Object Versioning

AWS Building Blocks

What do we use to build our application?

My Cloud Application, workflow or analysis pipeline

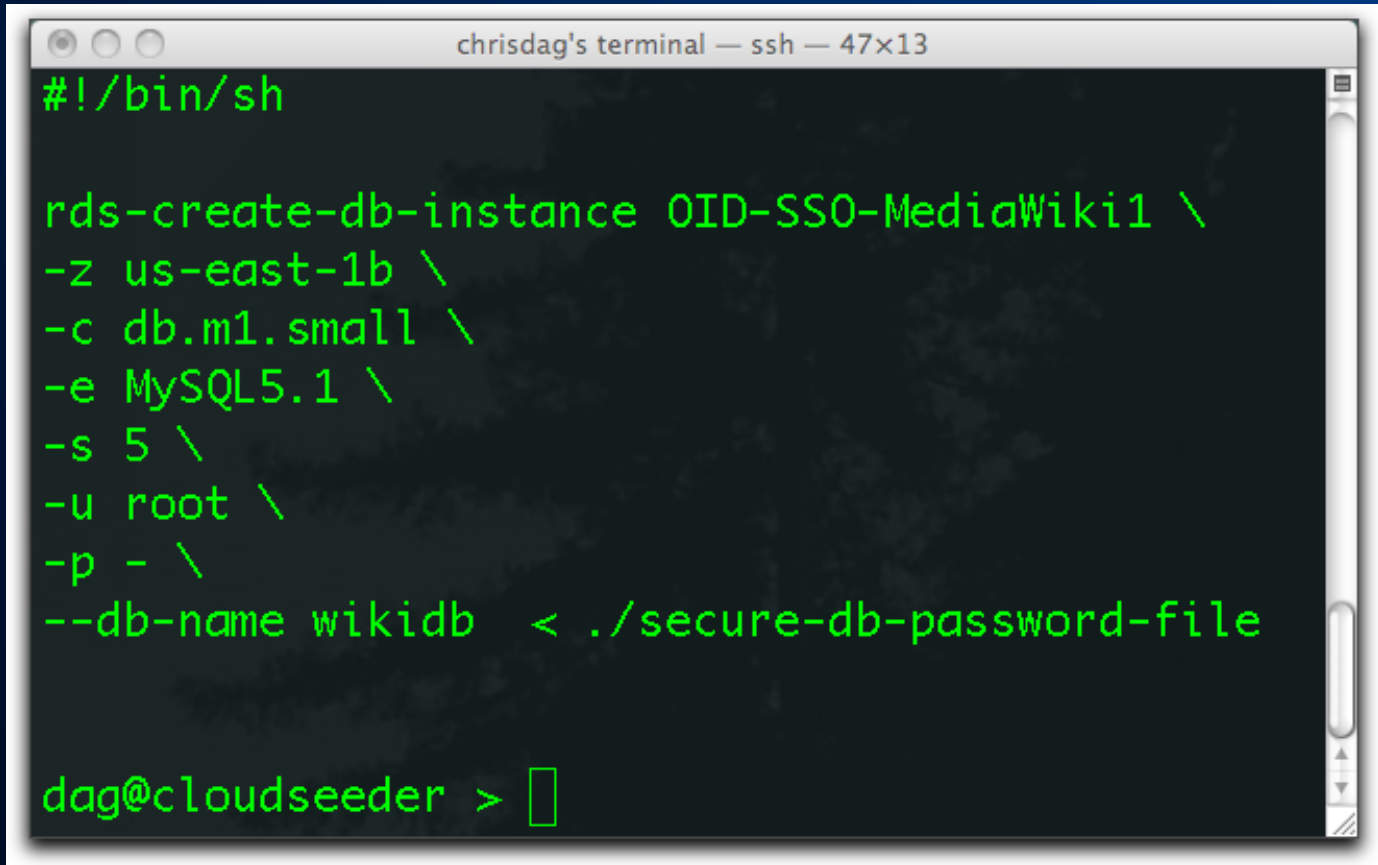
API Endpoints / AWS Web Management Console



Worldwide Physical Infrastructure
AWS Regions, Availability Zones & CloudFront Edge Locations

What's the big deal?

“Scriptable Infrastructure”

A terminal window titled "chrisdag's terminal — ssh — 47x13" with a dark background and green text. The prompt is "#!/bin/sh". The command being entered is "rds-create-db-instance OID-SS0-MediaWiki1 \" followed by several options: "-z us-east-1b \"", "-c db.m1.small \"", "-e MySQL5.1 \"", "-s 5 \"", "-u root \"", "-p - \"", and "--db-name wikidb < ./secure-db-password-file". The prompt at the bottom is "dag@cloudseeder > " followed by a cursor.

```
#!/bin/sh

rds-create-db-instance OID-SS0-MediaWiki1 \
-z us-east-1b \
-c db.m1.small \
-e MySQL5.1 \
-s 5 \
-u root \
-p - \
--db-name wikidb < ./secure-db-password-file

dag@cloudseeder > 
```

This single command will start a 5GB managed MySQL database in the Amazon cloud for \$0.11/hour. The database is **automatically** patched, managed and backed up. Planned enhancements include auto-scaling & snapshots. **THIS IS A BIG DEAL.**

Scriptable Infrastructure

- AWS contains just about everything you would find in your own datacenter(s)
 - ▶ Except:
 - ◆ Your IT now 100% automated & scriptable
 - ◆ Deploy servers, software & services in *minutes*
 - ◆ Scale way bigger than you can handle locally
 - ◆ Your apps & data can now span continents
 - ◆ Services delivered cheaper than local cost*
 - * Requires honest accounting though...
- Anyone can drive this stuff, especially motivated researchers. **This is a big deal.**

What can I do today?

What works today (easy)

- Lowest hanging fruit
 - ▶ Software dev & test environments
- Other
 - ▶ Running legacy Linux apps is easy
 - ▶ Databases of moderate size
 - ▶ Exchanging data with collaborators
 - ▶ Building Grid Engine & LSF clusters and compute farms is pretty trivial
 - ▶ Self-contained workflows & pipelines
 - ▶ Any workflow built using cloud best practices

What works today (moderate)

- Moderate difficulty
 - ▶ “Cloud bursting”
 - ▶ Extending your current cluster “into the cloud”
 - ▶ Data heavy applications
 - ▶ Terabyte scale data mining with Hadoop
 - ▶ Large-scale data movement in general
- The main issue
 - ▶ Networking & VPN overlay networks

What is painful today (hard)

- Massive MPI applications
- Any latency-sensitive parallel application
- Large pipelines or workflows that can't be decomposed easily
- Overwhelmingly IO-bound applications
- When massive two-way data transit is required

Enough cheerleading

Lets talk about the problems & challenges

Three Main Challenges

- For HPC & Informatics in the Cloud:
- Architectural
 - ▶ *Science != Facebook*
- Technical
 - ▶ *Mapping informatics to clouds often non-trivial*
- Political
 - ▶ *You didn't think we could keep politics out?*

Architectural Challenges

Infrastructure clouds were not built for people like us

Architectural Challenges

- Cloud designed for large internet-scale services
- Delivered via:
 - ▶ Loosely coupled, asynchronous services
 - ▶ Significant replication & load balancing tricks
 - ▶ Eventual consistency model
- Not ideal for our needs:
 - ▶ We are used to tightly coupled & fast systems
 - ▶ We happily trade reliability & availability for additional performance & throughput
 - ▶ Scientists see eventual consistency as evil

Architectural Challenges

- Virtual everything is slow
 - ▶ Performance is sacrificed to provide the foundational services required by the extreme internet-scale Web 2.0 crowd
 - ▶ Particularly problematic in life science informatics where we are often performance bound by the speed of our storage systems

Architectural Challenges

- Radical effect on HPC & Grid Computing:
 - ▶ Many of us use large HPC clusters & compute grids within our organization
 - ◆ Large systems *shared* by multiple users, groups, workflows & projects; Platform LSF or Sun Grid Engine software to enable the shared infrastructure resource
 - ▶ Clouds allow *dedicated resources for every user, problem, workflow & project*
 - ◆ Turns traditional methods & practices upside down

Technical Challenges

Data movement & HPC hassles in the cloud ...

Technical Challenges

- This subject is worth a talk of its own ...
- No time to get really deep
- Brief comments on
 - ▶ Data Movement
 - ▶ Networks
 - ▶ Storage
 - ▶ Documentation & How-To pitfalls

Technical Challenges

■ Data Movement

- ▶ #1 issue/concern
- ▶ Internet vs. FedEx?
- ▶ One-way or bidirectional?
- ▶ Not just the size of your pipe ...
 - ◆ Physical location matters as well

■ Networks

- ▶ No control over topology
- ▶ Some nasty surprises for HPC people & software
- ▶ Software VPNs for unifying network space work
 - ◆ ... but it's an insane hassle to set up and manage
 - ◆ Amazon VPC not quite there yet

Technical Challenges

■ Storage

- ▶ It's slow. Absolute fact.
- ▶ Various methods to mitigate or work around
- ▶ Among top 3 implementation challenge in most workflows we've seen

■ “Bad” Documentation

- ▶ Just like the “beowulf cluster” days
- ▶ Most material written for an entirely different audience
 - ◆ Following some ‘best practice’ advice can actually hinder scientific workflows

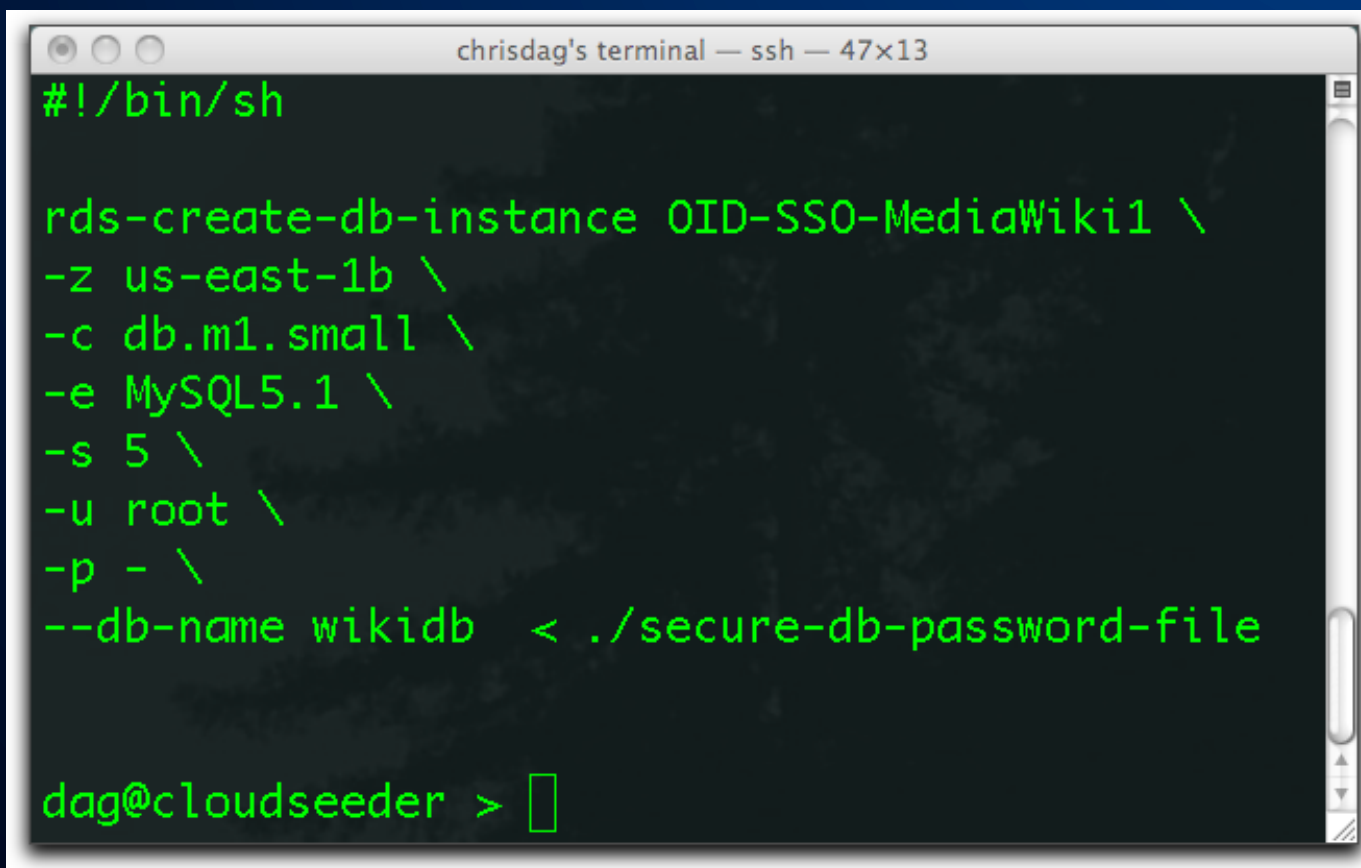
Political Challenges

Now it gets really complicated ...

Political Issues

- Clouds raise significant internal issues
 - ▶ CapEx vs. OpEx issues
 - ▶ Who pays? How do we pay? Who monitors?
 - ▶ When do you port legacy apps to “the new cloud way” ?
 - ▶ What does the support model look like?
 - ▶ What does the development model look like?
- Often encounter these issues:
 - ▶ IT staff protecting internal empires
 - ▶ Incredibly difficult to accurately track *true fully loaded internal costs* of local infrastructure
 - ◆ And if you can't do this, how can you claim the cloud will save money?

Remember This?

A terminal window titled "chrisdag's terminal — ssh — 47x13" with a dark background and green text. The prompt is "#!/bin/sh". The command being entered is "rds-create-db-instance OID-SS0-MediaWiki1 \" followed by several options: "-z us-east-1b \"", "-c db.m1.small \"", "-e MySQL5.1 \"", "-s 5 \"", "-u root \"", "-p - \"", and "--db-name wikidb < ./secure-db-password-file". The prompt at the bottom is "dag@cloudseeder > " followed by a cursor.

```
chrisdag's terminal — ssh — 47x13
#!/bin/sh

rds-create-db-instance OID-SS0-MediaWiki1 \
-z us-east-1b \
-c db.m1.small \
-e MySQL5.1 \
-s 5 \
-u root \
-p - \
--db-name wikidb < ./secure-db-password-file

dag@cloudseeder > 
```

Politics & Scriptable IT

- *What happens to IT roles when anyone with a web browser can instantly launch (and manage) a complex cluster, software pipeline or massive database?*
- Radical restructuring of the lines between
 - ▶ Research staff & Investigators
 - ▶ IT Operations Staff
 - ▶ IT Support Staff

Scriptable Infrastructure

- For the first time some of our IT infrastructure might be 100% virtual and entirely controllable via scripts and APIs
- Anyone can drive this stuff, especially motivated researchers
- My prediction:
 - ▶ The role of “Systems Administrator” is going to change
 - ▶ More focus on toolsmithing, scripting, troubleshooting
 - ▶ Significant focus on enabling end users to be effective and self-supporting (as much as possible)
 - ▶ Interesting times ahead ...

Quick Security Thoughts ...

Quick Security Thoughts ...

1. Microsoft, Google & Amazon have better operating, audit and network security controls than you do.
2. I am suspicious of people demanding cloud security practices that they themselves have failed to deploy on their own infrastructure
3. Cloud providers will happily answer your deepest technical security questions

One new announcement

- BioTeam has formalized it's science-centric Amazon Cloud training materials:
 - ▶ 2-Day hands-on training classes - “*Mastering Amazon Web Services for Science & Engineering*”
 - ▶ Two dates already announced, more coming.
 - ◆ September 2010 – Providence, RI USA
 - ◆ November 2010 – Hannover, Germany
 - ▶ I hate doing sales pitches, shoot me an email or find me afterwards if you are interested.

End;

- Thanks!
- Cloud Training info:
 - ▶ <http://bioteam.net/aws>
- Presentation slides will appear here:
 - ▶ <http://blog.bioteam.net>
- Comments/feedback:
 - ▶ chris@bioteam.net