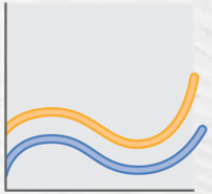
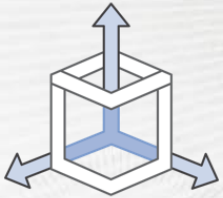
Secure, Cost Effective Compute for Genomics on Amazon Web Services

Vincent Quah
Head, Education/Research/Not For Profit
Asia Pacific and Japan
vquah@amazon.com

Why do researchers love using AWS?



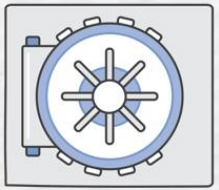
Elastic
Easily add or remove capacity



Scalable
Access to effectively limitless capacity

Time to Science
Access research infrastructure in minutes

Globally Accessible
Easily Collaborate with researchers around the world



Secure
A collection of tools to protect data and privacy

Low Cost
Pay-as-you-go pricing

amazon web services™

Global Footprint

Everyday, AWS adds enough new server capacity to support Amazon.com when it was a $7 billion global enterprise.

Over 1 million **active** customers across 190 countries

2300+ government agencies

7,000+ educational institutions

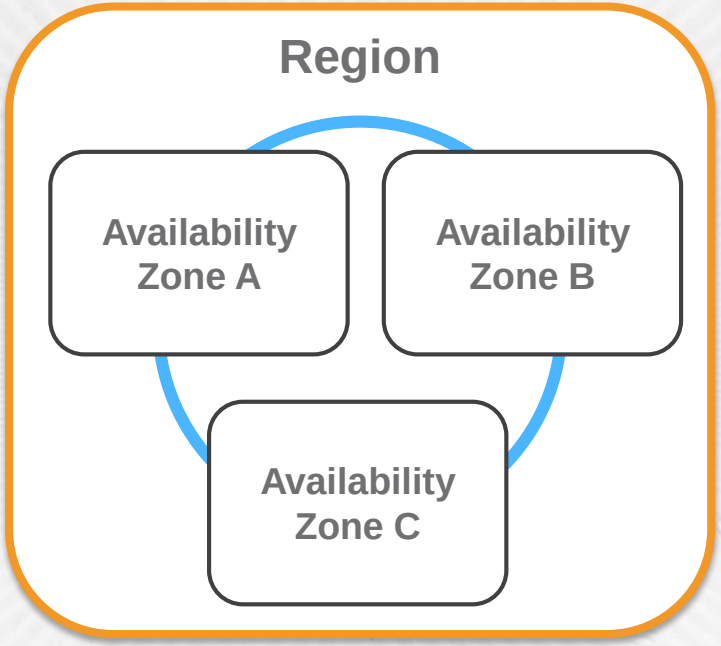
13 regions

35 availability zones

56 edge locations

amazon web services™

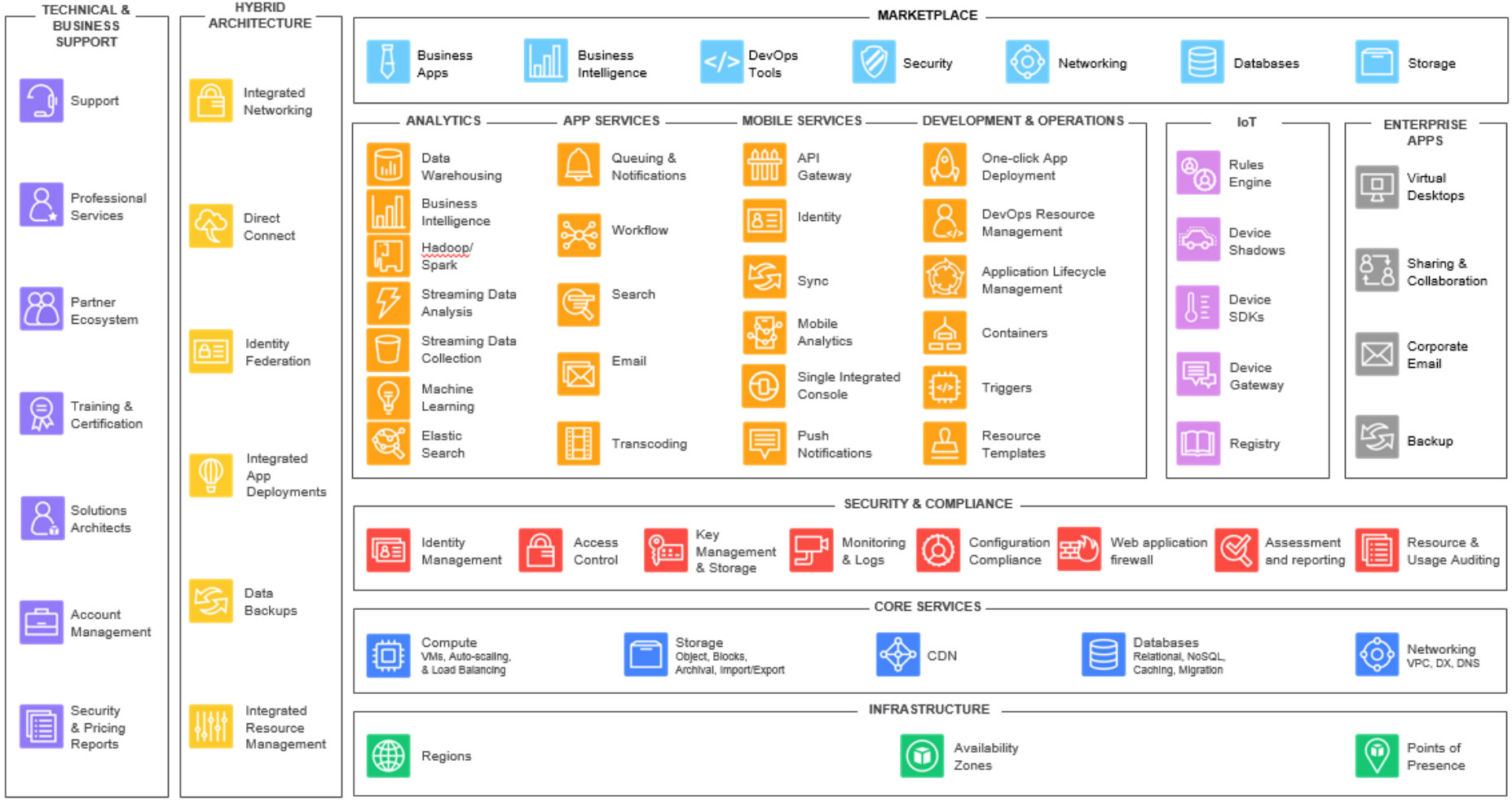
AWS Regions and Availability Zones



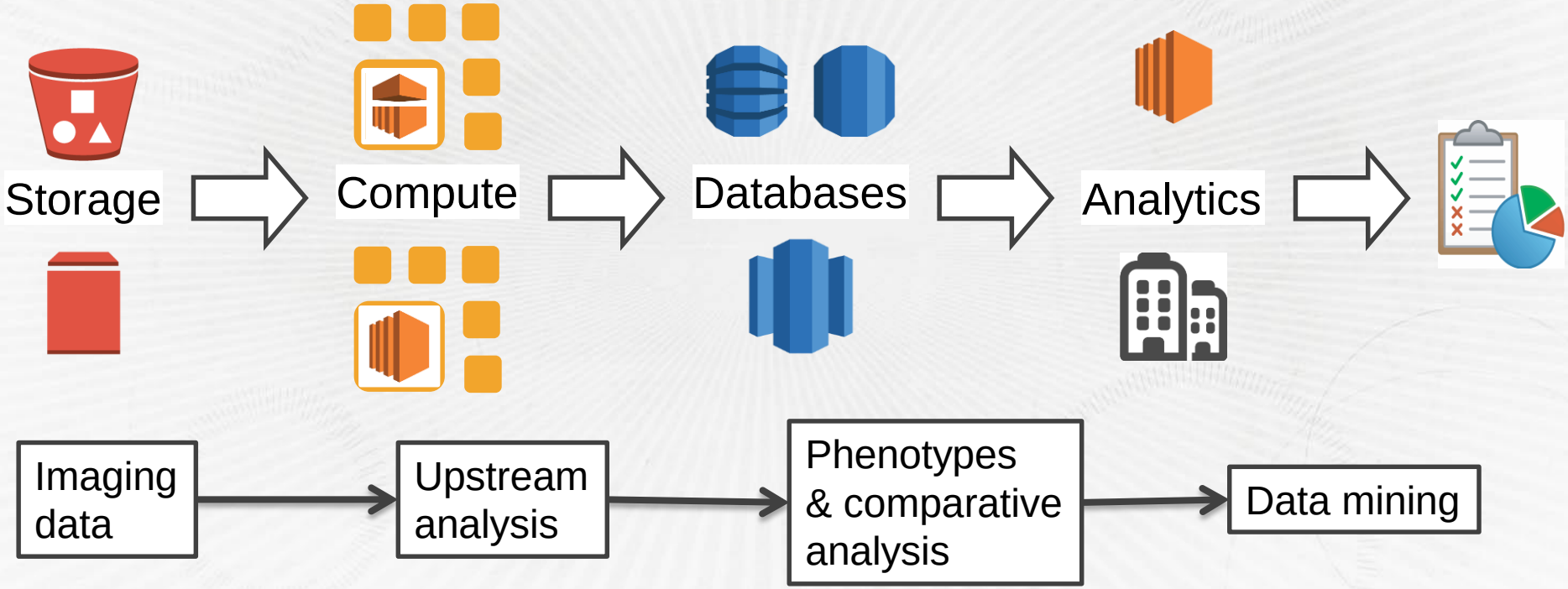
Customer Decides Where Applications and Data Reside

Note: Conceptual drawing only. The number of Availability Zones may vary.

amazon web services™


TECHNICAL & BUSINESS SUPPORT
Support
Professional Services
Partner Ecosystem
Training & Certification
Solutions Architects
Account Management
Security & Pricing Reports
HYBRID ARCHITECTURE
Integrated Networking
Direct Connect
Identity Federation
Integrated App Deployments
Data Backups
Integrated Resource Management
MARKETPLACE
Business Apps
Business Intelligence
DevOps Tools
Security
Networking
Databases
Storage
ANALYTICS
Data Warehousing
Business Intelligence
Hadoop/ Spark
Streaming Data Analysis
Streaming Data Collection
Machine Learning
Elastic Search
APP SERVICES
Queuing & Notifications
Workflow
Search
Email
Transcoding
MOBILE SERVICES
API Gateway
Identity
Sync
Mobile Analytics
Single Integrated Console
Push Notifications
DEVELOPMENT & OPERATIONS
One-click App Deployment
DevOps Resource Management
Application Lifecycle Management
Containers
Triggers
Resource Templates
IoT
Rules Engine
Device Shadows
Device SDKs
Device Gateway
Registry
ENTERPRISE APPS
Virtual Desktops
Sharing & Collaboration
Corporate Email
Backup
SECURITY & COMPLIANCE
Identity Management
Access Control
Key Management & Storage
Monitoring & Logs
Configuration Compliance
Web application firewall
Assessment and reporting
Resource & Usage Auditing
CORE SERVICES
Compute
VMs, Auto-scaling, & Load Balancing
Storage
Object, Blocks, Archival, Import/Export
CDN
Databases
Relational, NoSQL, Caching, Migration
Networking
VPC, DX, DNS
INFRASTRUCTURE
Regions
Availability Zones
Points of Presence

Full stack sequence analysis platform



Amazon Elastic Compute Cloud (EC2)

- Resizable compute capacity in >25 instance types
- Reduces the time required to obtain and boot new server instances to minutes or seconds
- Scale capacity as your computing requirements change
- Pay only for capacity that you actually use
- Choose Linux or Windows
- Deploy across Regions and Availability Zones for reliability
- Support for virtual network interfaces that can be attached to EC2 instances in your VPC

Amazon EC2 Instance Families

General Purpose
(Burstable or Fixed Performance)

Compute Optimized

Memory Optimized

GPU Instances

Storage Optimized

Amazon Simple Storage Service (Amazon S3)

- Durable and secure object storage
  - 99.999999999% durability SLA
  - Encrypt in transit and at rest, fine-grained access permissions
- Low cost
  - 3¢ per Gb per month, pay only for what you use (region dependent)
  - Reduced redundancy storage for lower cost
- Horizontally scalable performance
  - Multipart uploads
  - Access across instances sustained performance

Amazon Elastic Block Store (Amazon EBS)

- Storage volumes for use with Amazon EC2 instances
  - Create, attach, backup, restore and delete raw and formatted device volumes
- Reliable, secure block device storage
  - Volume level AES-256 encryption
- Provisioned amount and performance
  - 1GiB to 16TiB*
  - Magnetic (1TiB), General Purpose and Provisioned IOPS SSD (16TiB)
- Snapshots are durably saved to Amazon S3

DynamoDB

Scalable NoSQL Data Store

Amazon Relational Database Service

Managed Relational Database Service

MySQL, PostgreSQL, MS SQL, Oracle, Amazon Aurora

ElastiCache

In-Memory Cache

Memcached or Redis

Amazon Elastic Map Reduce (EMR)

Managed Hadoop Framework

Apache Hadoop, MapR

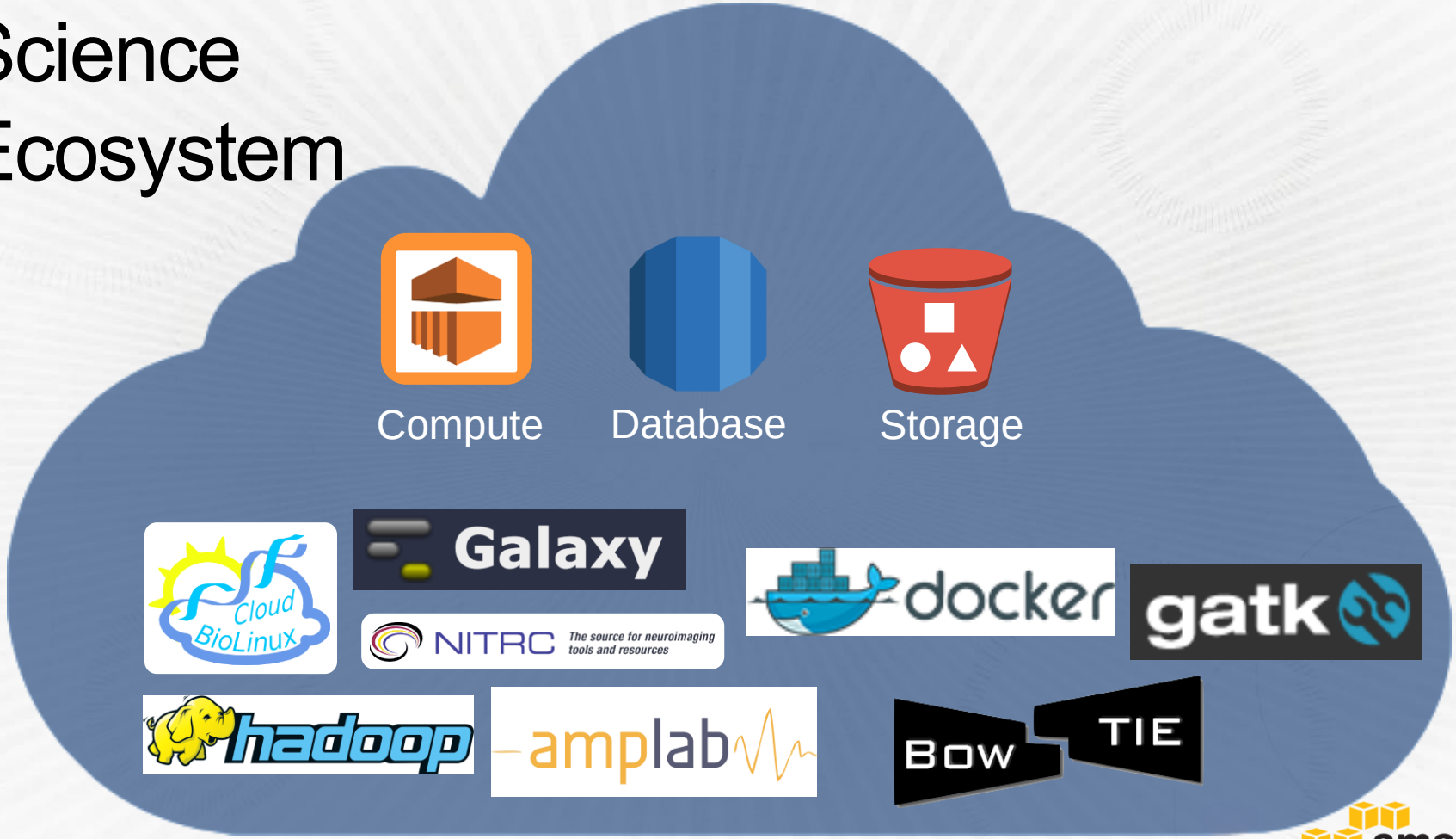
Amazon Redshift

Managed Petabyte-Scale Data Warehouse Service

Columnar data store

PostgreSQL ODBC or JDBC drivers

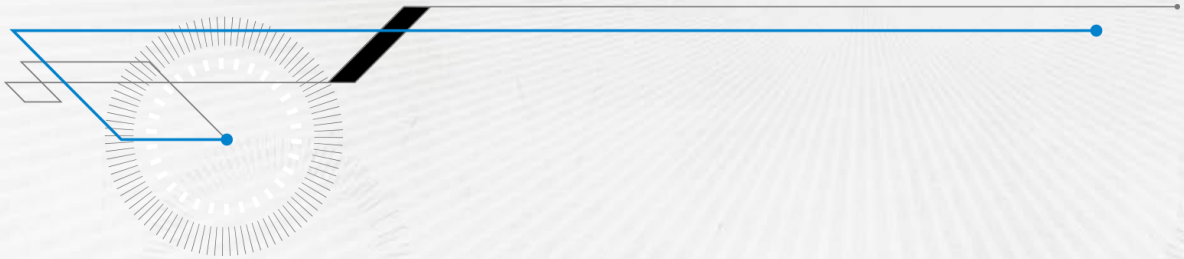
Science Ecosystem



amazon web services™

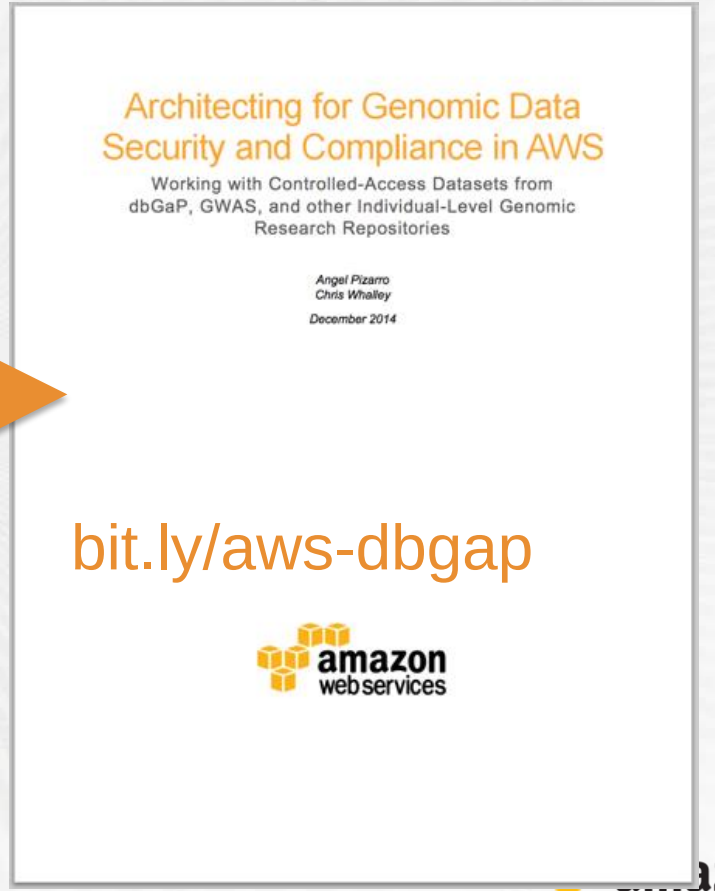
Partner Ecosystem

Genomics Data Security



amazon web services™

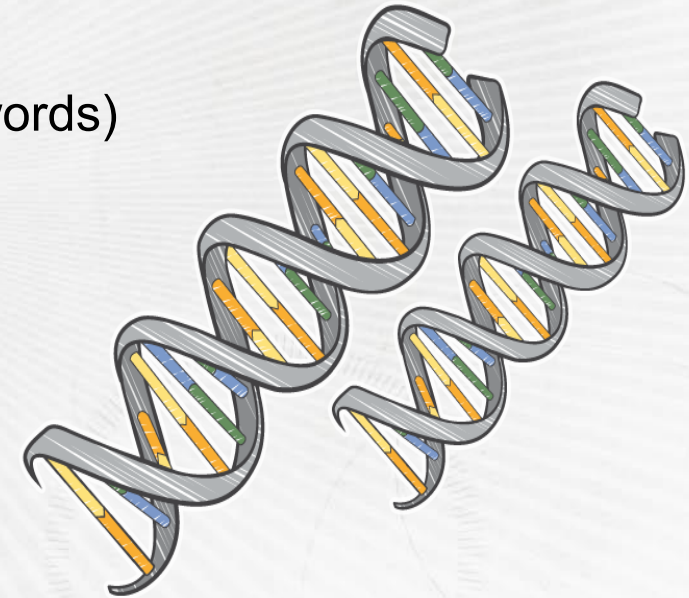
Store and analyze restricted-access genomics on AWS



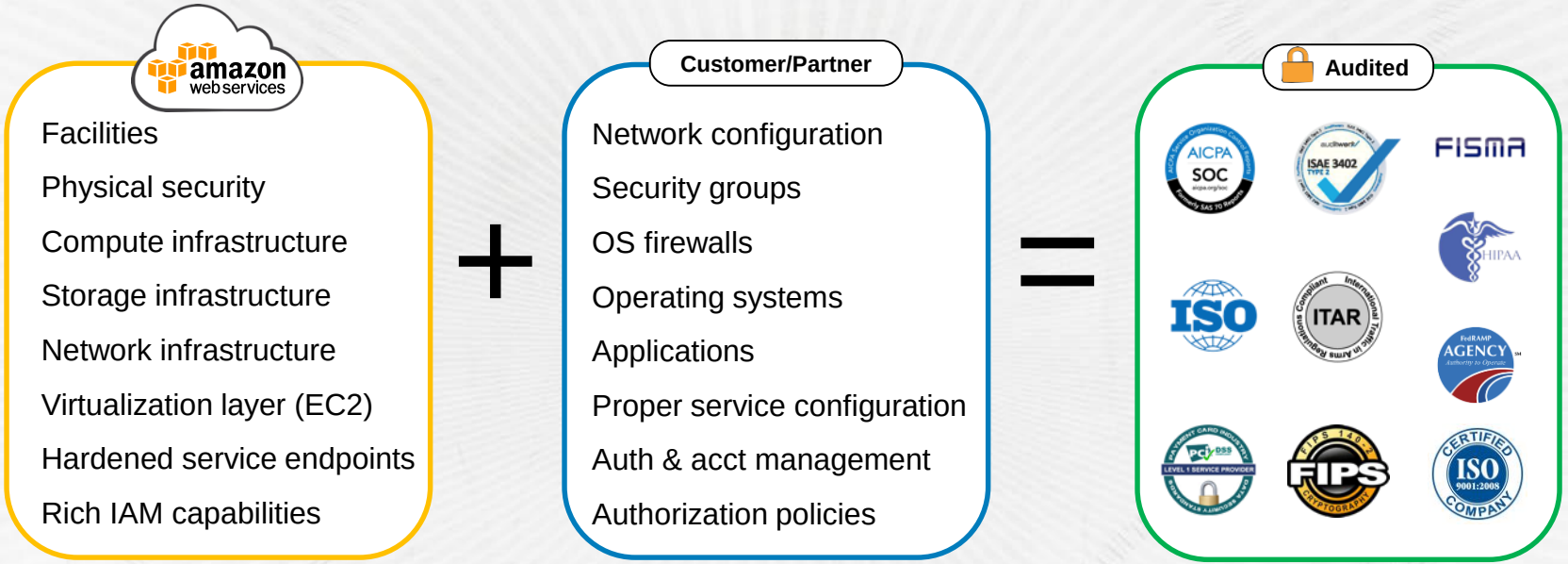
amazon web services™

NIH dbGaP security best practices

- Physical security
  - Data center access and remote administrator access
- Electronic security
  - User account security (for example, passwords)
  - Use of Access Control Lists (ACLs)
  - Secure networking
  - Encryption of data in transit and at rest
  - OS and software patching
- Data access security
  - Authorization of access to data
  - Tracking copies; cleaning up after use



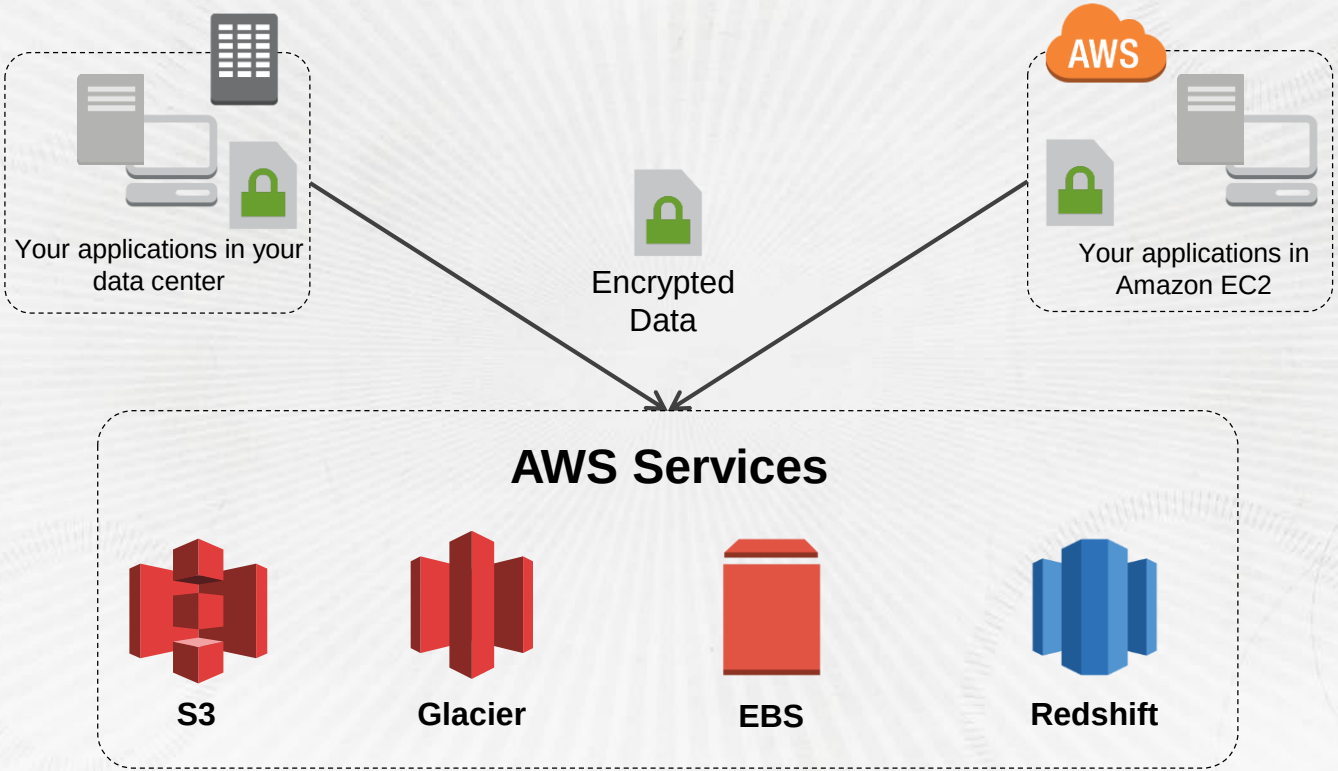
The Shared Responsibility model



- Re-focus your security professionals on a subset of the problem
- Take advantage of high levels of uniformity and automation

amazon web services™

Encrypt your data prior to sending to AWS



Encryption of AWS storage services

Amazon S3

- HTTPS
- AES-256 server-side encryption
- AWS **or** customer provided **or** customer managed keys
- Each object gets its own key

Amazon EBS

- End-to-end secure network traffic
- Whole volume encryption
- AWS **or** customer managed keys
- Encrypted incremental snapshots
- Minimal performance overhead (utilizes Intel AES-NI)

AWS Key Management Service

A service that enables you to provision and use encryption keys to protect your data

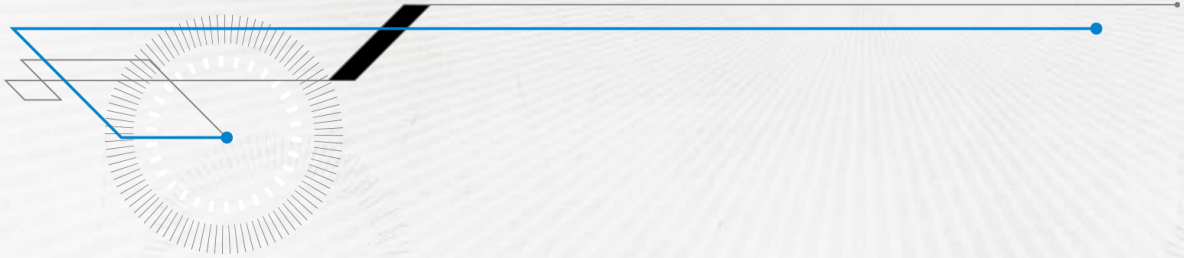
Allows you to create, use, and manage encryption keys from within…

- *Your own applications via AWS SDK*
- *Supported AWS services (S3, EBS, Redshift)*

Available in all commercial regions

Can be used in a key hierarchy to secure data encryption keys protecting PHI

Cost effective computing

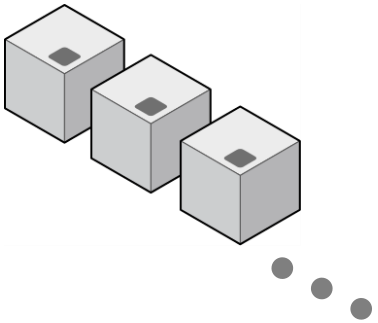


amazon web services™

On-Demand Instances

Pay as you go for computing power

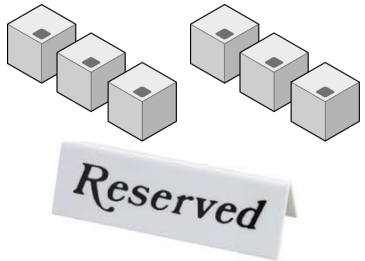
Flat hourly rate, no upfront commitments



Reserved Instances

Pay an upfront fee (or not) to secure **discounted hourly pricing** and a **capacity reservation** for up to 3 years

Buy RIs at a discount or sell underutilized RIs via the **RI Marketplace**

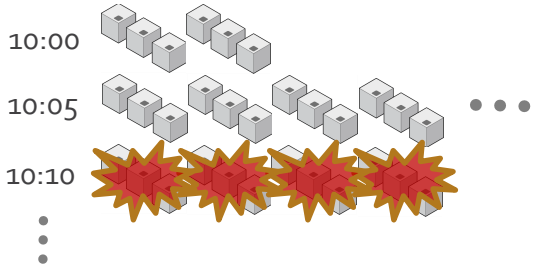


Spot Instances

Bid for *spare* EC2 capacity, access 1,000s of instances at up to 90% off the OD price

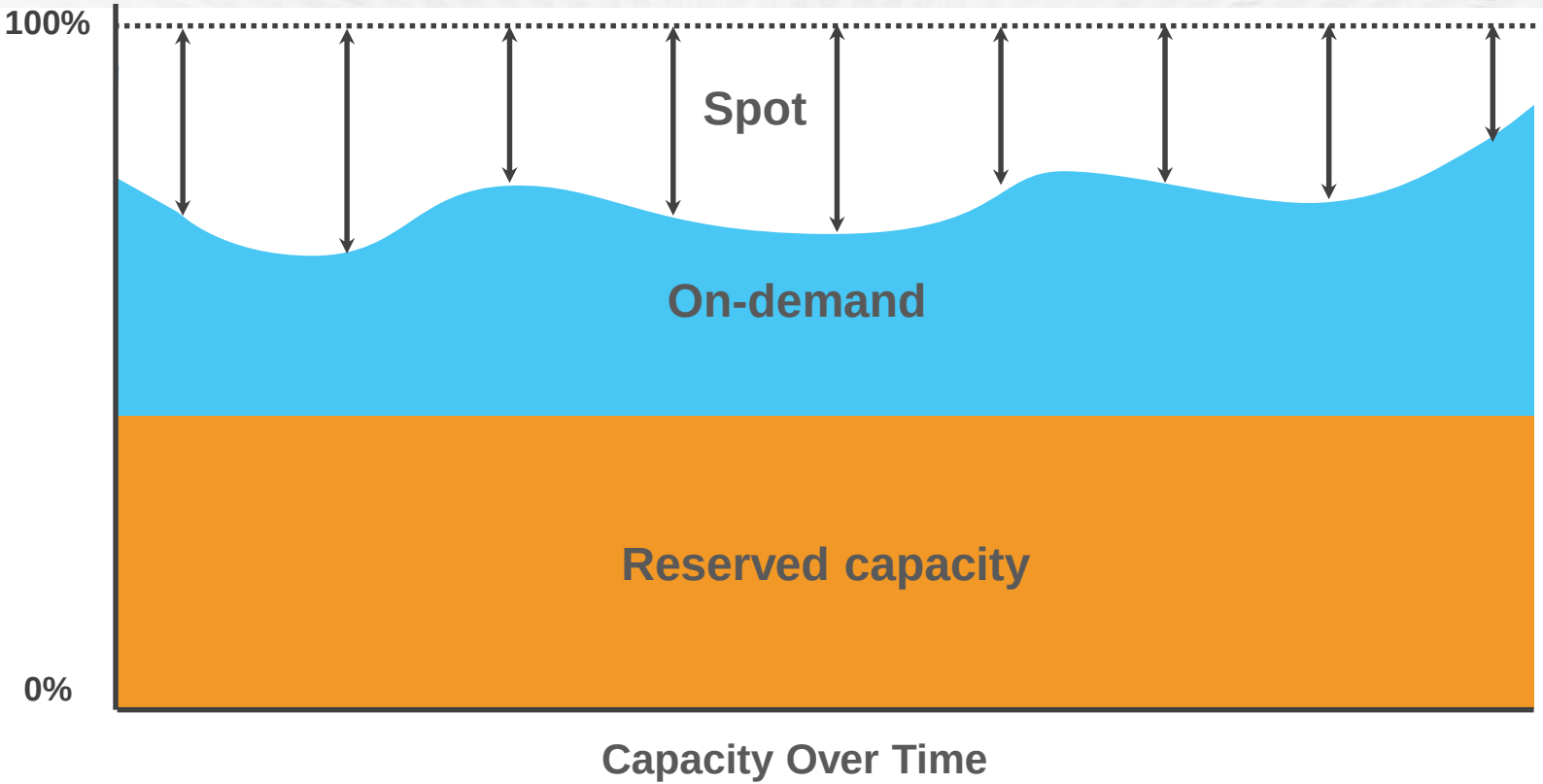
Spot instances run if your bid price > Spot price

If capacity is constrained, your instances may be **evicted**



AWS Spot Market

Achieving economies of scale

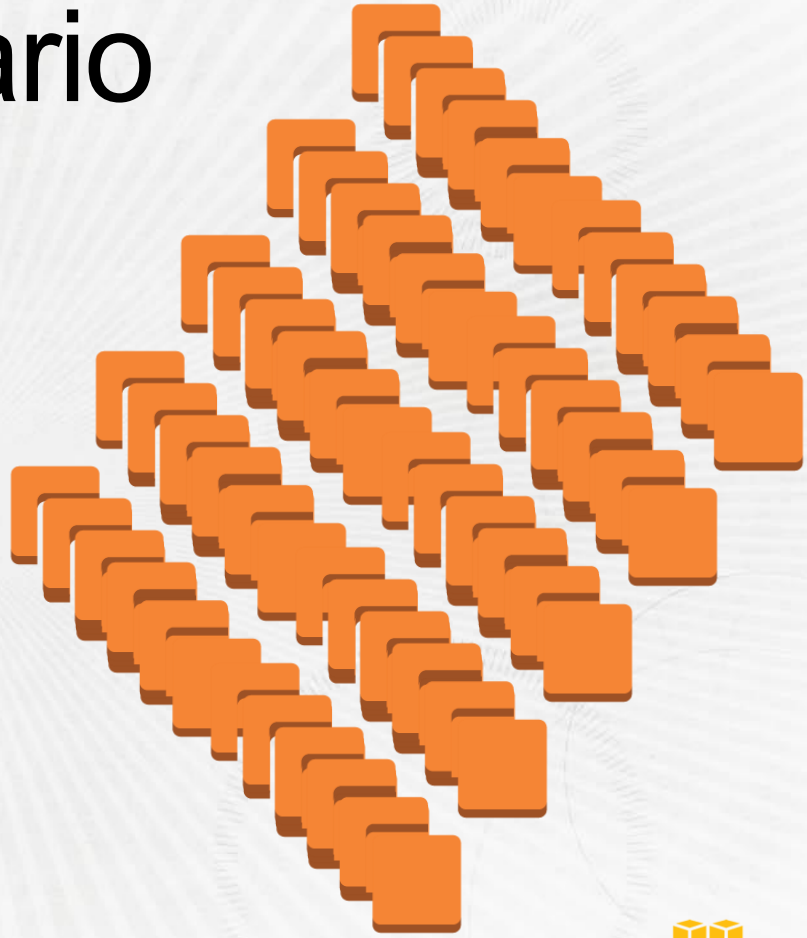


Consider a scenario

You have a need for 10,000 highly powered cores for 48 hours to run your NGS analysis pipeline.

You do some math:

- Your datacenter doesn't have that capacity .. <Math can't help>
- Cloud has the capacity
- Quick back of the napkin calculation – doable for around $50K.. Awesome! But, what if we could do it for less?

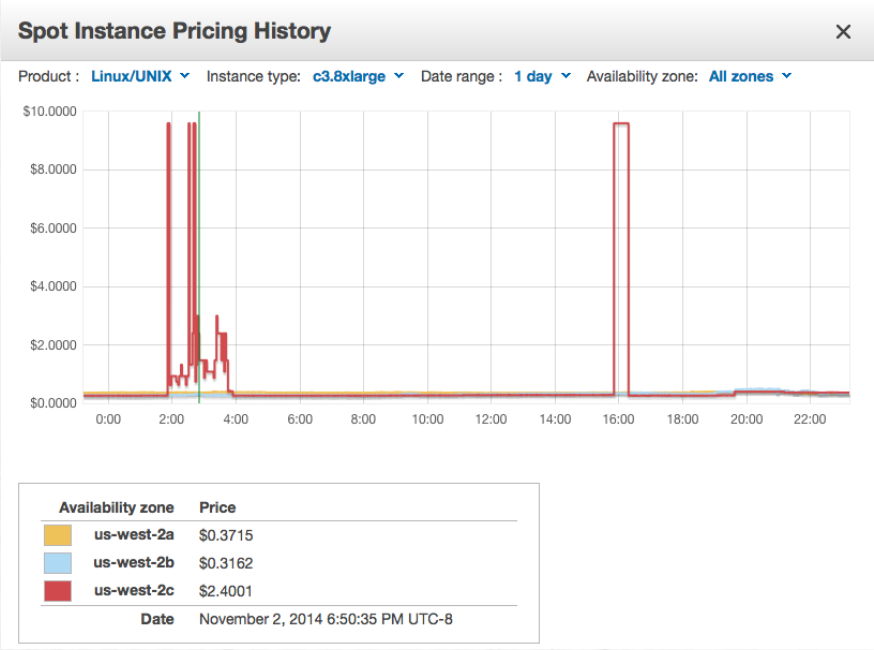


amazon web services™

EC2 Spot Instances…

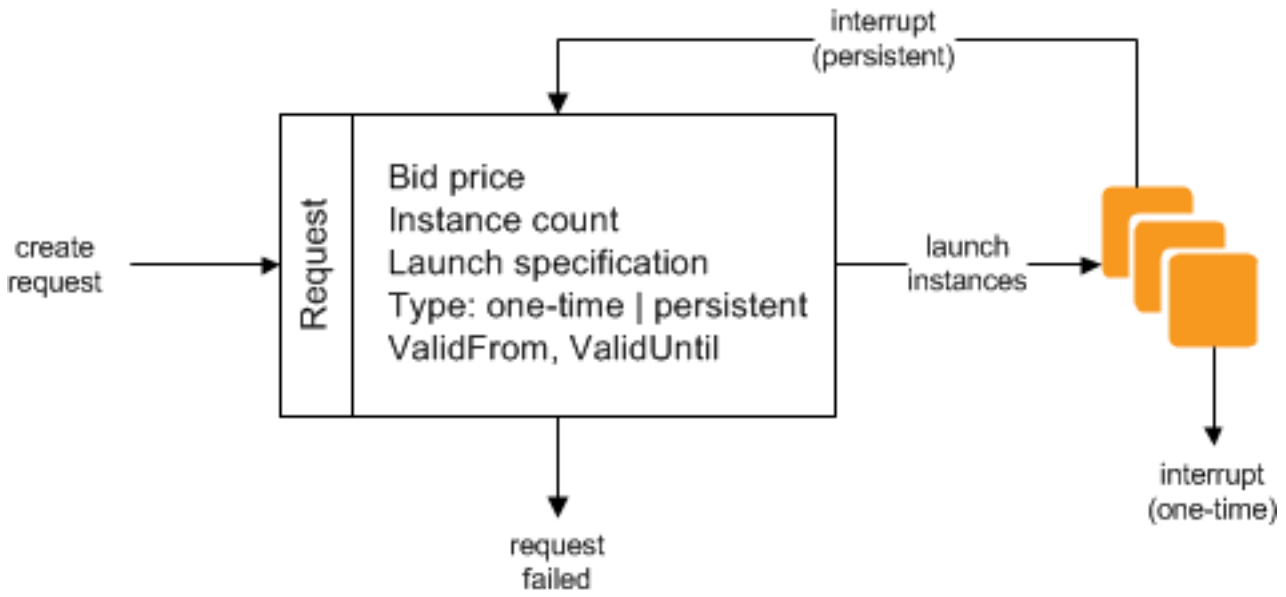
{A market where you can bid your own price for the compute.}

So.. you look at the Spot Market and you estimate that you could do the project for around $8000 **… a saving of over 80%** over an already cost effective choice



amazon web services™

How do Spot instances work?



2 minute warning

Spot friendly workloads

- ***All* Spot applications should be…**
  - **Time-flexible**: You can't expect Spot to always be available, so they can wait until it is, or use OD/RI instances
  - **Fault-tolerant**: the application can gracefully handle Spot interruptions
  - (Because Spot may not be available and Spot Instances may be interrupted)

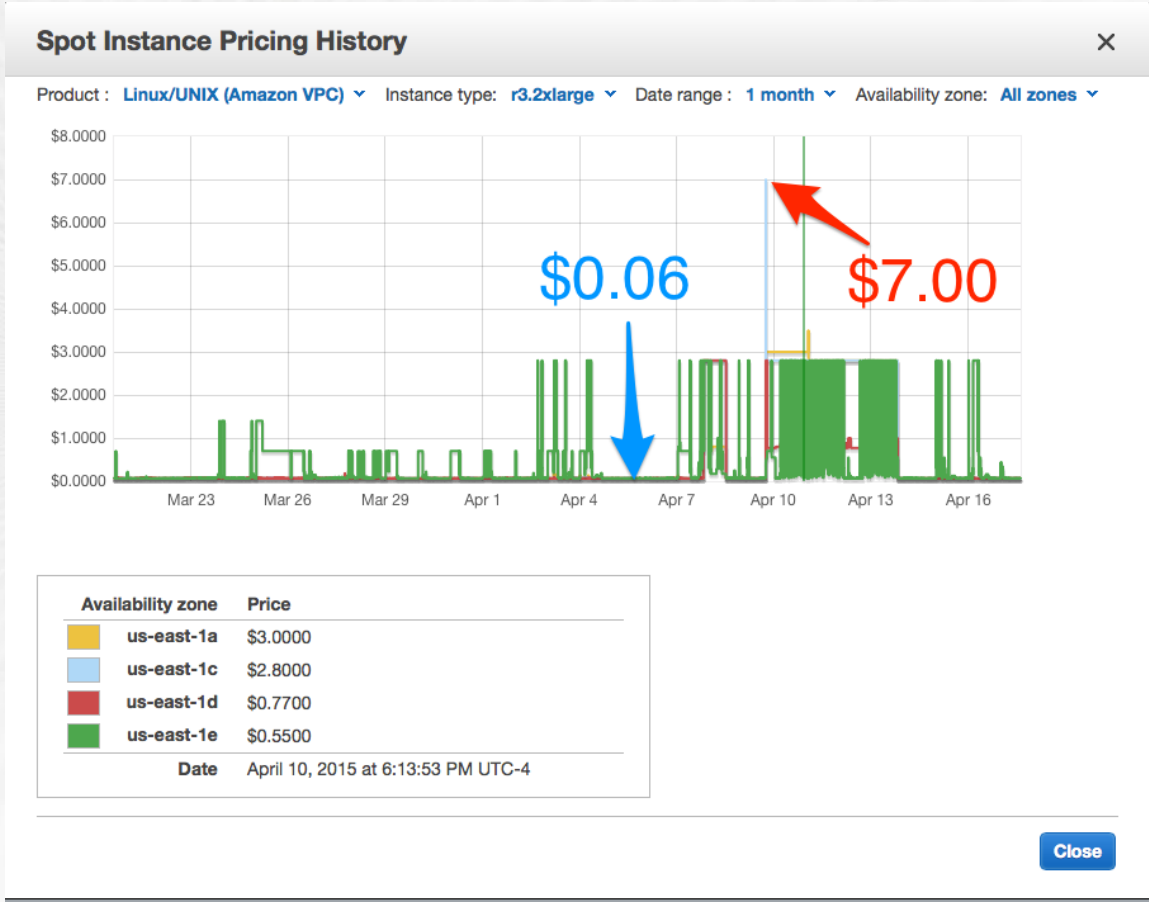
- ***Great* Spot applications are…**
  - **Distributed**: the application has jobs that can be spread out across many Spot Instances, instance types, AZs, even regions
  - **Scalable**: the application can pile on more Spot Instances in parallel to get its job(s) done faster
  - (Access up to 10,000s of Spot Instances worldwide)

Spot friendly architectures

- Fault tolerant
  - Worker resources are ephemeral
- De-centralized workflows
  - Worker resources can run in multiple Availability Zones
- Stateless applications
  - Worker resources grab work items and data from resilient data stores
  - Amazon SQS, Amazon SWF, Amazon S3

Spot price market examples

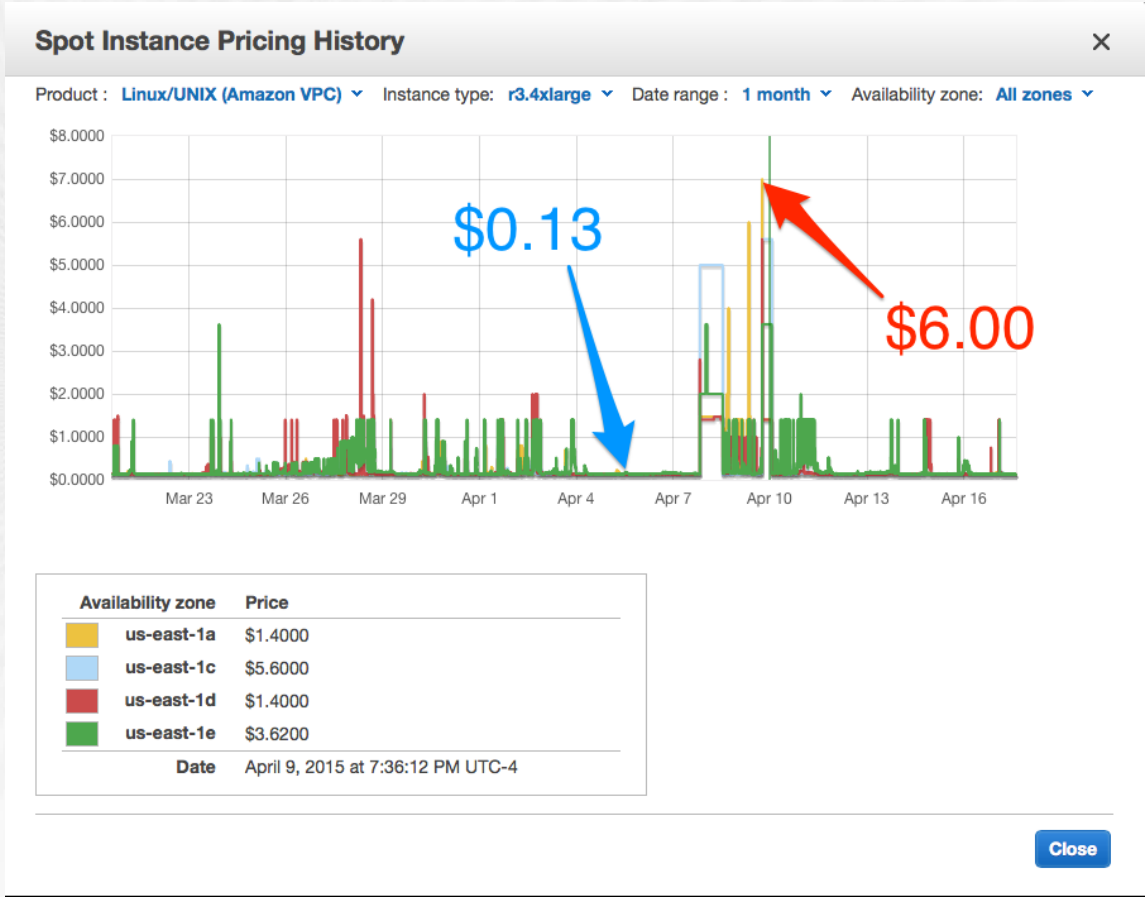
- r3.2xlarge
  - 8 vCPU
  - 61GB RAM
  - 1 x 160GB D
  - $0.70 per hour



* Prices on April 17, 2015

Spot price market examples

- r3.4xlarge
  - 16 vCPU
  - 122GB RAM
  - 1 x 320GB SSD
  - $1.40 per hour



* Prices on April 17, 2015

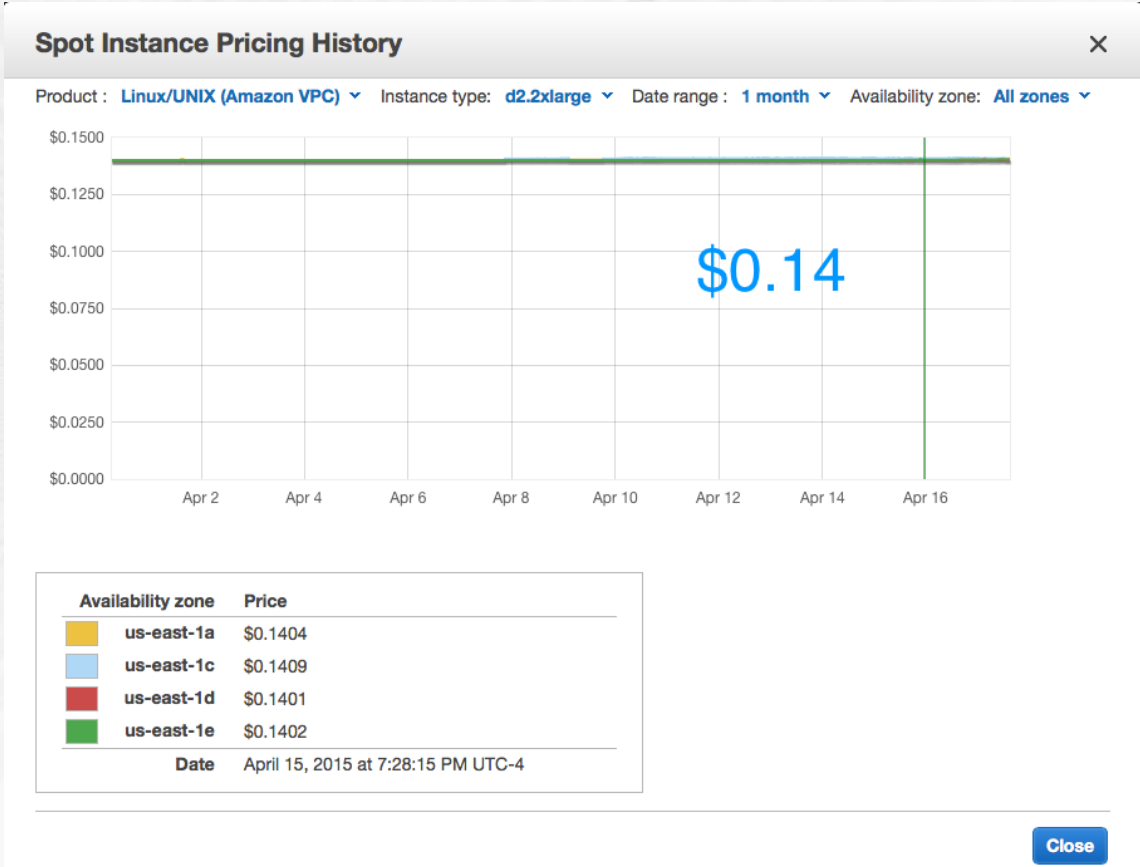
Spot price market examples

- c3.8xlarge
  - 32 vCPU
  - 60GB RAM
  - 2 x 320 SSD
  - $1.68 per hour

* Prices on April 17, 2015

Spot Price Markets – region + instance

- d2.2xlarge
  - 8 vCPU
  - 61GB RAM
  - 6 x 2000GB HDD
  - $0.84 per hour



* Prices on April 17, 2015

Using Spot Effectively – Diversify

- Regions
  - US-East1, US-West2, EU-West2, …
- Availability Zones
  - us-east1a, us-east1b, us-east1c, …
- Instance families & size
  - 1 x r3.8xlarge vs. 4 x r3.2xlarge

Using Spot effectively – normalize application requirements

- CPU Generation
- Memory/core
- Networking
- VPC or Classic EC2

Using Spot effectively – bidding strategies

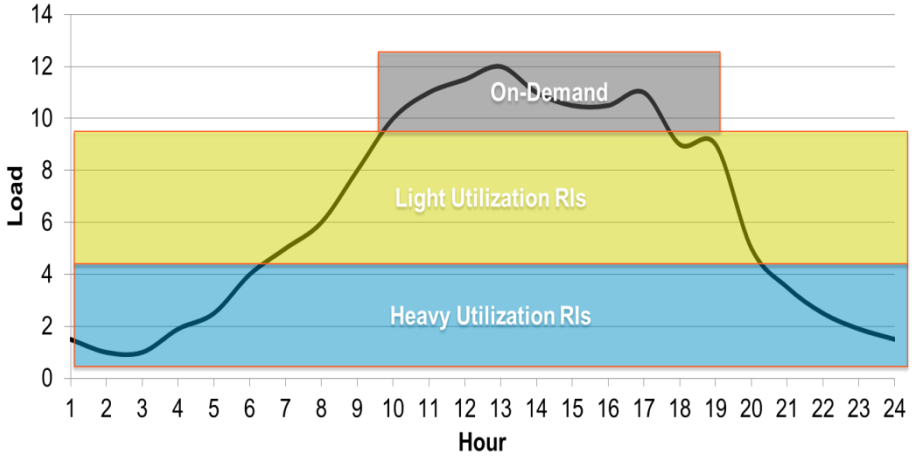
- You only pay what current Market price
- **But,** bid what you are *willing* to pay

Bid only what you are willing to pay.

(by default, bid limited to 4 * On Demand Price)

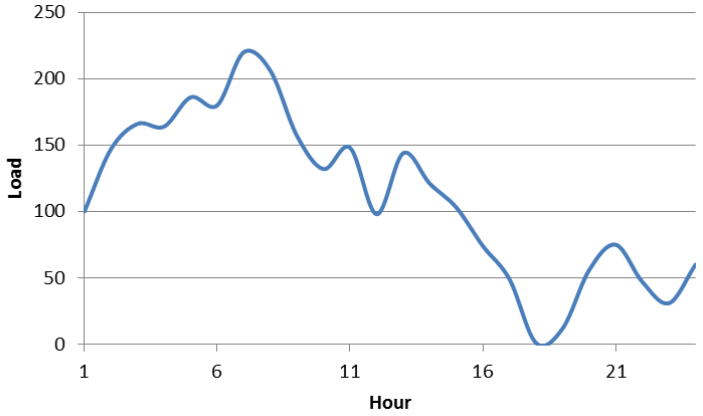
Use On-Demand & Spot

Frontend Applications
on On-Demand/Reserved Instances



+

Backend Applications*
on Spot Instances



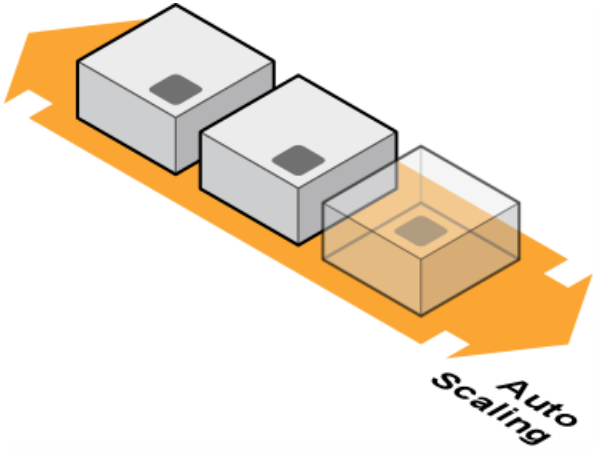
* e.g., image segmentation, genomic alignment

amazon web services™

Amazon EC2 Auto Scaling ♥ Spot

```
aws autoscale create-launch-configuration
   --launch-configuration-name spotlc-5cents
   --image-id ami-e565ba8c
   --instance-type d2.2xlarge
   --spot-price “0.25”

aws autoscale create-auto-scaling-group
   --auto-scaling-group-name spotasg
   --launch-configuration spotlc-5cents
   --availability-zones “us-east-1a,us-east-1b”
   --max-size 16
   --min-size 1
   --desiredcapacity 3
```



http://aws.amazon.com/cli/

amazon web services™

Harvard Medical School

The Laboratory of Personal Medicine

Run EC2 clusters to analyze entire genomes

"The AWS solution is stable, robust, flexible, and low cost. It has everything to recommend it."

Dr. Peter Tonellato, LPM, Center for Biomedical Informatics, Harvard Medical School

Leverage Spot instances in workflows
1 days worth of effort
resulted in
50% savings in cost

http://aws.amazon.com/solutions/case-studies/harvard/

https://cosmos.hms.harvard.edu/

Novartis: Pre-clinical research and drug discovery

> We completed the equivalent of thirty-nine years of computational chemistry in just under 9 hours for a cost of around $4200.
>
> **Steve Litster**
> Global Head of Scientific Computing, Novartis

- Scientists at Novartis had identified a target molecule and needed to screen 10 million compounds against it in a computational model
- Existing infrastructure was not available
- New infrastructure would have cost approximately $40 million to build
- Novartis built a virtual high-performance computing data center in the cloud to run the experiments
- Dramatic increase in the speed of science – able to receive an answer in a fraction of the time at a fraction of the cost

amazon web services™

Bristol-Myers Squibb: Development

> [We could] reduce the number of subjects from 60 to 40 [in a Phase I clinical trial]….the length of the study is reduced by almost 1 year.

Russell Towell
Senior Solutions Specialist, Bristol-Myers Squibb

- PK group at BMS is a group of 40 scientists responsible for clinical simulations
- With two servers it took them 60 hours to run 2000 simulations
- Using a portal built on AWS, the group can now spin up 256 servers simultaneously
- The same amount of work can be done in 1.2 hours for $336

Merck: Manufacturing and Distribution

> We came up with a model that demonstrated, quantifiably, that specific fermentation performance traits are very important to yield.

Jerry Megaro
Director of Manufacturing, Advanced Analytics and Innovation, Merck

- In the summer of 2012, managers at Merck were noticing higher-than-usual discard rates on certain vaccines
- Using a “spreadsheet approach”, on-premises storage and memory limitations meant only 1-2 batches at a time could be analyzed

- A Hadoop distribution running on AWS was used to combine 16 data sources into a data lake
- 1.5 billion batch-to-batch comparisons were performed
- Conclusive answers about production yield variances were produced in 3 months

amazon web services™

Baylor School of Medicine Uses AWS to Accelerate Analysis and Discovery

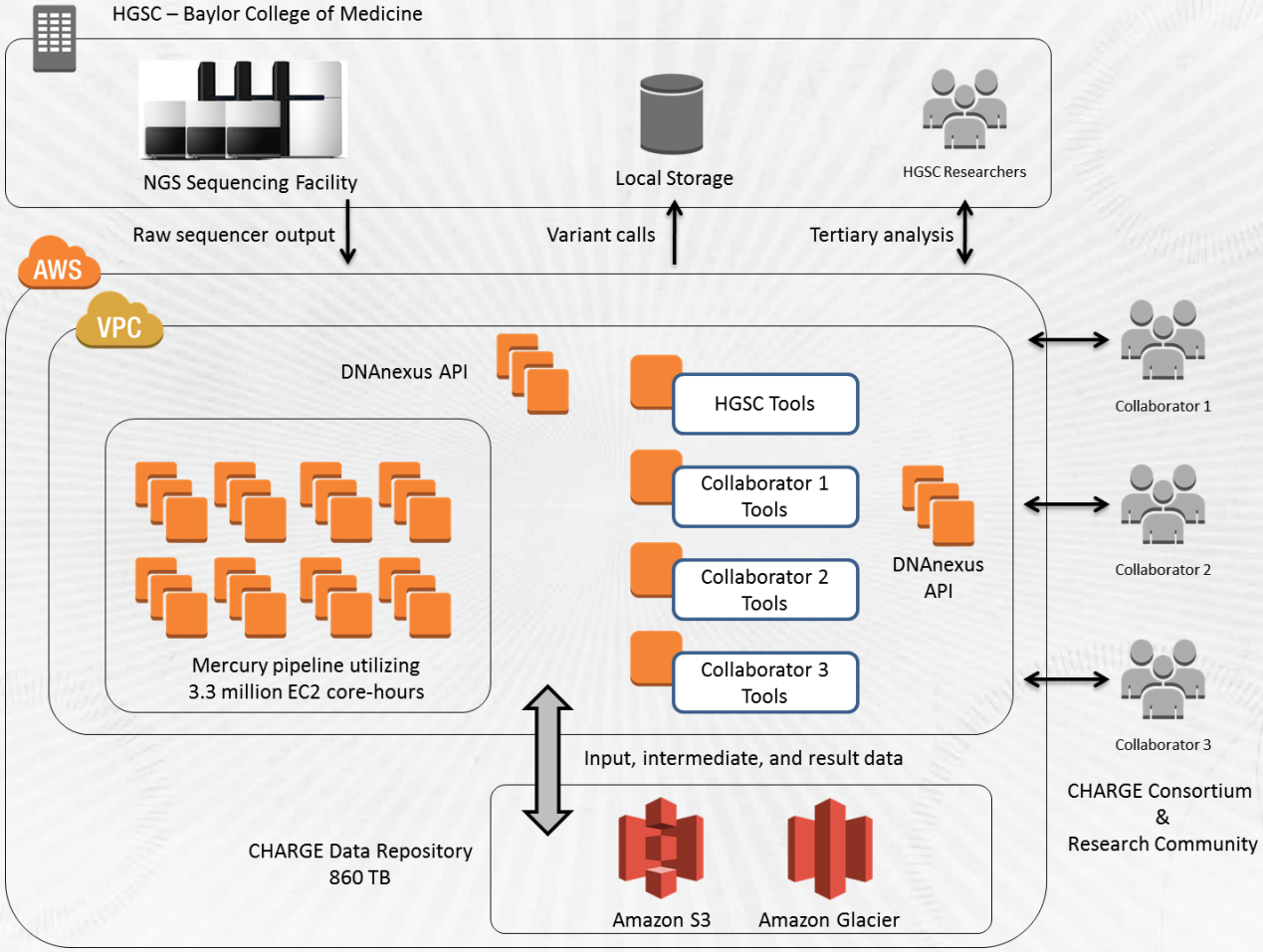
> We are able to power ultra large-scale clinical studies that require computational infrastructure in a secure and compliant environment at a scale not previously possible.
>
> **– Omar Serang**
> DNAnexus Chief Cloud Officer, DNAnexus

BCM
Baylor College of Medicine

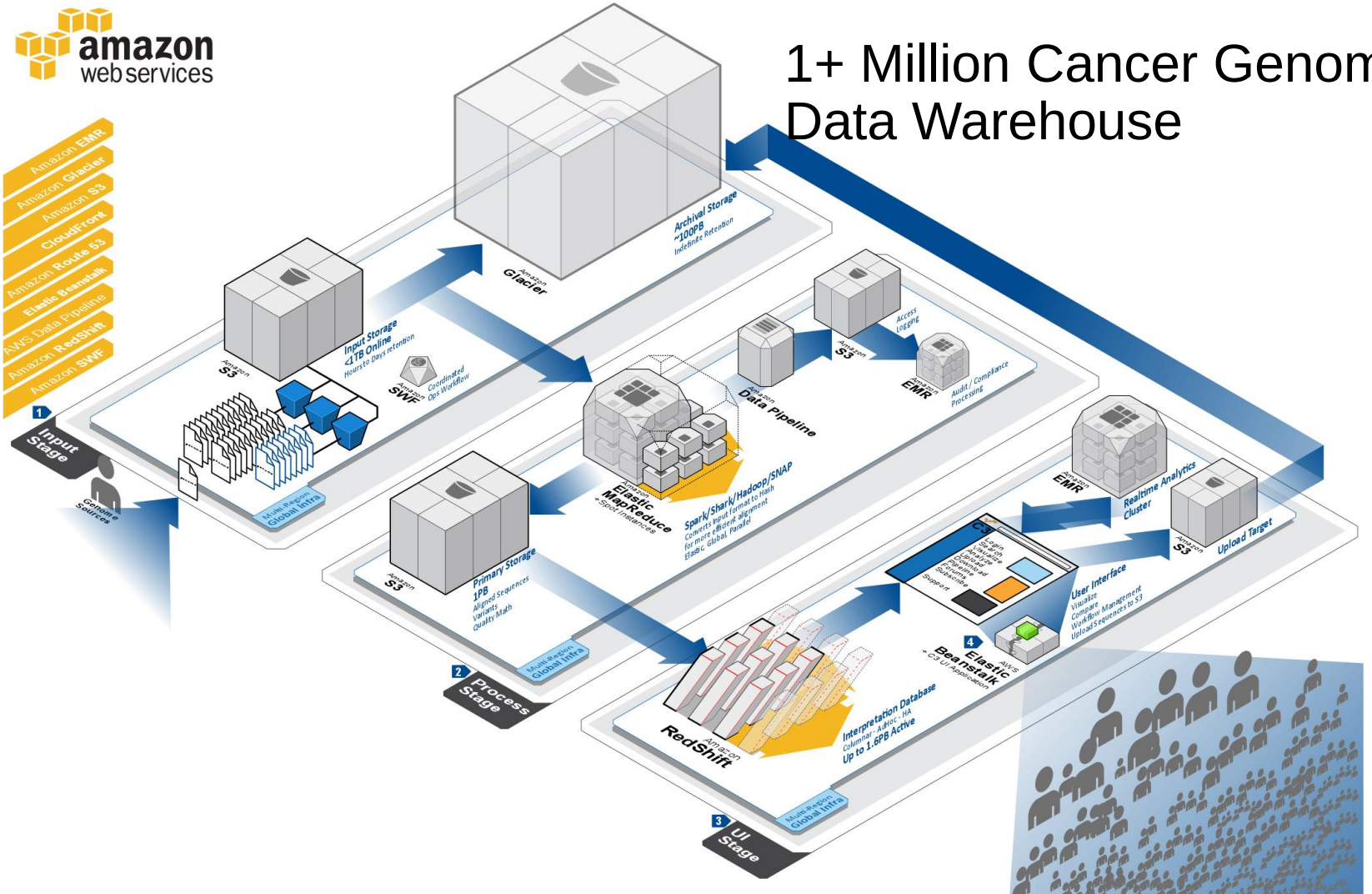
The Baylor College of Medicine is a leading contributor to the CHARGE Project, a group of 200+ scientists who are working to identify genes that contribute to aging and heart disease.

- Offers customers video surveillance and free video storage for 7 days in a private, secure cloud
- Baylor's collaboration with the CHARGE investigators required a secure, scalable genomic analysis platform. They partnered with DNAnexus to use their PaaS for genomic analysis, built on AWS.
- Stores more than 430 TB of genomic result data
- Analyzes the genome sequences of more than 14,000 individuals—5 times faster than with the previous infrastructure
- Enables more than 200 scientists worldwide to share tools and data quickly

DNAnexus


HGSC – Baylor College of Medicine
NGS Sequencing Facility
Local Storage
HGSC Researchers
Raw sequencer output
Variant calls
Tertiary analysis
AWS
VPC
DNAnexus API
HGSC Tools
Collaborator 1 Tools
Collaborator 2 Tools
Collaborator 3 Tools
DNAnexus API
Mercury pipeline utilizing
3.3 million EC2 core-hours
Input, intermediate, and result data
CHARGE Data Repository
860 TB
Amazon S3
Amazon Glacier
Collaborator 1
Collaborator 2
Collaborator 3
CHARGE Consortium
&
Research Community

1+ Million Cancer Genome Data Warehouse



Thank you!

Architecting for Genomic Data Security and Compliance in AWS

http://bit.ly/aws-dbgap