Scalability in the Cloud

HPC Convergence with Big Data in Design, Engineering, Manufacturing

July 7, 2014

David Pellerin, Business Development Principal
Amazon Web Services

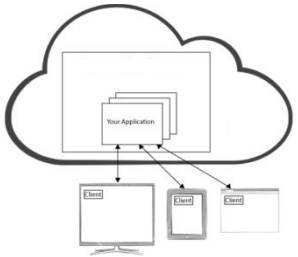
What Do We Hear From Customers?

Want to run larger HPC workloads, more often

- Need to scale up and scale out for faster results
- Want to run more jobs for higher quality of results

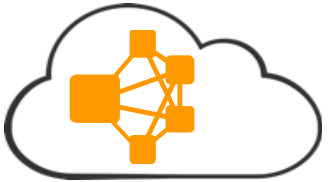
Need to globally collaborate on critical data

- To enhance the productivity of distributed teams
- To operate globally, with greater data governance



Need to manage large, diverse datasets

- Ingest and analyze, query, and take action
- Including connected devices, Internet-of-Things



A “Cloud First” Strategy at HGST

HGST
a Western Digital company

Zero to Cloud in 6+ Month

HGST
a Western Digital company

By 31 Oct 2013:

- ✓Cloud eMail – Microsoft Office365
- ✓Cloud eMail archiving/eDiscovery
- ✓External SingleSignOn (off VPN)
- ✓Cloud File/Collaboration – BOX
- ✓Cloud CRM – Salesforce.com
  - ✓ Integrated to save files in BOX
- ✓Cloud–High Performance Computing (HPC) on Amazon AWS
- ✓Cloud – Big Data Platform on Amazon AWS

Office 365 · box · salesforce · amazon web services

“HGST is using the cloud for a higher performance, lower cost, faster deployed solution vs buying a huge on-site cluster.”

- Steve Philpott, CIO

HGST HPC and big data roadmap:

- ✓ Molecular dynamics simulations
- ✓ Collaboration tools for engineering
- ✓ Big data for manufacturing yield analysis
- ✓ CAD, CFD, EDA – “cloud first”

Every application presents unique challenges… some technical, some business

HGST’s Amazon HPC Platform

Base HPC Platform

- Scalable to thousands of instances to support numerous simultaneous simulations

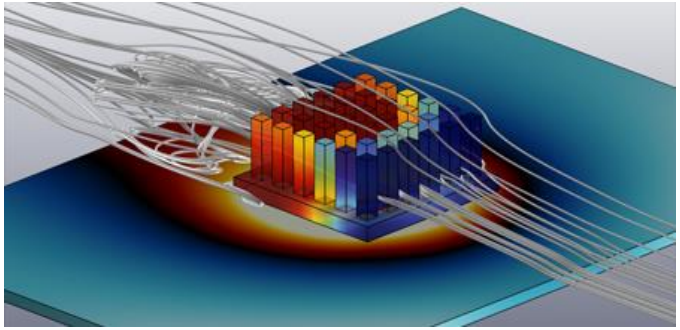
Cloud Scalability for Simulations

Computer-Aided Design, Simulation, Analysis, Visualization

- Across industries, the trend is Simulation-Driven Design and Discovery

Examples in Design and Manufacturing

- Computer-Aided Design
- Electronic Design Automation
- Computational Fluid Dynamics
- Molecular Modeling
- Seismic Modeling and Reservoir Simulations



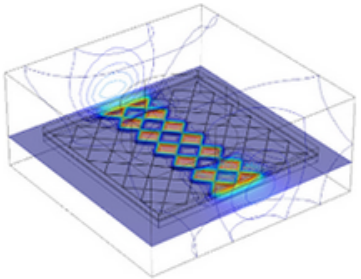
Cloud enables scalable simulations for global manufacturers

Cloud also enables HPC convergence with big data

Simulations for Electronic Product Design

A. Gourevitch
Cypress Semiconductor Corp., San Jose, CA, USA

We report an implementation of parallel computing on Amazon Web Services™ (AWS) for touch-sensor modeling. COMSOL Multiphysics® was used to simulate an electromagnetic field distribution in a capacitive sensor assembly. Multiple COMSOL jobs were deployed on separate AWS instances and were executed in parallel. The simulation results indicate that implementation of parallel computing for COMSOL simulations can significantly reduce the computational time required for optimization of capacitive touch sensor patterns.

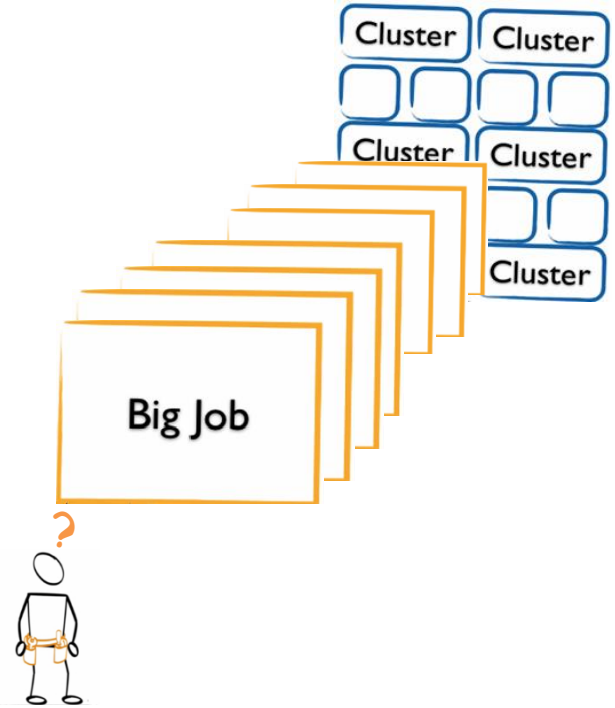


Files Available for Download
Abstract
Paper

No job queue… *Using cloud, simulation sweeps are possible in hours instead of weeks*

Courtesy of Cypress Semiconductor

Job Queues Are Evil!



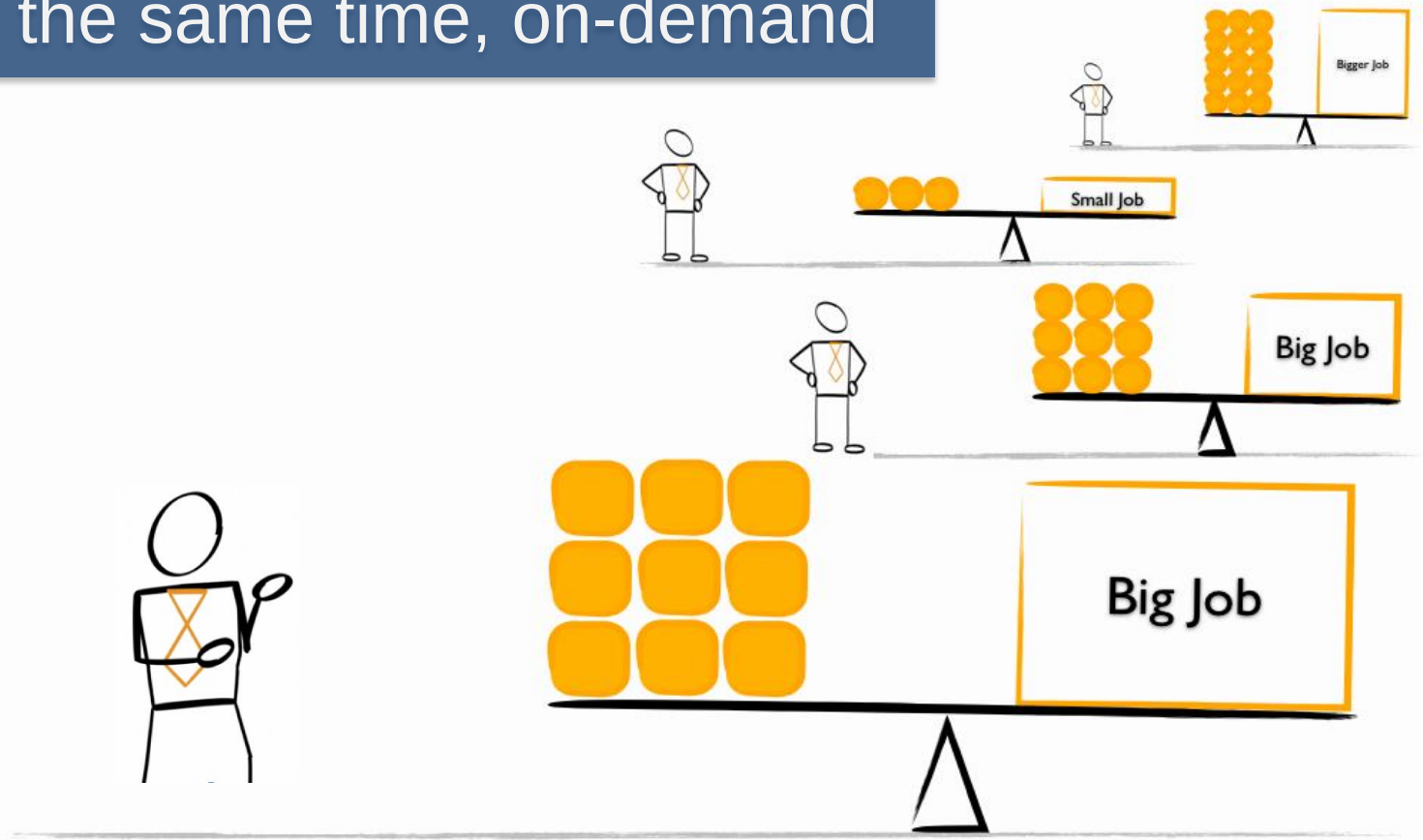
Conflicting goals

- HPC users seek fastest possible time-to-results
- IT support team seeks highest possible utilization

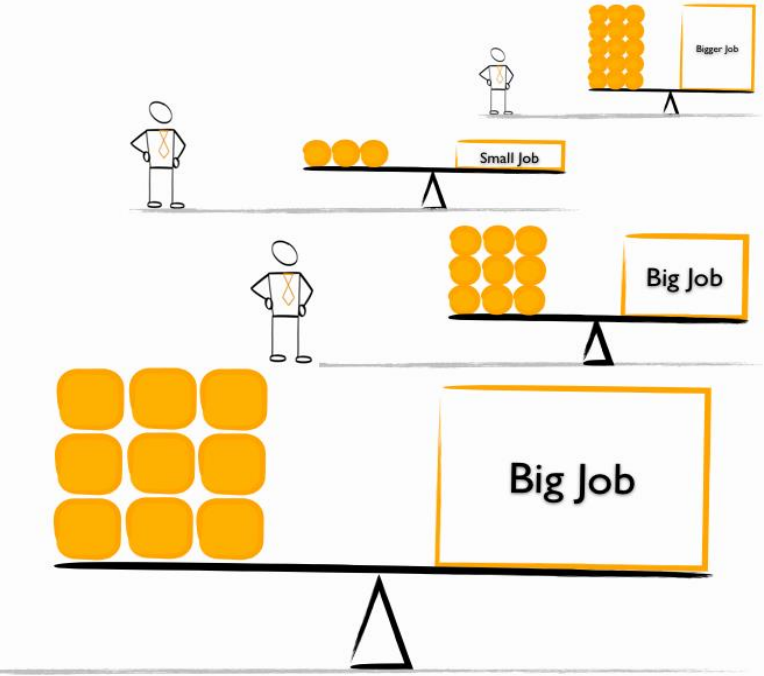
Result:

- The job queue becomes the capacity buffer
- Users are frustrated and run fewer simulations, or they try to game the system
- Money is being saved, but for the wrong reasons!

Instead, run multiple clusters at the same time, on-demand



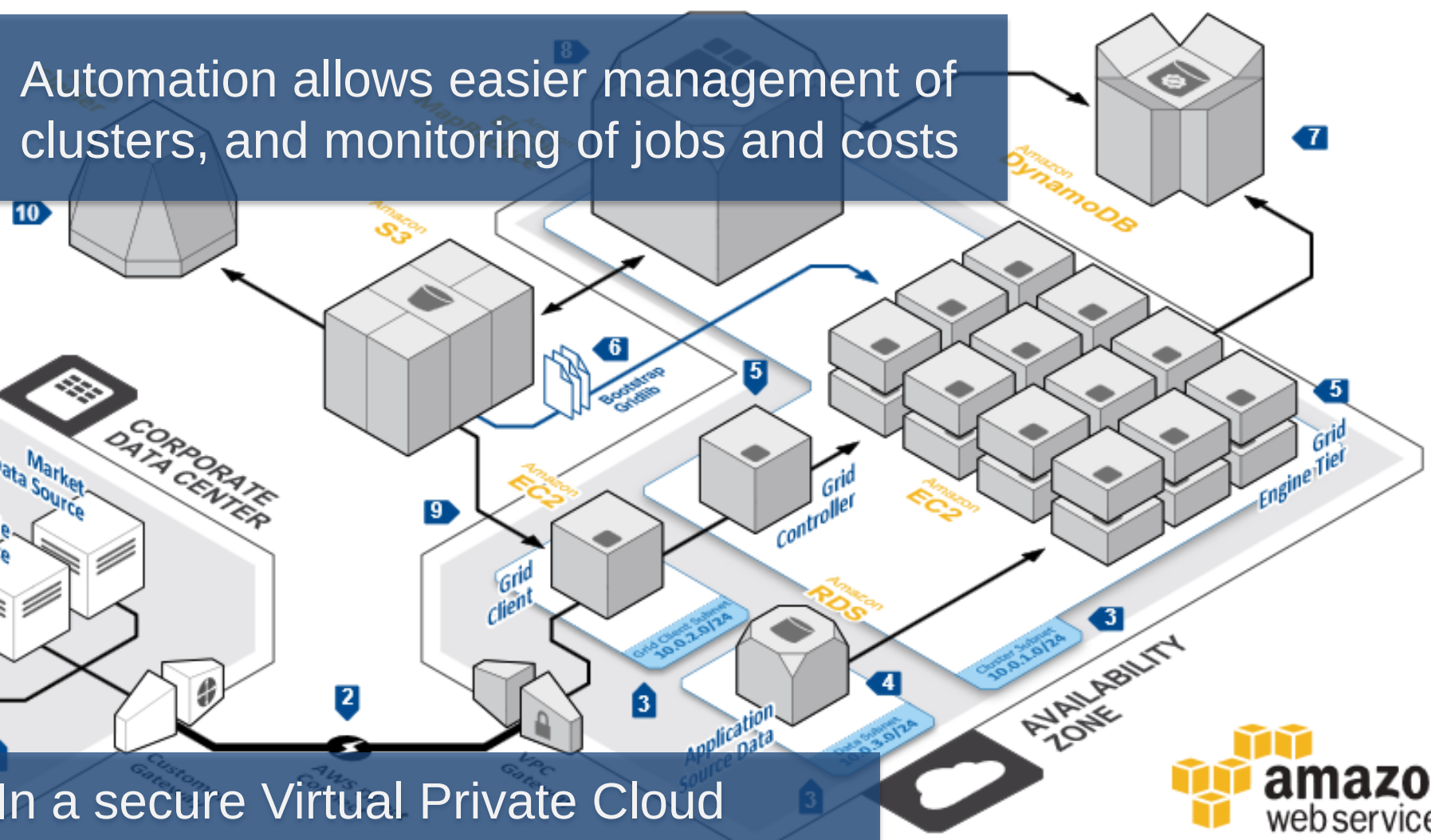
Match the Architectures to the Jobs



Instance Family	Instance Type	Processor Arch	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	EBS-optimized Available	Network Performance
Compute optimized	c3.large	64-bit	2	7	3.75	2 x 16 SSD	-	Moderate
Compute optimized	c3.xlarge	64-bit	4	14	7.5	2 x 40 SSD	Yes	Moderate
Compute optimized	c3.2xlarge	64-bit	8	28	15	2 x 80 SSD	Yes	High
Compute optimized	c3.4xlarge	64-bit	16	55	30	2 x 160 SSD	Yes	High
Compute optimized	c3.8xlarge	64-bit	32	108	60	2 x 320 SSD	-	10 Gigabit[4]

Instance Family	Instance Type	Processor Arch	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	EBS-optimized Available	Network Performance
Storage optimized	i2.xlarge	64-bit	4	14	30.5	1 x 800 SSD	Yes	Moderate
Storage optimized	i2.2xlarge	64-bit	8	27	61	2 x 800 SSD	Yes	High
Storage optimized	i2.4xlarge	64-bit	16	53	122	4 x 800 SSD	Yes	High
Storage optimized	i2.8xlarge	64-bit	32	104	244	8 x 800 SSD	-	10 Gigabit[4]

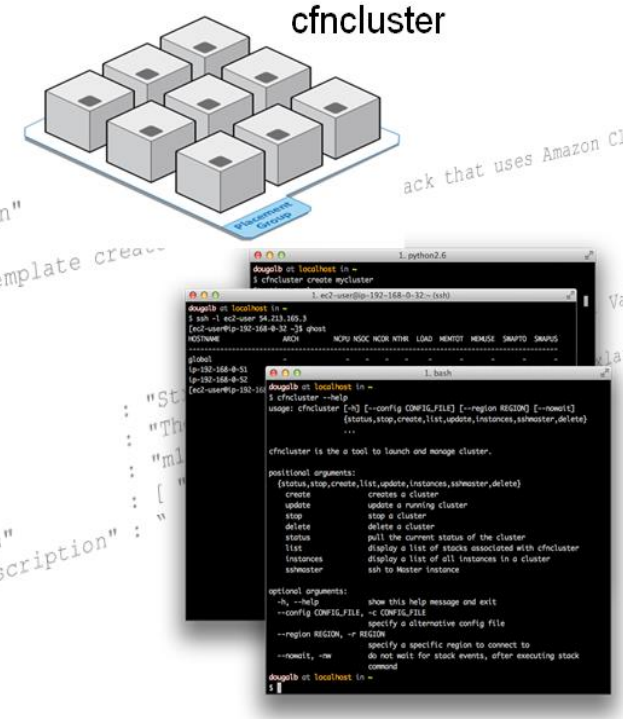
Automation allows easier management of clusters, and monitoring of jobs and costs



In a secure Virtual Private Cloud

Many HPC Deployment Methods

- Traditional HPC schedulers and cluster managers
  - MOAB/Torque, Bright
  - LSF/OpenLava, Grid Engine, SLURM, Condor
- “Born in the cloud” tools
  - MIT StarCluster
  - Cycle Computing CycleServer
- AWS-provided tools and APIs
  - Cloudformation, Auto Scaling
  - cfncluster (github.com/awslabs/cfncluster)



Multiple Consumption Models

On-Demand

Pay for compute capacity by the hour with no long-term commitments

For spiky workloads, or to define needs

Reserved

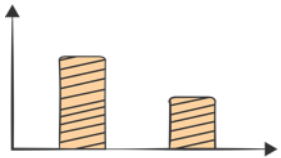
Make a low, one-time payment and receive a significant discount on the hourly charge

For committed utilization

Spot

Bid for unused capacity, charged at a Spot Price which fluctuates based on supply and demand

For time-insensitive or transient workloads



Dedicated

Launch instances within Amazon VPC that run on hardware dedicated to a single customer

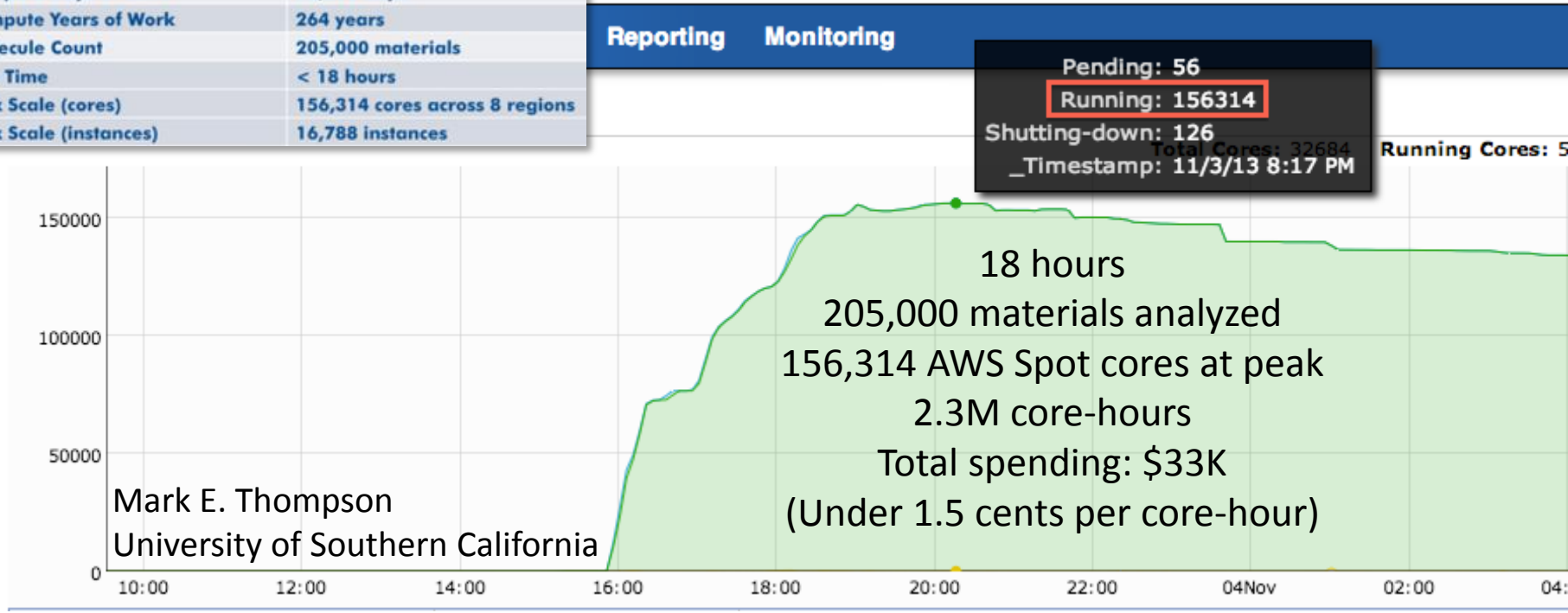
For highly sensitive or compliance related workloads

Spot Instances for Scale-Out Computing

amazon web services

CYCLECOMPUTING
LEADER IN CONDOR GRID COMPUTING SOLUTIONS

Metric	Count
Compute Hours of Work	2,312,959 hours
Compute Days of Work	96,373 days
Compute Years of Work	264 years
Molecule Count	205,000 materials
Run Time	< 18 hours
Max Scale (cores)	156,314 cores across 8 regions
Max Scale (instances)	16,788 instances

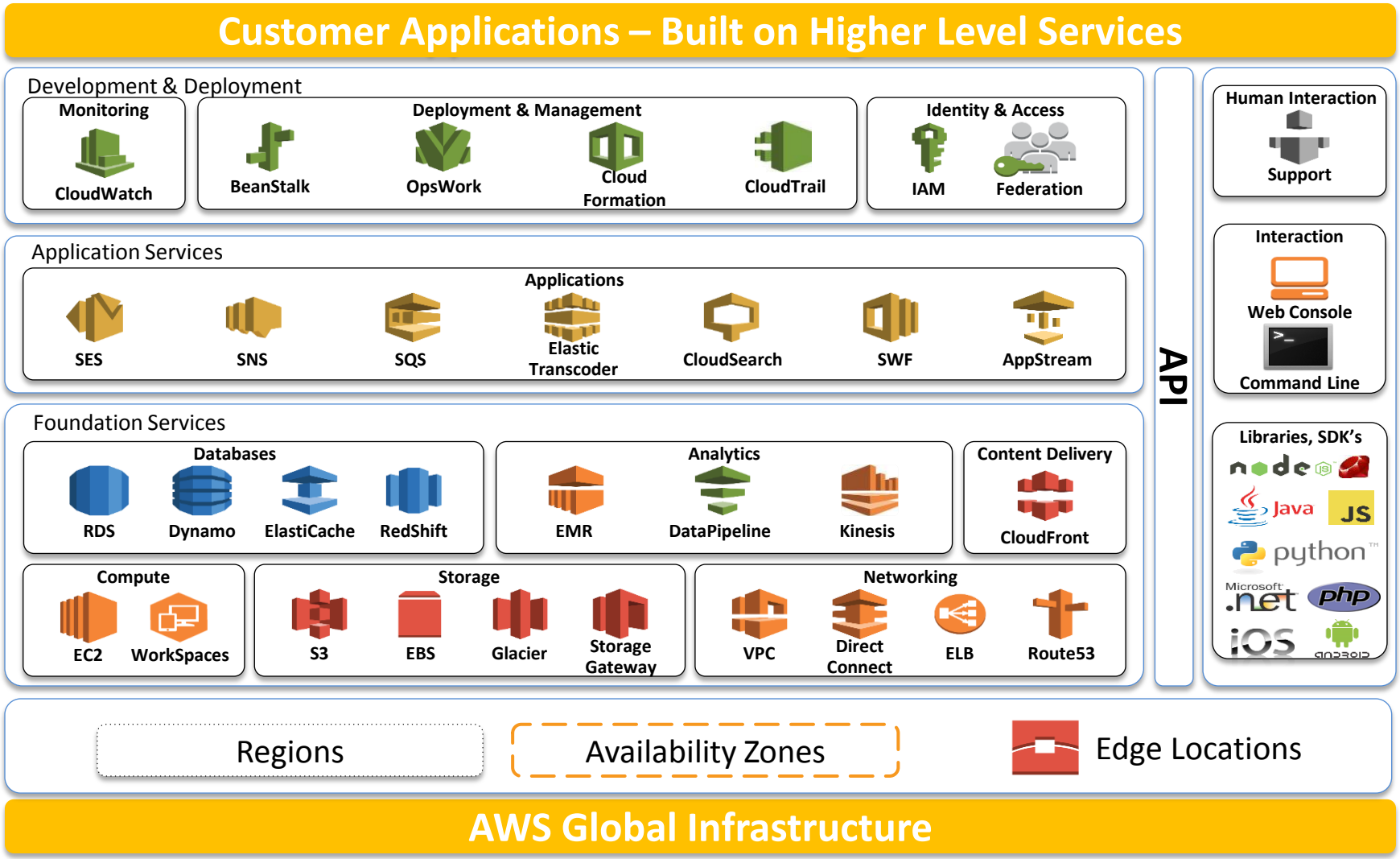


Mark E. Thompson
University of Southern California

The Democratizing of HPC

?

- Deployed in one AWS region (we have 10)
- In one availability zone in that region
- Using one placement group in that AZ
- Consisting entirely of one new instance type (c3.8xlarge)
- Brought up, benchmarked, and torn down in hours

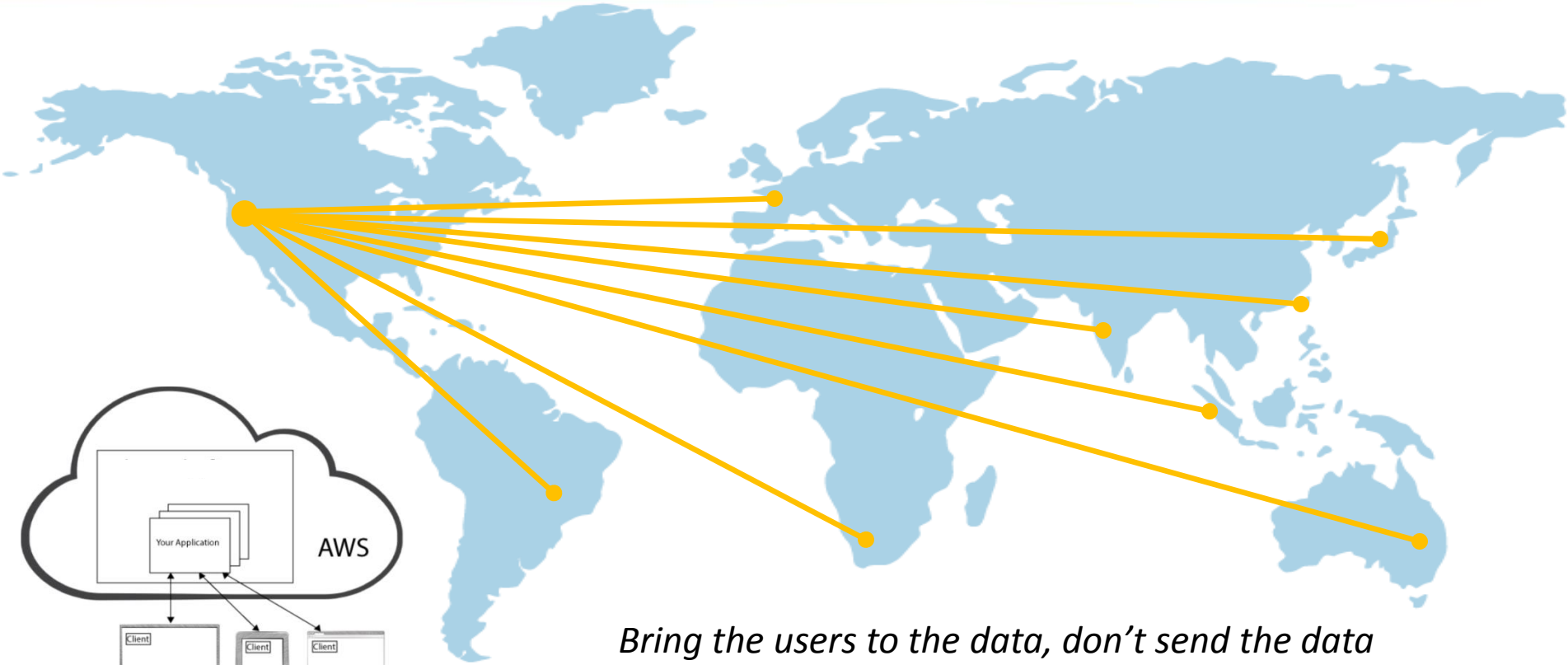

Customer Applications – Built on Higher Level Services
Development & Deployment
Monitoring
CloudWatch
Deployment & Management
BeanStalk
OpsWork
Cloud Formation
CloudTrail
Identity & Access
IAM
Federation
Application Services
Applications
SES
SNS
SQS
Elastic Transcoder
CloudSearch
SWF
AppStream
Foundation Services
Databases
RDS
Dynamo
ElastiCache
RedShift
Analytics
EMR
DataPipeline
Kinesis
Content Delivery
CloudFront
Compute
EC2
WorkSpaces
Storage
S3
EBS
Glacier
Storage Gateway
Networking
VPC
Direct Connect
ELB
Route53
API
Human Interaction
Support
Interaction
Web Console
Command Line
Libraries, SDK's
Regions
Availability Zones
Edge Locations
AWS Global Infrastructure

Cloud Collaboration

is

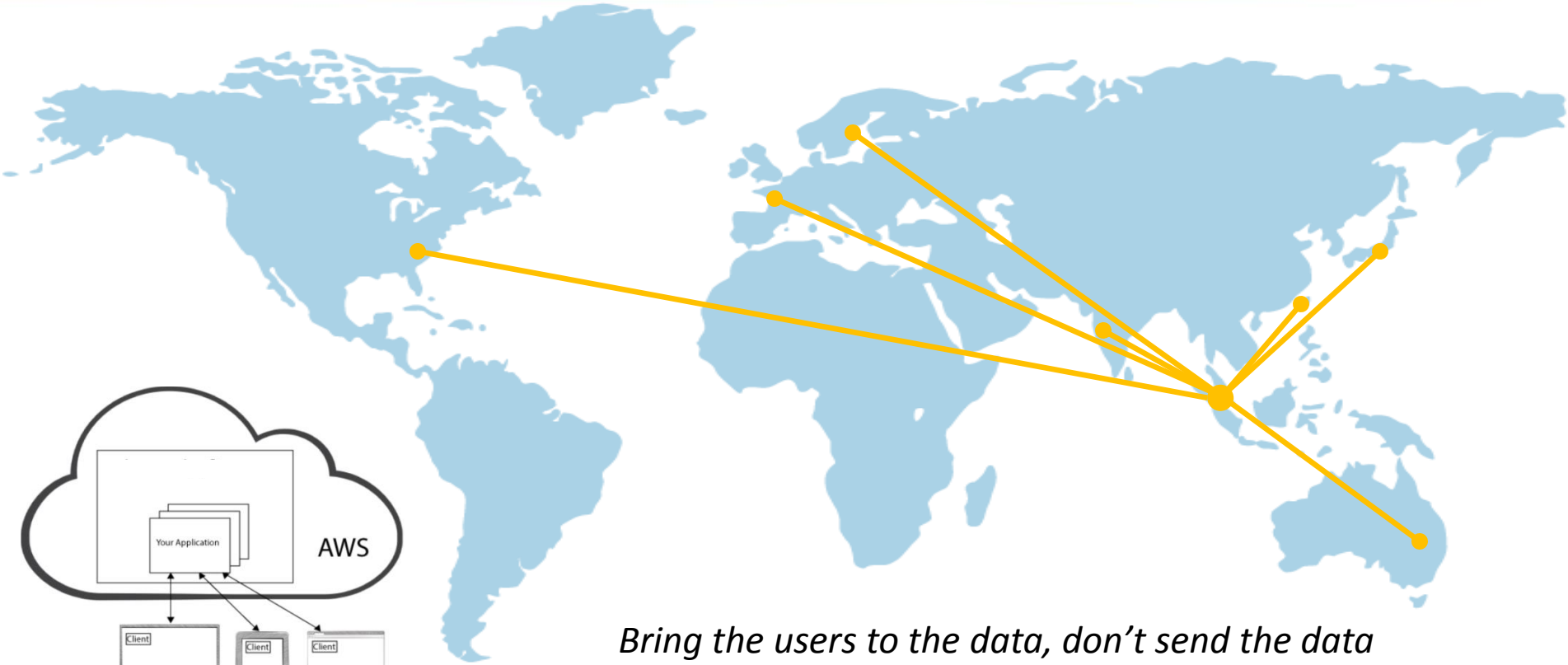
Secure Collaboration

Collaboration is More Secure in the Cloud

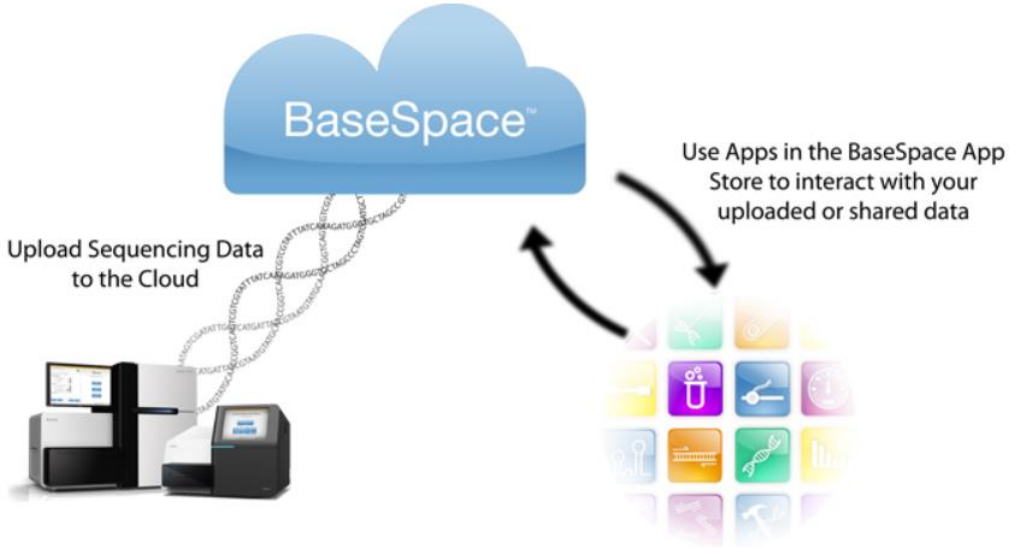


Bring the users to the data, don't send the data to the users

Collaboration is More Secure in the Cloud



Bring the users to the data, don't send the data to the users



Technical Note: Sequencing

illumina

BaseSpace™ Data Security

Data is safe in Illumina's cloud analysis and storage platform

Next-generation sequencing has fundamentally altered biomedical research, generating large amounts of analytical data. This large-scale data requires a scalable, robust, and secure storage and analysis solution. BaseSpace is a genomics cloud analysis built by Illumina on Amazon's Web Services (AWS). Illumina has chosen to work with AWS as the leader in cloud-based infrastructure, hosting customer-facing services and critical operations for both private industry and government departments including Treasury, DOE, and State.

Data security is a key concern in making the decision to move to cloud-based genomic storage and analysis. Illumina provides a combination of Amazon's comprehensive and well-tested approach to platform security, overlaid with Illumina's own security testing and procedures, to provide a cloud genomics solution that meets or exceeds the security provided by many institutional IT infrastructures.

The BaseSpace Data Model

Processing a flow cell on a sequencing instrument produces a variety of files, collectively referred to as a run. A run contains log files, instrument health data, run metrics, and base call information (*.bcl files), which are demultiplexed in BaseSpace to create the samples used in secondary analysis.[1]

Samples are analyzed by launching BaseSpace apps. BaseSpace apps are processing software and routines that interact with BaseSpace data through the API. User-level authentication and in-flight data encryption are enforced for every app that requests access to BaseSpace data. Files that are output from apps are stored in an object called AppResults. For example, when a resequencing app executes alignment and variant calling, an AppResult is created for each sample. AppResults can be used as inputs to apps as well. App sessions are created to record every time an app is launched. Finally, projects are simple containers that store samples and AppResults (Fig. 1).[1]

Security of Data in Flight

Data transfer is the major part of communication between the genomic sequencing instruments and the data analysis and storage servers. Illumina has implemented several security measures to make sure your data is protected in flight.

Secure Connection to Instrument

The decision to send data to BaseSpace is always initiated by the user during run set-up. If BaseSpace is chosen, the run is authenticated against and tracked to a user's BaseSpace / MyIllumina account. The

Collaboration with Secure Remote Access

Data and computation hosted in a secure, customer-managed virtual private cloud, with controlled access via a wide variety of client devices.

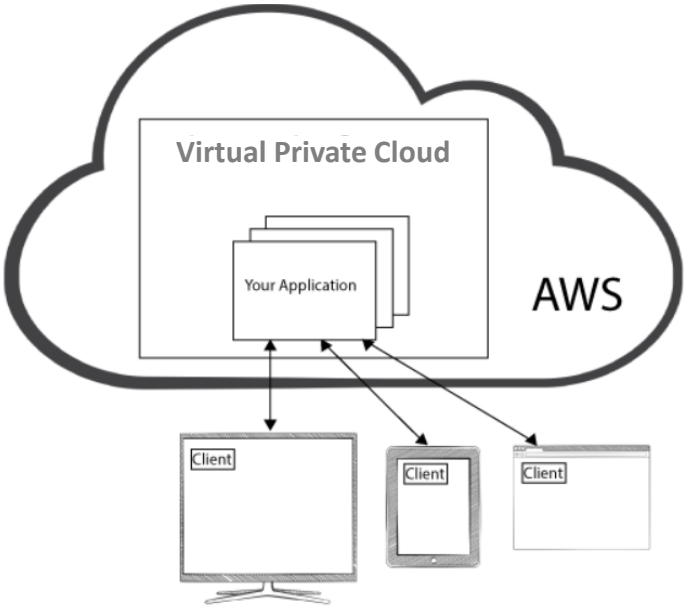
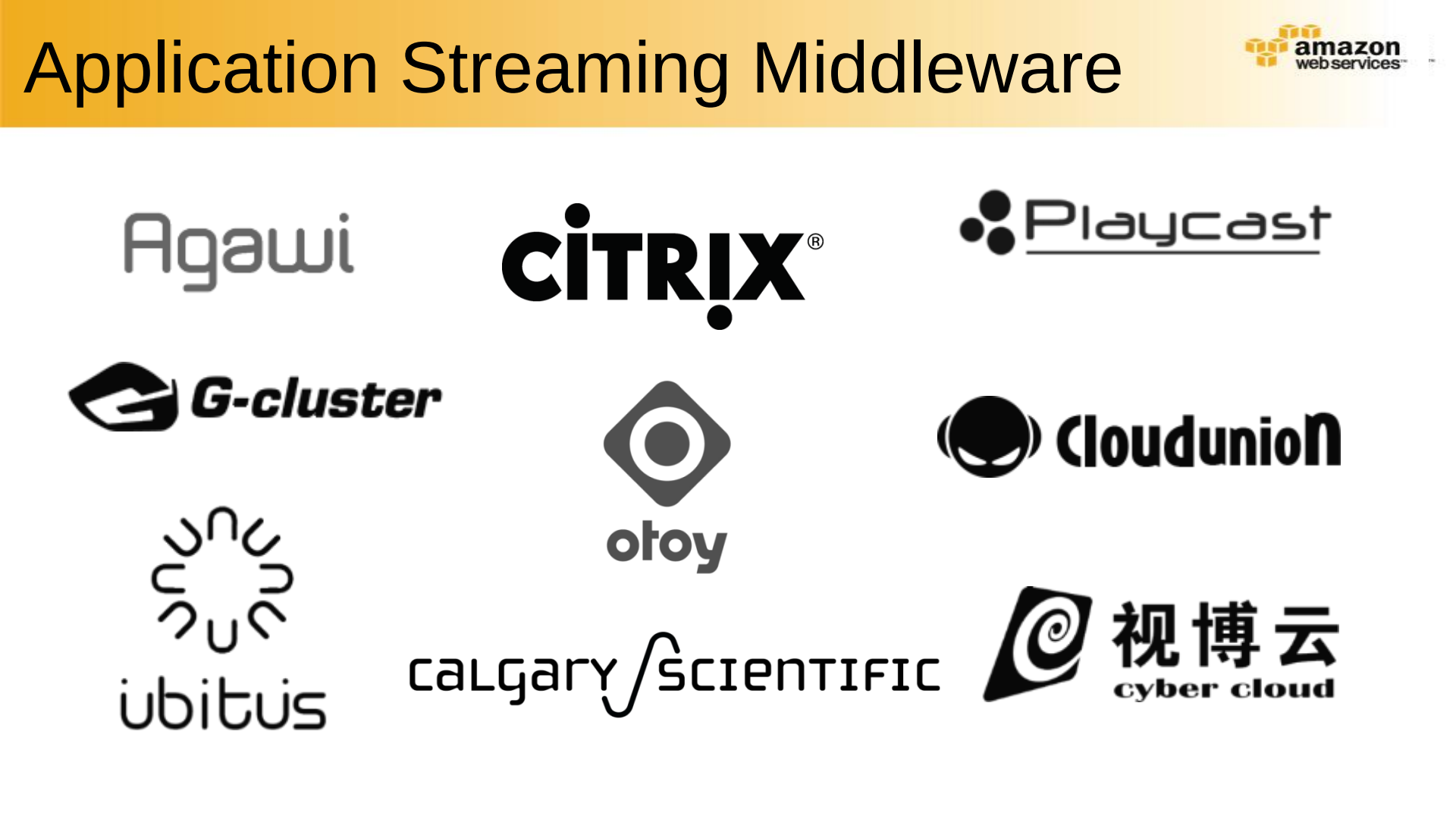


Image courtesy of Calgary Scientific

NVIDIA GRID K520 in AWS Cloud

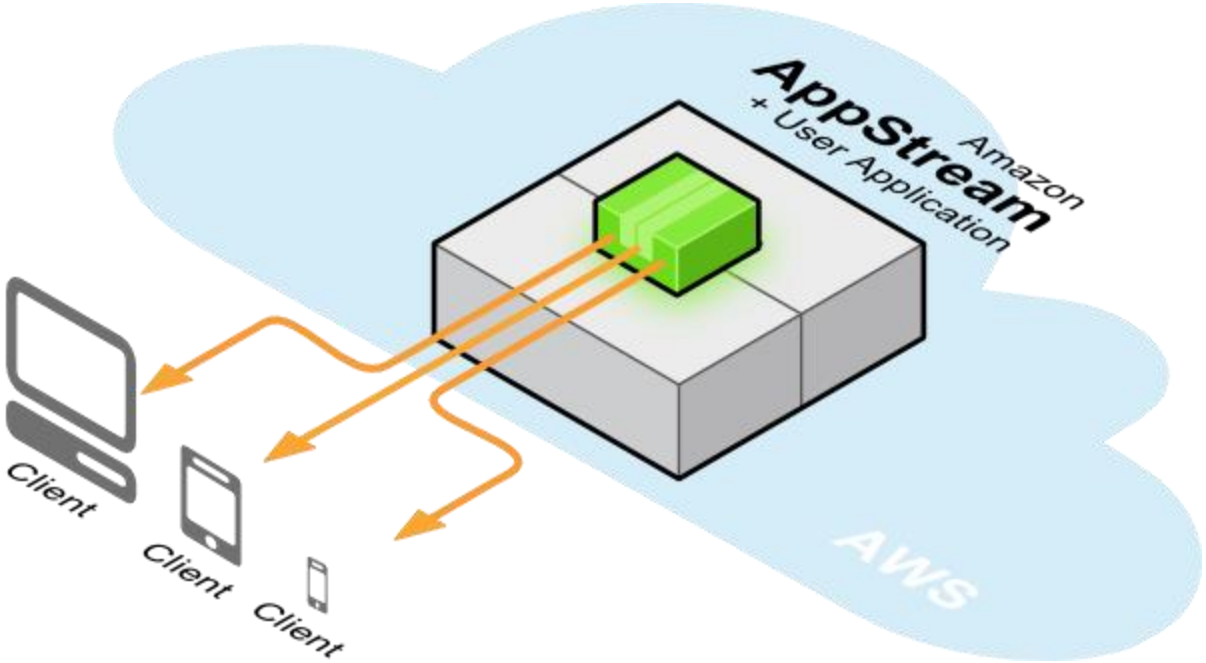
Product Name	GRID K520
GPUs	2 x GK104 GPUs
CUDA cores	3,072 (1,536 per GPU)
Core Clocks	800 MHz
Memory Size	8GB GDDR5 (4GB per GPU)
HW Video Encoder	2x h.264 (1 per GPU)
Power Consumption	225W
Supported APIs	OpenGL 4.3, DirectX 9, 10, 11, CUDA 5.5, OpenCL 1.1, NVFBC, NVIFR, NVENC

Application Streaming Middleware



Amazon AppStream

- Application Streaming
- Remote visualization
- Thin client 3D applications



Cloud Plus Big Data Equals Awesome

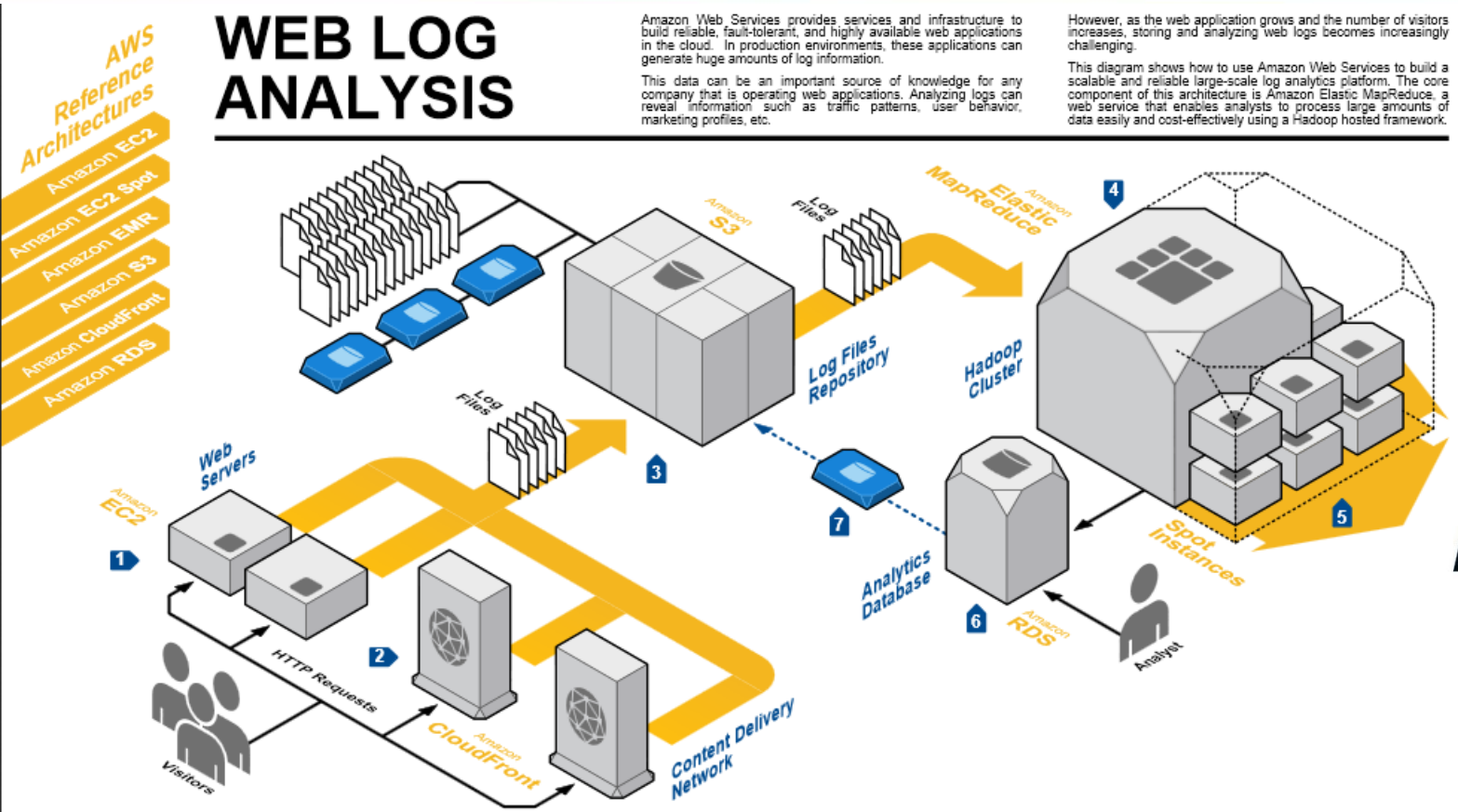
Integration of Big Data and HPC via Cloud

GE on AWS

GE's Global Business Integration Technologies Laboratory wanted to advance traditional manufacturing and create a dynamic network of people and machines that would allow collaboration, rapid prototyping, and product development for complex systems. GE had to adhere to U.S. International Traffic in Arms Regulations (ITAR) regulations and other compliance requirements. By using AWS GovCloud (US), GE developed a revolutionary manufacturing platform, Crowd-driven Ecosystem for Evolutionary Design (CEED), which connects people, materials, models, simulation, and equipment in an ITAR-compliant, secure, and distributed global environment.

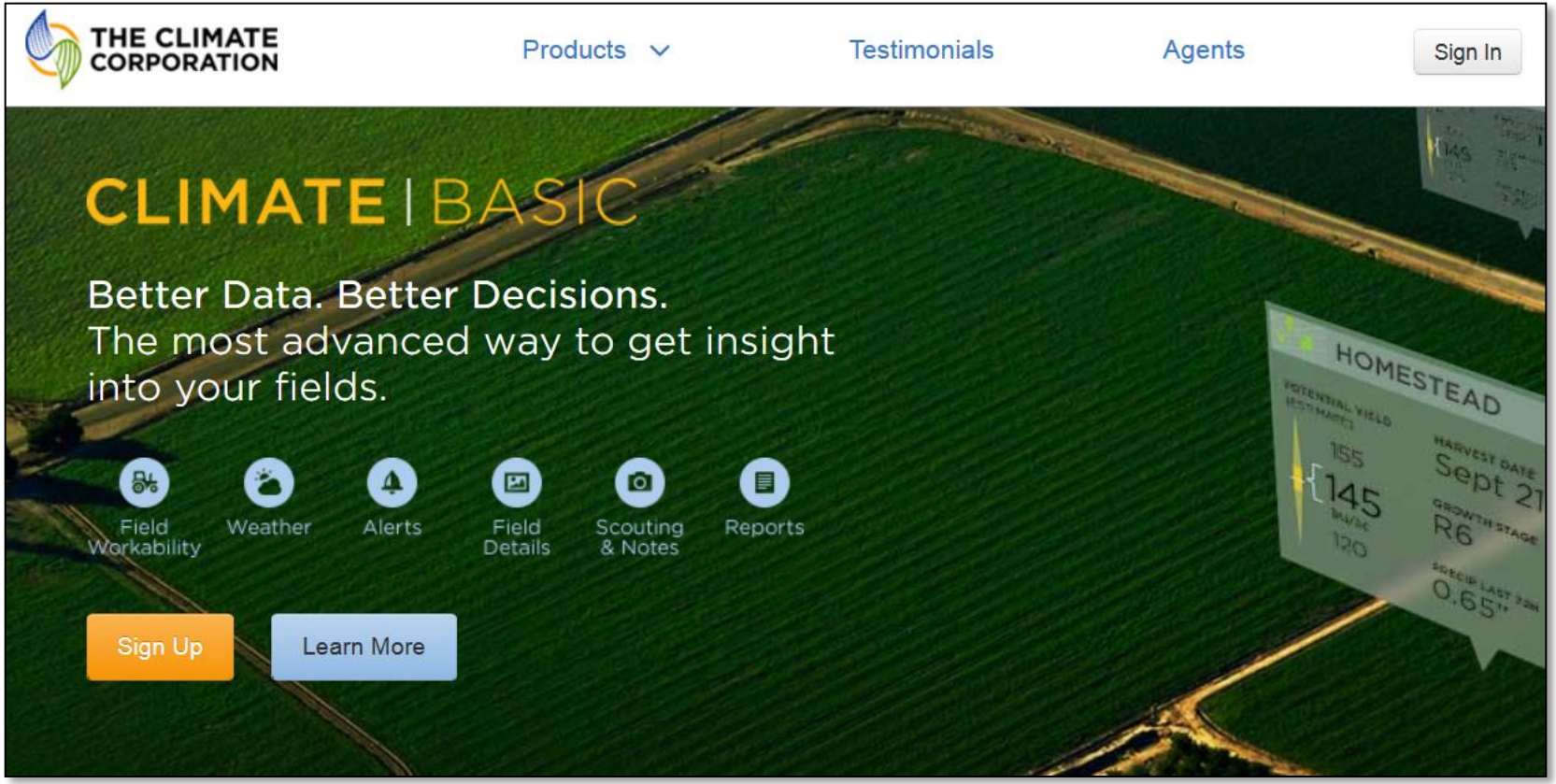
Cloud provides a global, distributed, secure, and scalable environment for collaborative design and manufacturing

AWS Has Always Been About Big Data



hadoop

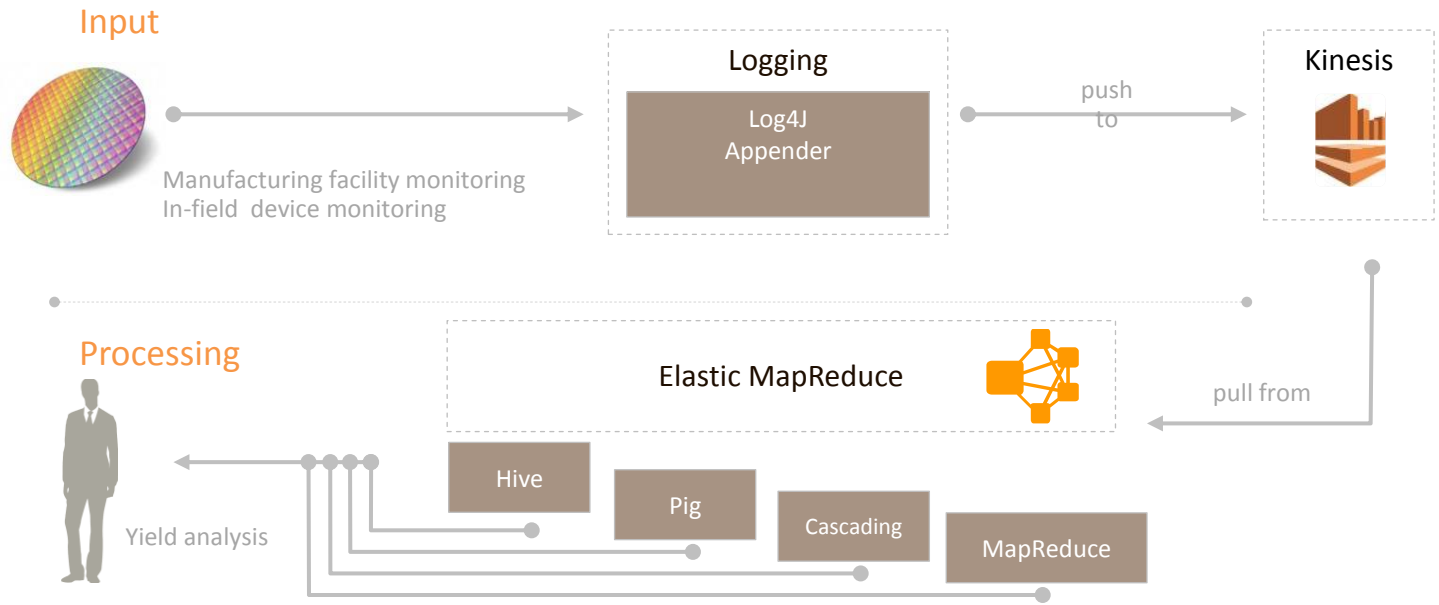
Big Data is Everywhere!



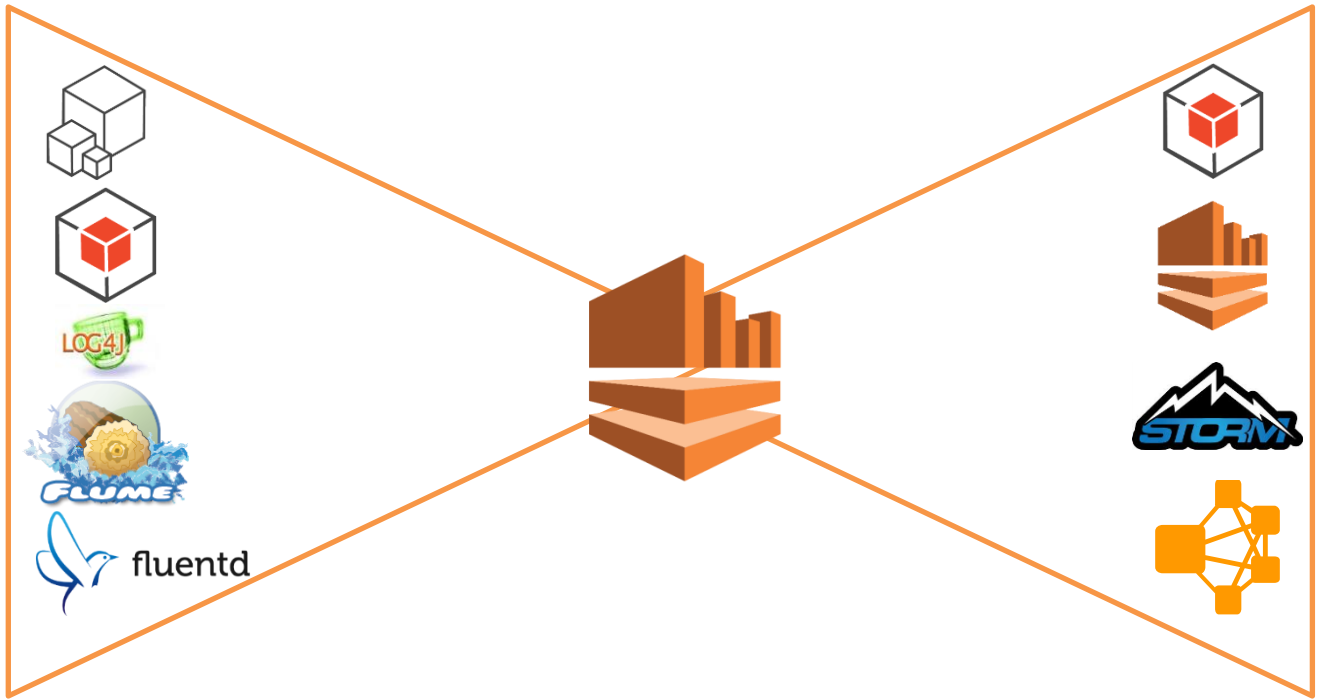
Big Data in Manufacturing

Managing big data for competitive advantage

- For design, engineering, production environments
- Quality and Yield Analysis, Statistical Process Control
- Structured and unstructured queries



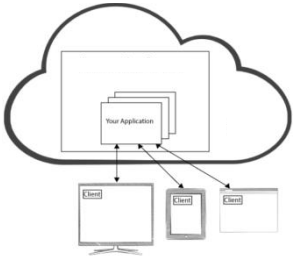
Sending & Reading Data from Kinesis Streams



Summary

Cloud for Scalability

Cloud for Global Collaboration



Cloud for Big Data