Compliant research data architecture and data sharing

Columbus, OH | May 9th 2024

Kiran Palsam

Solutions Architect
State and Local Government
Amazon Web Services (AWS)

Vishanth Davidar

Senior Solutions Architect
State and Local Government
Amazon Web Services (AWS)

Agenda

- How Amazon Web Services (AWS) can help research
- Research for health on AWS
- How to get started
- Conclusion

How AWS can help research

How AWS can help research

Science, not servers

Use latest compute when you need it to do large scale analysis

Collaboration

Federate datasets across institutions and platform for reproducing science

Security

A collection of tools to protect data and privacy

Baylor College of Medicine's Human Genome Sequencing Center uses AWS to innovate

Challenge

One of the projects Baylor HGSC is involved with is the Cohorts for Heart and Aging Research in Genomic Epidemiology project (CHARGE). Baylor needed a cost efficient, easily maintainable solution that would enable it to provide safe, effective, worldwide collaboration without delays caused by setting up a physical infrastructure. The solution also needed to meet clinical standards and HIPAA requirements.

Solution

Baylor decided to partner with DNAnexus, which provides an API-based Platform as a Service (PaaS) that enables clinical and research enterprises to efficiently and securely move their analysis pipelines and data into AWS.

Benefits

- Completed its first analysis in 10 days—five times faster than with the local infrastructure—and was able to share the findings quickly.
- The scalability of AWS helps CHARGE scientists gain more predictive power over the conditions they are studying.

BCM
Baylor College of Medicine

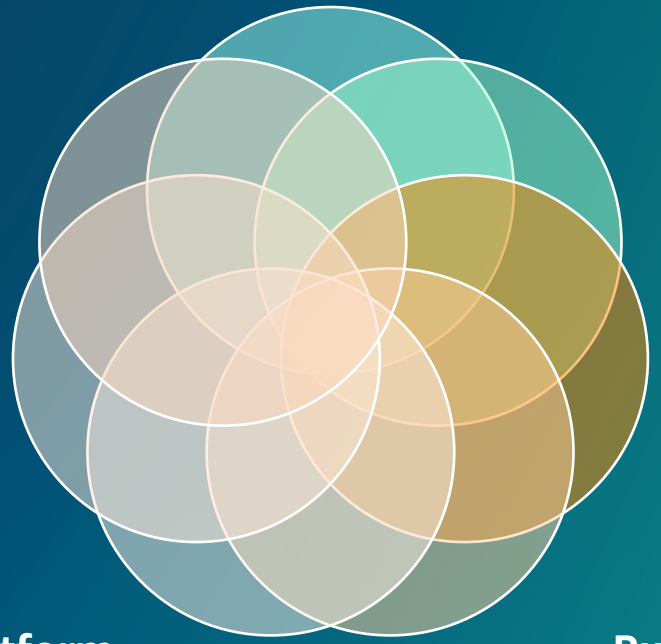
"There are all kinds of limitations in our ability to find the horizons of science. But now, thanks to AWS and DNAnexus, we can focus on the science instead of the infrastructure."

Research for health on AWS provides

Secure and Compliant Researcher Enclave

Cohort building and querying

Federated data and compute infrastructure



End to end AI/ML platform

High-performance next-generation GPU, FPGA, quantum compute

Open data platform for OMOP, FAIR, open source, NIH DMS ready

Purpose built services for genomics, FHIR, and medical imaging

Research for health

Secure, self-service research from research IT for researchers

Research data platform	Modern omics	HPC as a service	ML for research	Trusted Research Enclave	Next-gen research compute	Data sharing and federation
Singular OMOP data platform for all research data	High scale, high performance genomics cloud services	Centralized large scale Batch and HPC research on demand clusters	One platform for all ML and data science needs	Secure and isolated research enclaves for PIs	Massive compute scale with latest generation compute	Secure data sharing, and data cleanroom

Research use cases

Multi-modal research data, de-identified data, research data meshes, genomics storage, imaging storage	Secondary Analysis, Tertiary Analysis, Genomics workflows, native Genomics CLI support	Research HPC clusters, SLURM, Genomics, Massive Batch computing	Deep learning, machine learning Imaging AI, AI assisted annotation, PyTorch, TensorFlow, CNN, DNN	Enclaves for researcher workbenches for data, compute, data science, and visualization	Nextgen NVIDIA GPUs, FPGAs, ARM, Intel, AMD, and Quantum	NIH DMS 2023 sharing, research consortia, federated learning, federated queries

Children's Hospital of Philadelphia (CHOP) accelerates pediatric research using AWS-powered data resource

Challenge

As medical researchers generate more and more clinical data, they're faced with the challenge of storing and organizing that data so that researchers can access, study, and cross-reference it to facilitate medical breakthroughs.

Benefits

CHOP provided the research community with access to genomic and associated clinical data and increased KFDRC's collaborative potential.

CHOP stored 26 billion occurrences of 215 million unique genomic variants from 5,000 participants, while meeting the FHIR industry standard.

Solution

CHOP built the Gabriella Miller Kids First Data Resource Center (KFDRC), a data source that brings genomics, clinical and imaging data as an open resource for researchers to focus on discovers in pediatric cancer and structural birth defects.

Children's Hospital of Philadelphia

"All of our system is currently built on AWS. . . . We went from zero to managing a few petabytes of genomic data within a year using this setup."

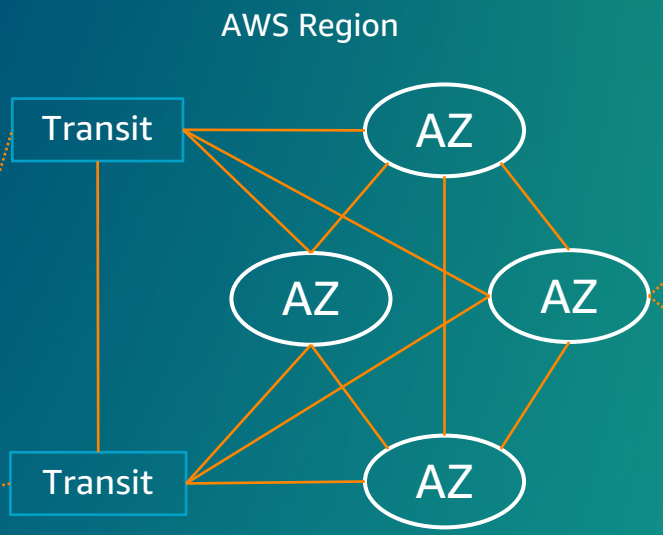
Allison Heath
Director of Data Technology and Innovation, Center for Data-Driven Discovery in Biomedicine

AWS Global Infrastructure for research

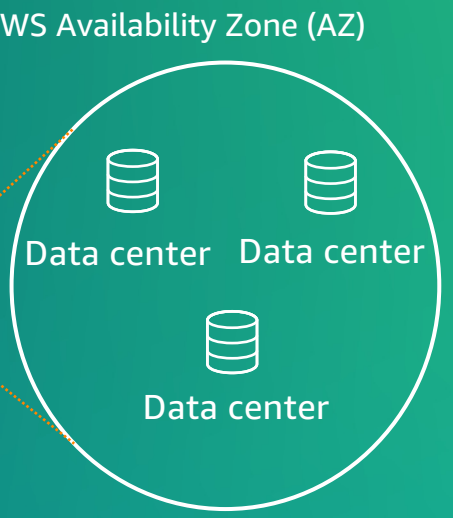
Region & number of Availability Zones

Announced Regions

Auckland, Canada, Israel, Malaysia and Thailand

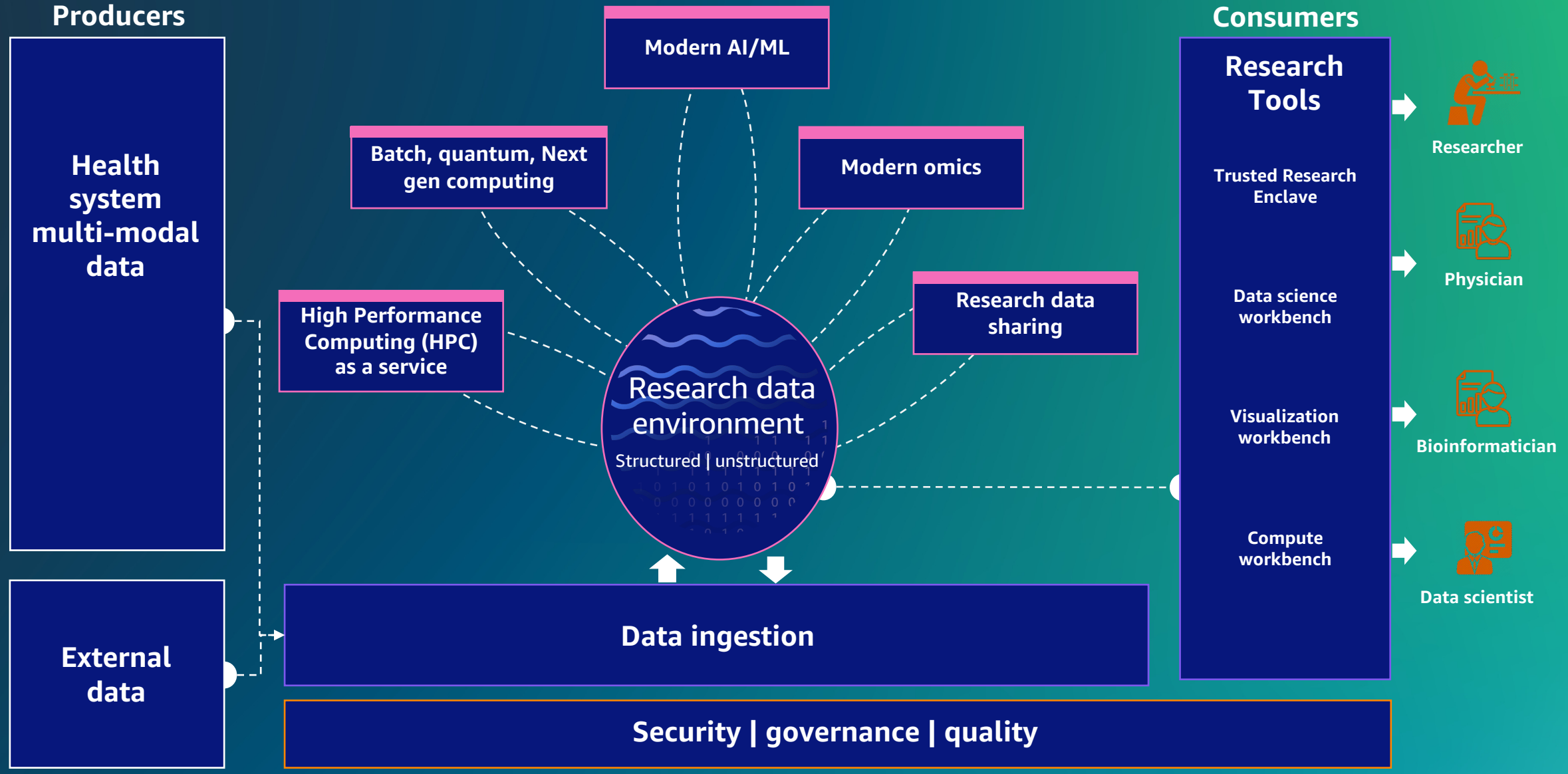


A **Region** is a physical location in the world where we have multiple **Availability Zones**



Availability Zones consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities

Research for health on AWS



What is AWS Batch?

Job scheduler

- Schedules and runs jobs asynchronously
- Manages dependencies

Resource orchestrator

- Manages and optimizes compute resources
- Scales up/down as needed
- Utilizes the right compute resources for the job

Fully managed

Integrated with AWS services

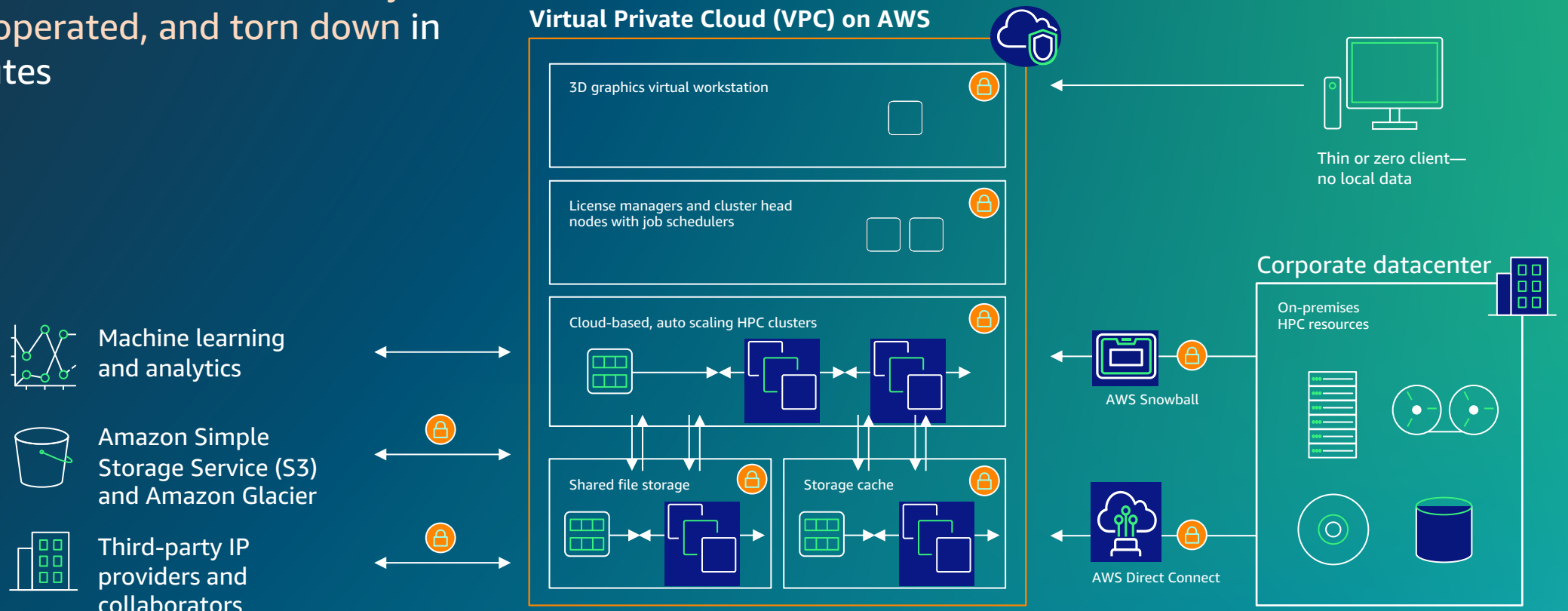
Massive scalability

Optimized resource provisioning

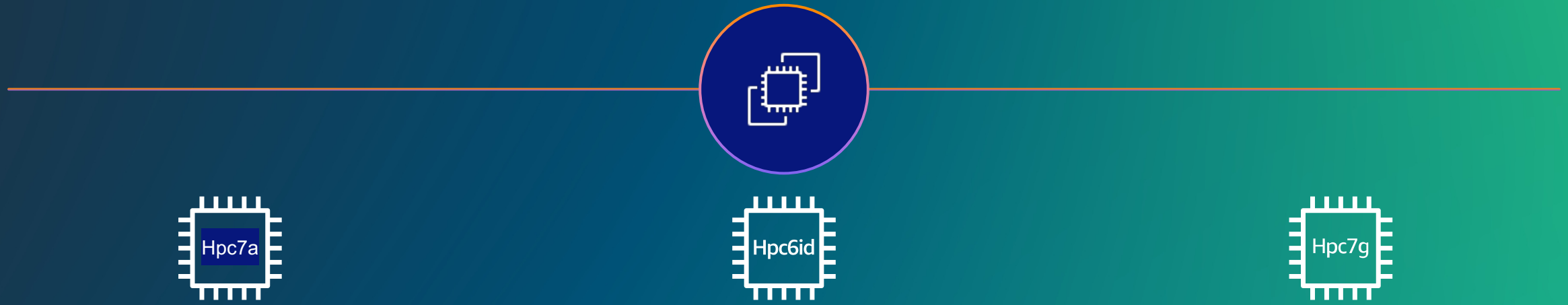
Cost efficient

High Performance Computing (HPC) on AWS ParallelCluster

On AWS, secure and well-optimized HPC clusters can be automatically created, operated, and torn down in just minutes



HPC-optimized Amazon Elastic Compute Cloud (EC2) instances



Powered by fourth-generation AMD EPYC processors

Amazon EC2 Hpc7a instances

Compute-intensive applications like Computational Fluid Dynamics and Numerical Weather Prediction

3.6Ghz 192 cores AMD EPYC 9R14

768GB RAM

300Gbps EFA

Powered by third-generation Intel Xeon Scalable processors

Amazon EC2 Hpc6id instances

Memory-bound and data-intensive workloads like Finite Element Analysis and seismic simulations

3.5Ghz 64 cores Intel Xeon

1024GB RAM | 15.2TB NVMe

200Gbps EFA

Powered by the next-generation AWS Nitro System

Amazon EC2 Hpc7g instances

Based on custom AWS Graviton3E processors with low latency, and high network performance for MPI-based applications

2.6Ghz 64 cores Graviton3E

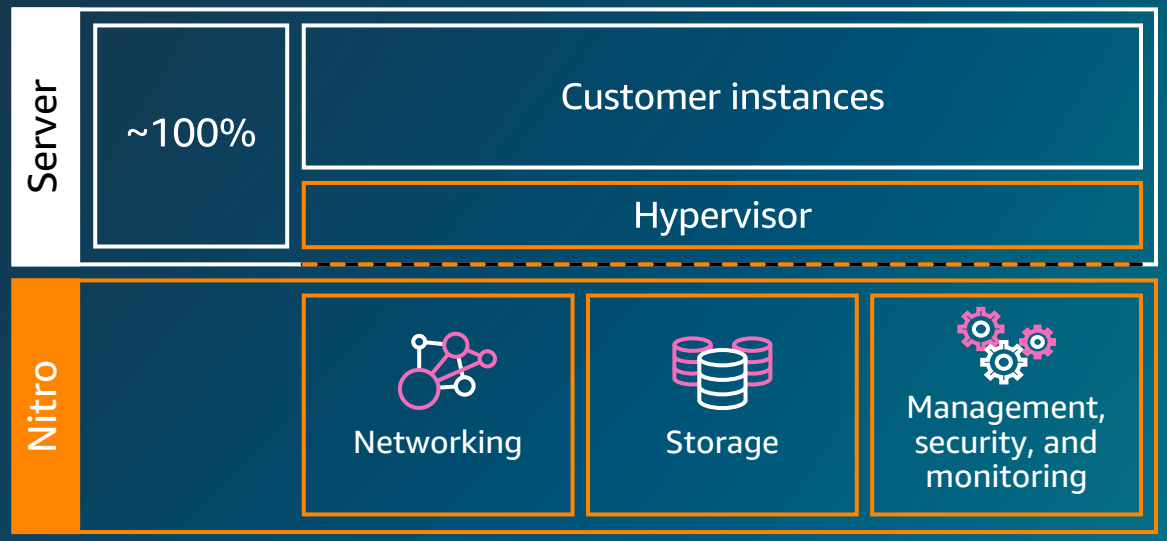
128GB RAM

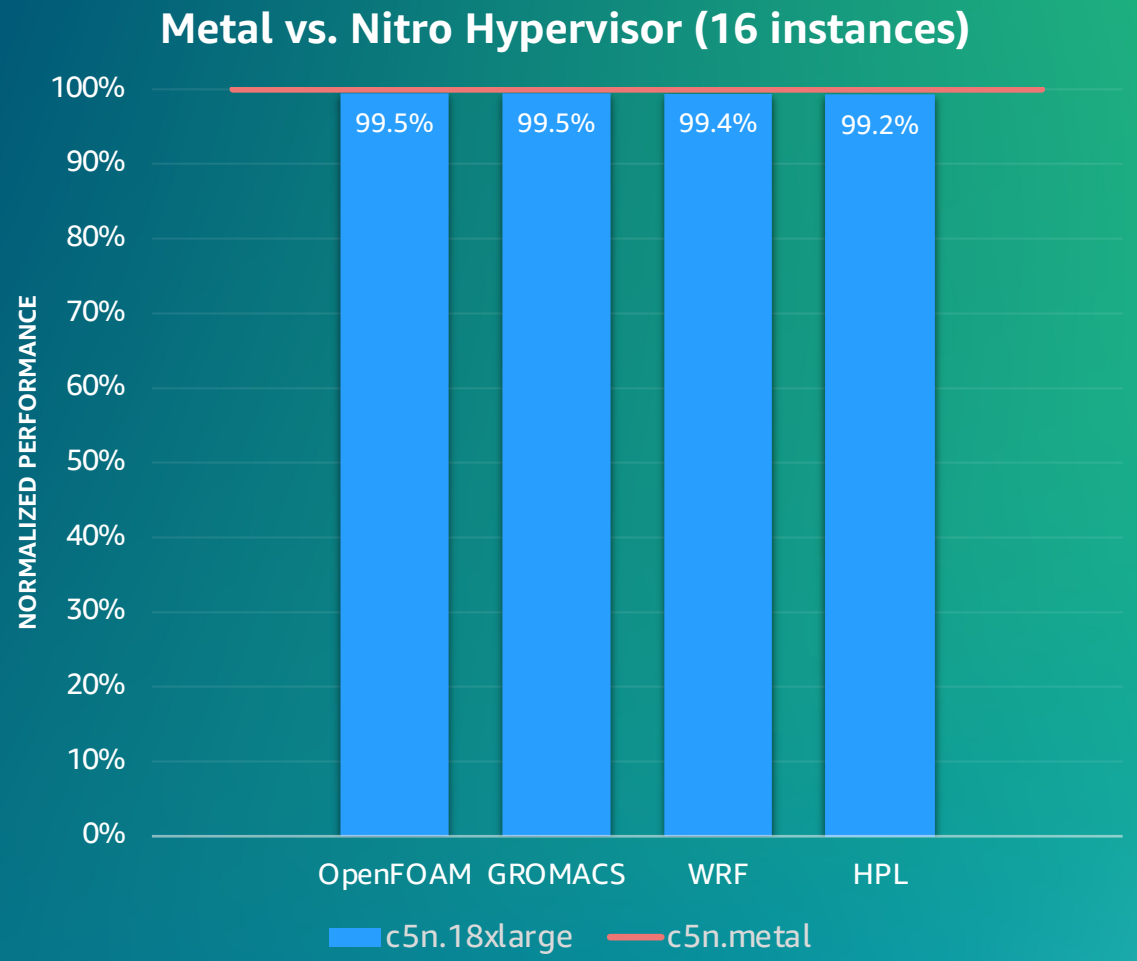
200Gbps EFA

aws

The AWS Nitro System

Engineered with a **hardware-based root of trust** using the Nitro Security Chip, allowing for the system to be continuously measured and validated





Amazon Braket – the AWS quantum computing service

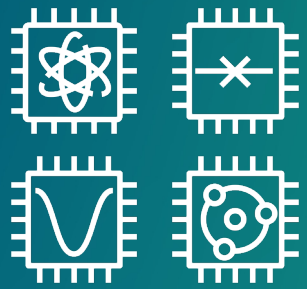
A fully-managed service that makes it easy for scientists and developers to explore quantum computing

Build

- Amazon Braket Python software development kit (SDK)
- Jupyter notebooks
- Command Line Interface (CLI)

Test

- Local simulators for rapid testing
- High-performance simulators



Run

- Access multiple quantum computers
- Combine quantum and classical resources

Analyze

- Monitor algorithms in almost real time
- Analyze algorithm results and performance

Research for health on AWS

Purpose-built health services

Services dedicated to healthcare and life sciences customers

AWS HealthOmics

Transform genomic, transcriptomic, and other omics data into insights

AWS HealthLake

Store, transform, transact, and analyze health data in minutes

AWS HealthImaging

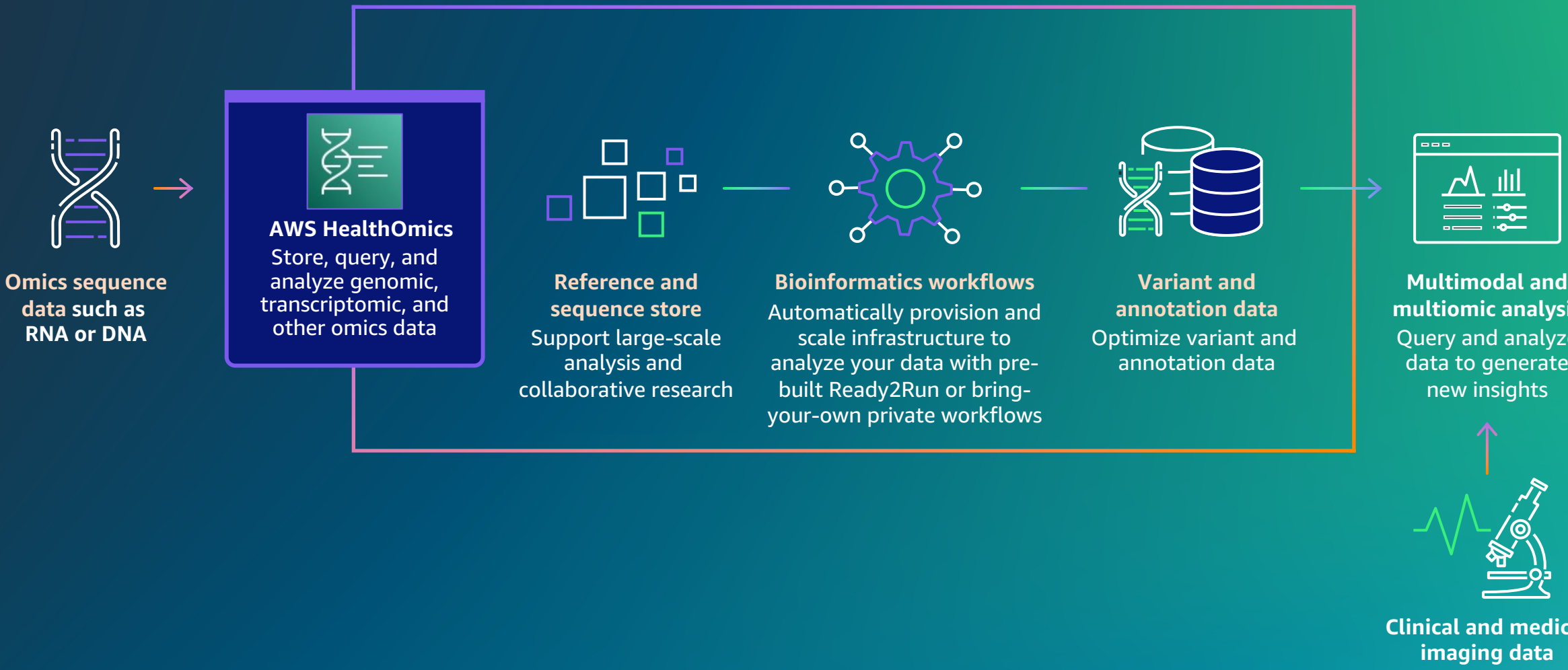
Store, analyze, and share medical images in the cloud at petabyte scale

AWS HealthScribe

Autogenerate clinical notes from patient-clinician conversations

AWS HealthOmics

Transform genomic, transcriptomic, and other omics data into insights



Amazon HealthLake

A managed, secure, HIPAA-eligible FHIR storage, transactional, and analytics service

Transactional FHIR server to power clinical applications

Store clinical/claims data using interoperability standards

Store patient medical history from multiple data sources in the normalized common data model (FHIR-based) format and leverage FHIR APIs to build transactional applications and patient 360 views

Build highly scalable interoperability solutions to meet regulatory needs

Leverage HealthLake Patient Access APIs and Bulk FHIR APIs with built-in support for US CORE and CARIN BB profile validation to meet the 21st Century Cures Act for patient access and interoperability requirements

Build secure applications using SMART on FHIR

Build patient360 end user clinical applications by integrating with OAuth2 compliant authorization service and securely access HealthLake data

AWS HealthImaging

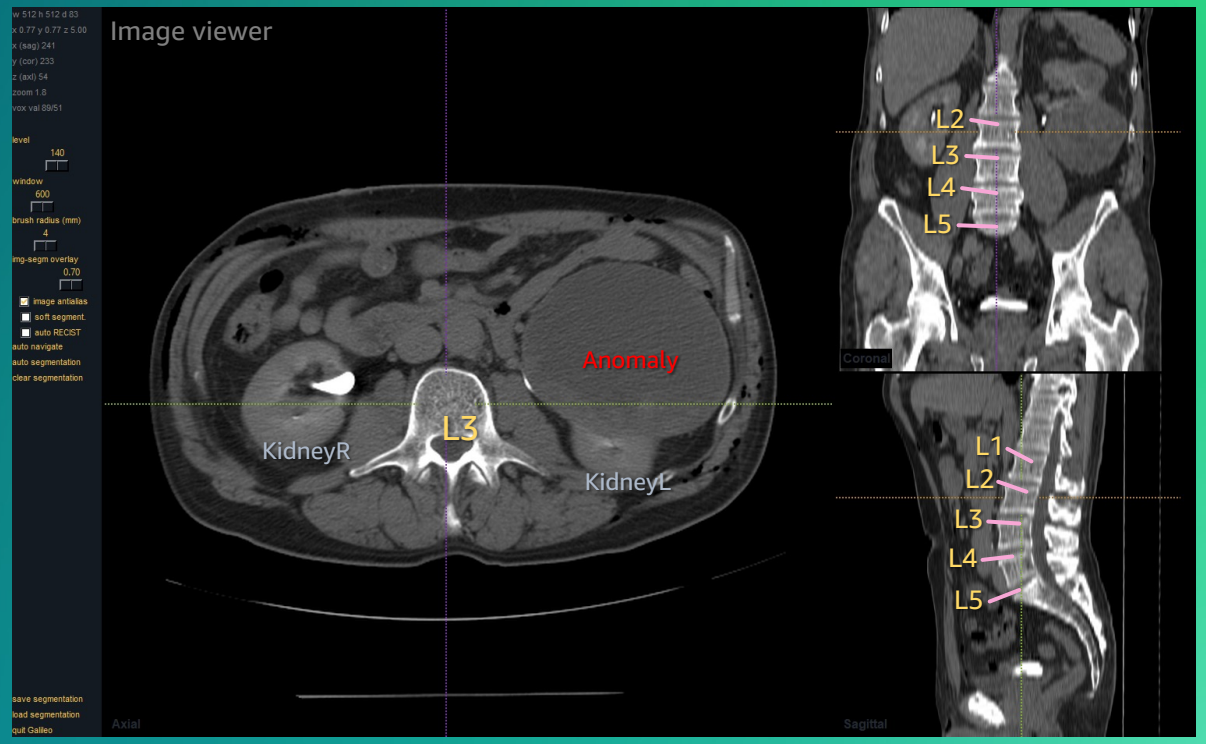
A HIPAA-eligible service for storing, analyzing, and sharing medical images at petabyte scale

A single copy of patient imaging data in the cloud

Reduce the cost of medical imaging storage **up to 40 percent** with AWS HealthImaging

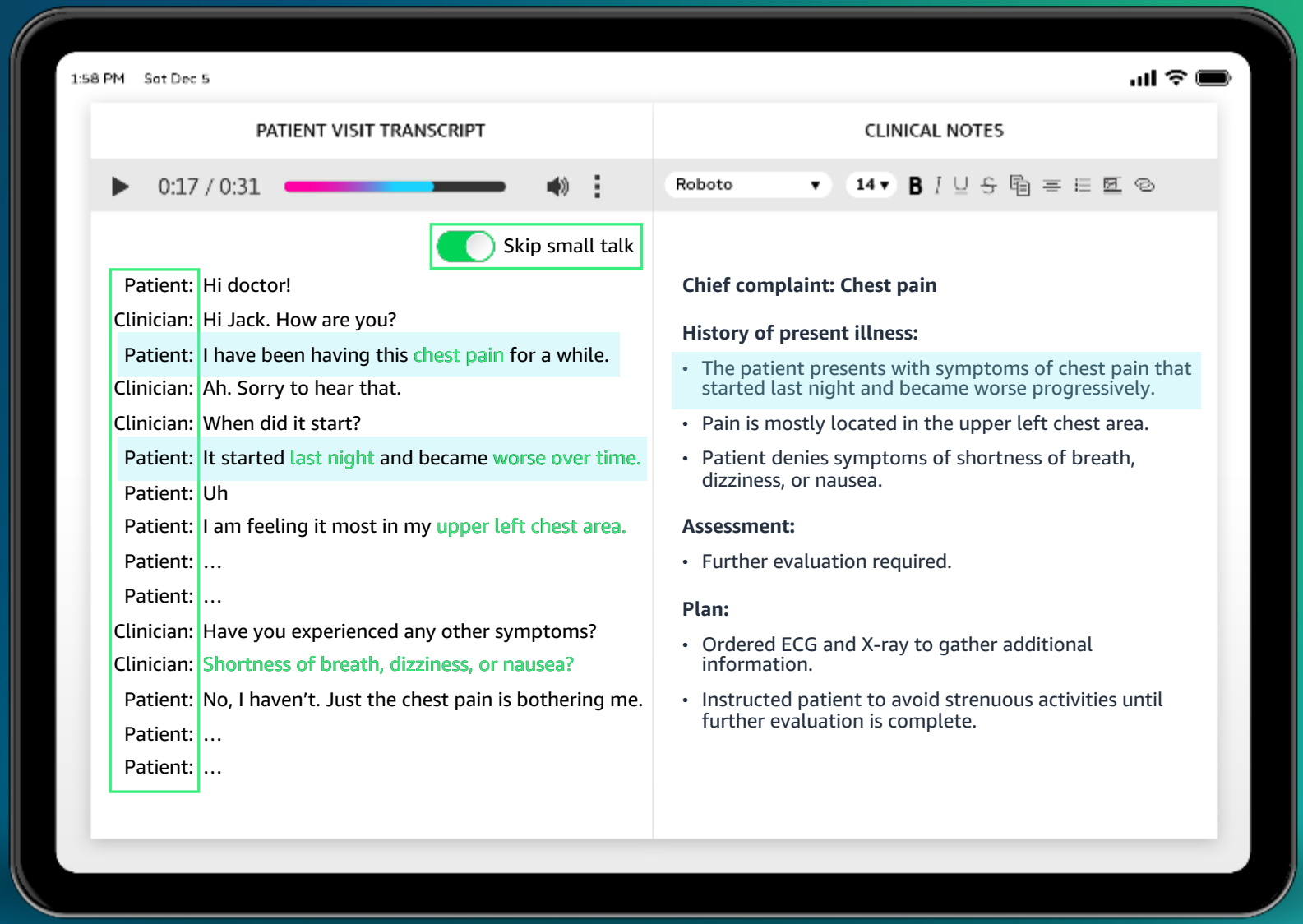
Sub-second image retrieval from anywhere

Fast access to images stored in the cloud



Amazon HealthScribe

- Rich consultation transcript
- Speaker role identification
- Dialogue classification
- Preliminary clinical notes
- References to transcript
- Extracted medical terms



Bring it all together on AWS

DATA GOVERNANCE

Multimodal assets

- Clinical records
- Omics
- Imaging

Data (such as FHIR) normalization

Data (such as variant) normalization

Metadata extraction

Analysis and query

- Amazon EMR
- Amazon Athena
- Amazon SageMaker
- Amazon QuickSight
- Third-party applications

Amazon QuickSight

Create persuasive insights and stories fast

Contextualize insights to drive actions

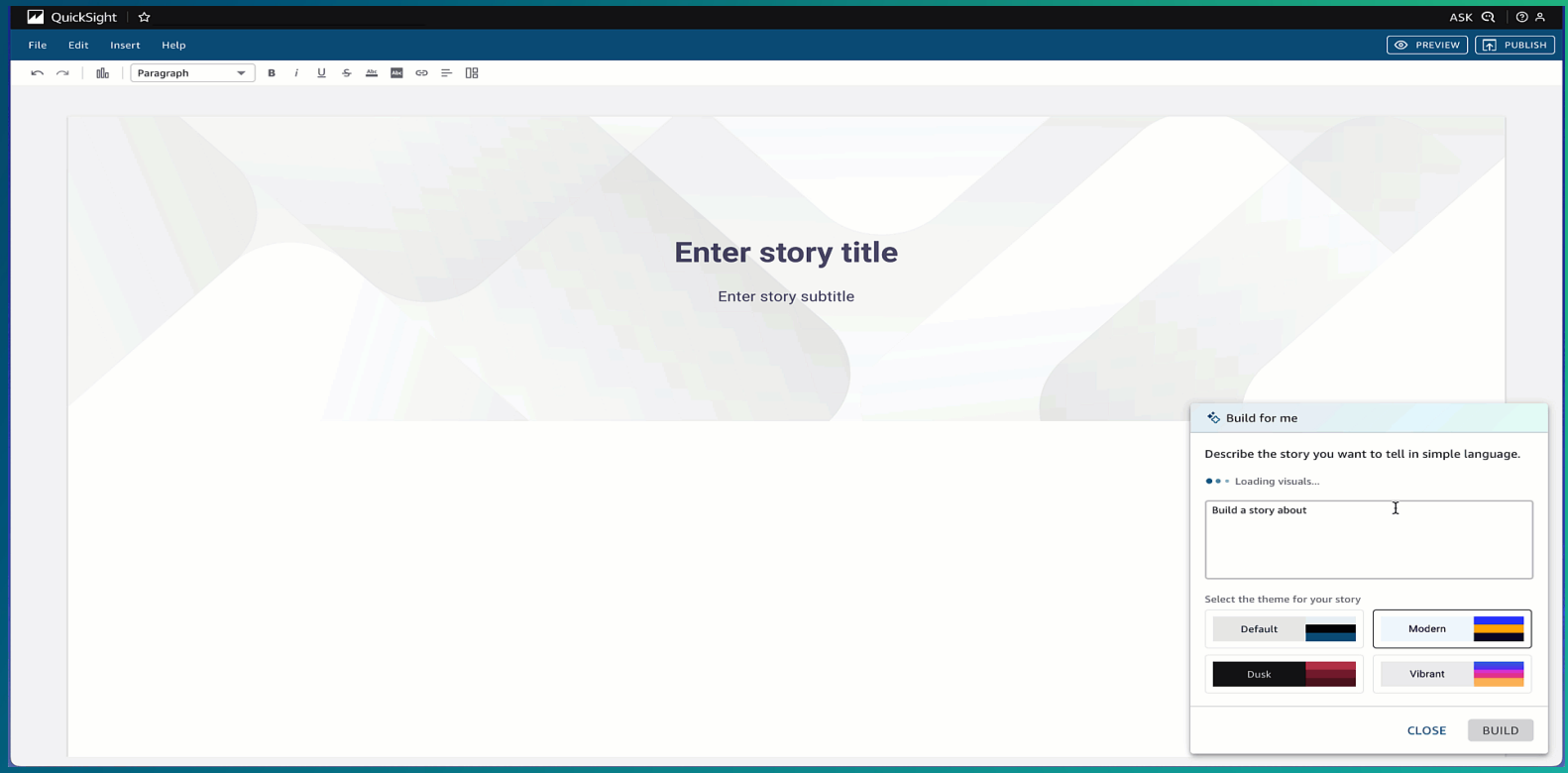
Interpret and share insights with visually compelling narratives

Generate stories using AI

With one sentence, save time by letting QuickSight build a powerful, data-driven story for you

Always up-to-date

Easily share your narrative with others, and refresh it with the newest data in just a few clicks



Amazon Bedrock

The easiest way to build and scale generative AI applications with foundational models (FMs)

Access a range of leading FMs through a single API

Privately customize FMs with your own data

Enable data security and compliance

Build agents that execute complex business tasks by dynamically invoking APIs

Extend the power of FMs with your data using retrieval augmented generation (RAG)

Get the best price performance without managing infrastructure

How it works

Amazon Bedrock

The easiest way to build and scale generative AI applications with FMs

Choose from leading FMs

Access a range of top FMs using a single API

Customize with your data

Privately fine-tune FMs with your data

Connect data sources

Connect FMs with your data sources for RAG

Enable task execution

Configure agents to dynamically invoke APIs

Amazon Bedrock supports leading foundation models

amazon

Amazon Titan

Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings, and search

AI21labs

Jurassic-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

ANTHROP\C

Claude 2

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems

cohere

Command

Text-generation model for business applications and embeddings model for search, clustering, or classification in 100+ languages

Meta

Llama 2

Fine-tuned models ideal for dialogue use cases and language tasks

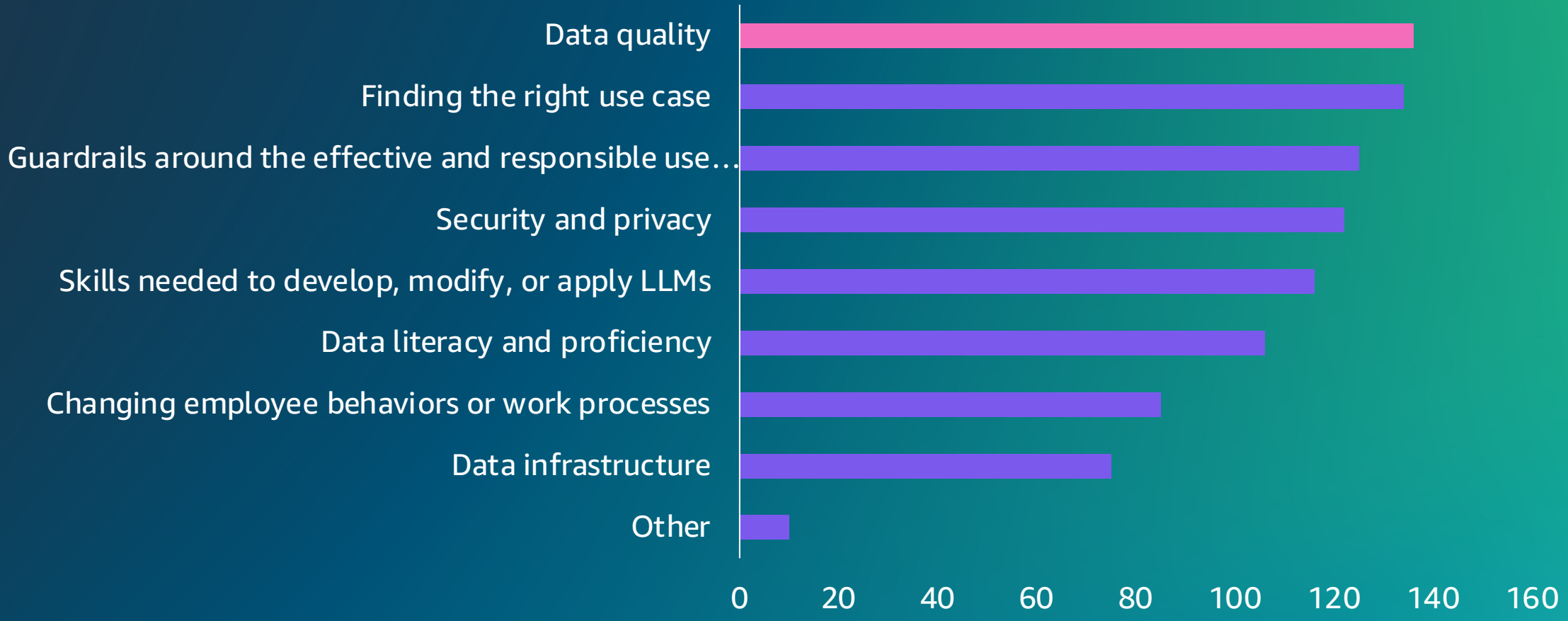
stability.ai

Stable Diffusion

Generation of unique, realistic, high-quality images, art, logos, and designs

Generative AI needs quality data

What is the biggest challenge in realizing the potential of generative AI?



AWS Glue Data Quality

Serverless and scalable

Serverless infrastructure powered by Deequ and tested at scale by Amazon teams to manage quality of 60 PB data lakes

Data quality rules and recommendations

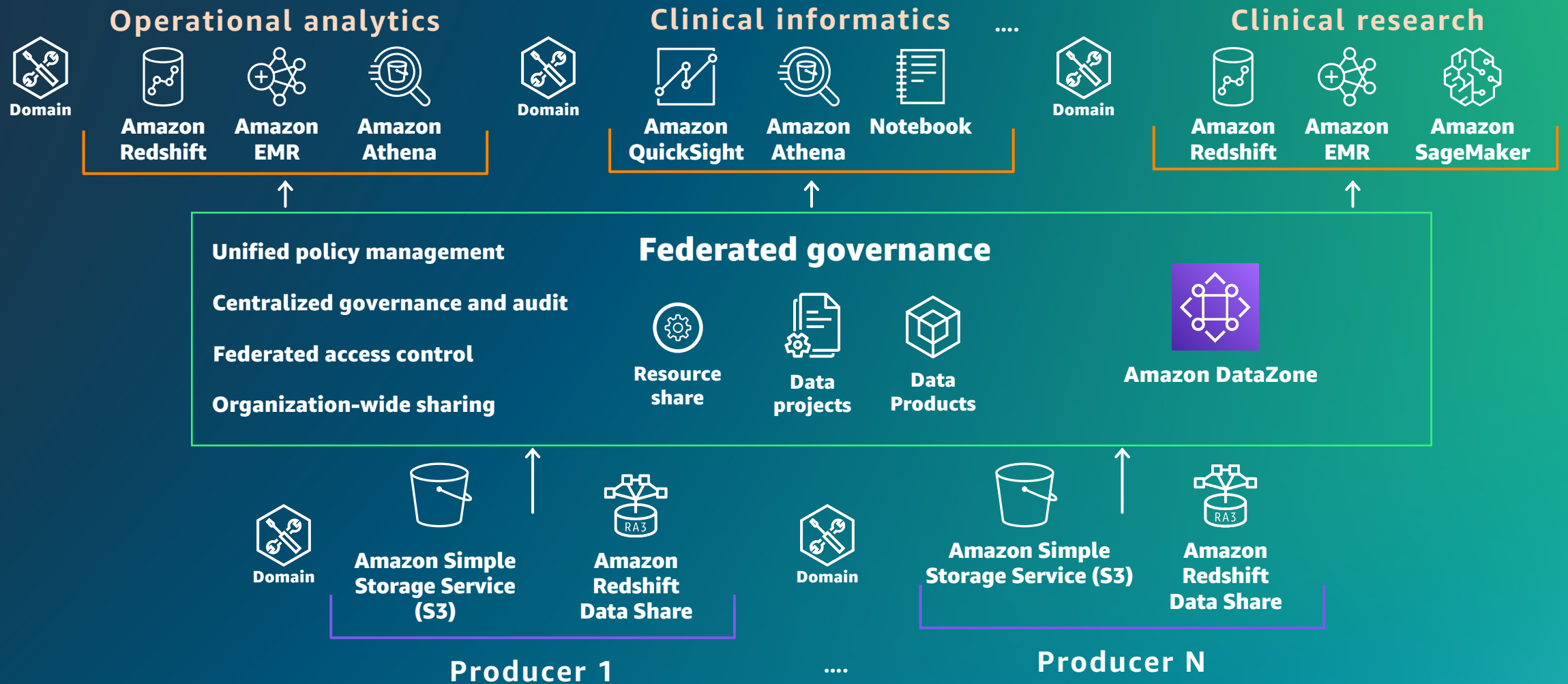
Recommendations you can tweak using out-of-the-box data quality rules and associate out-of-box actions to take when quality deteriorates

Multiple persona support

Data stewards can work on the data quality without worrying about the complexity, data engineers can easily integrate DQ in their data pipelines with the familiar APIs

Modern health data mesh architecture

Decentralized, lightweight federated governance across domain-oriented data systems to drive governed sharing



AWS Clean Rooms helps organizations collaborate on datasets without sharing underlying data

Multi-party collaborations

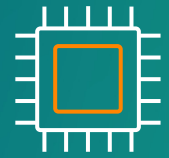
Collaborate with up to five parties in a single collaboration; extract insights from multiple companies

No AWS data movement

Use Amazon S3 data with direct permissioning and no AWS data movement

Query controls and enforcement

Configure analysis rules to restrict the type of analysis allowed on your data



Cryptographic computing

Pre-encrypt data so that it is encrypted at all times, including during query execution

Programmatic access

Automate and integrate functionality into existing workflows and products; create white-labeled clean room offering

AWS Registry of Open Data

AWS hosts a variety of public datasets that anyone can access for free

- 1000 Genomes Project
- The Cancer Genome Atlas
- International Cancer Genome Consortium
- 3000 Rice Genome
- Genome in a Bottle (GIAB)
- The Genome Modeling System
- Medicare Drug Spending
- The Human Connectome Project
- The Human Microbiome Project
- OpenNeuro
- Physionet
- Tabula muris
- gnoMAD

How to get started

Healthcare Data Maturity Assessment

Data-Driven Everything Program

AWS Immersion Day

Thank you!

Kiran Palsam
kpalsam@amazon.com

Vishanth Davidar
vdavidar@amazon.com

Track: Data and analytics
Session: Compliant research data architecture & data sharing management