Service Name

AWS HealthOmics

Service Overview

AWS HealthOmics is a HIPAA-eligible service that accelerates clinical diagnostic testing, drug discovery, and agricultural research by fully managing the complex infrastructure behind bioinformatics workflows. The service supports industry-standard workflow languages (WDL, Nextflow, CWL) and seamlessly scales bioinformatics infrastructure to support data from tens of thousands of tests per day with predictable cost-per-sample. HealthOmics consists of three main components: workflows for processing and analyzing genomics data, storage for genomics data and genome references, and analytics for transforming and analyzing genomics data using variant and annotation stores.

Key Use Cases

- Clinical Diagnostics: Build and scale diagnostic testing workflows with predictable costs and fully managed infrastructure that grows with testing volume, supporting variant calling, filtering, and annotation
- Drug Discovery: Accelerate therapeutic research by orchestrating biological foundation models at scale, enabling rapid iteration across millions of potential candidates using AI-powered analysis
- Agricultural Research: Enhance crop traits like drought tolerance and pest resistance through AI-powered workflows that improve food security and agricultural productivity using genomic analysis

Features

- **Workflows**: Process and analyze genomics data using WDL, Nextflow, or CWL with private and Ready2Run pre-built workflows
- **Storage**: Purpose-built storage for bioinformatics file formats with sequence stores for genomics data and reference stores for genome references
- **Analytics**: Transform and analyze genomics data using variant stores and annotation stores with population-scale data support
- **Auto-scaling**: Seamlessly scales across 100,000+ concurrent vCPUs to support tens of thousands of tests daily
- **Data Collaboration**: Secure sharing capabilities with fine-grained access controls, tagging, and permissions
- **HIPAA Compliance**: Built-in compliance features with comprehensive audit trails, data provenance tracking, and encryption

Value Proposition

AWS HealthOmics eliminates the complexity of managing bioinformatics infrastructure while providing predictable cost-per-sample pricing and enterprise-scale performance. The service handles technical complexities like managing compute resources and maintaining workflow engines, allowing organizations to focus entirely on scientific breakthroughs. Key differentiators include: fully managed infrastructure that scales to 100,000+ concurrent vCPUs, support for industry-standard workflow languages, HIPAA-eligible infrastructure with built-in compliance features, comprehensive data provenance and audit trails, seamless integration with AWS services ecosystem, and predictable pricing with no upfront costs.

Service Integration

AWS HealthOmics integrates extensively with core AWS services to provide a comprehensive bioinformatics platform. Primary integrations include Amazon S3 for data storage and input/output operations, Amazon Elastic Container Registry (ECR) for workflow container storage, AWS Lake Formation and Amazon Athena for analytics and querying genomics data through Apache Iceberg tables, Amazon SageMaker AI for machine learning workflows, Amazon EventBridge for event-driven architectures and workflow automation, AWS Lambda for custom processing, Amazon CloudWatch for monitoring and logging, and AWS IAM for access control and security management.

Onboarding and Offboarding

Getting started with AWS HealthOmics requires proper IAM permissions and roles for the service. Customers can begin by using Ready2Run workflows in the HealthOmics console, selecting pre-built workflows, entering run names and output locations, and creating service roles:

- The service supports Amazon Q CLI for natural language workflow management
- Setup involves creating AWS accounts, securing root users with MFA, creating administrative users with IAM Identity Center, configuring IAM permissions for HealthOmics, and optionally connecting with external code repositories through AWS CodeConnections for GitHub integration

Getting Started Guide: https://docs.aws.amazon.com/omics/latest/dev/getting-started.html

Data Removal

Data removal in AWS HealthOmics involves deleting resources through the console, CLI, or APIs. Customers can delete sequence stores, reference stores, workflows, variant stores, and annotation stores. Data in storage follows minimum retention periods of 30 days with prorated charges for early deletion. Cross-account sharing capabilities require

proper permission management during offboarding to ensure data access is properly revoked from shared accounts.

Backup/Restore and Disaster Recovery

AWS HealthOmics provides data durability through its purpose-built storage architecture with automatic data integrity verification using ETags. The service maintains data relationships for provenance tracking and supports cross-region capabilities. Data in ACTIVE state is readily accessible, while data not accessed for 30+ days moves to ARCHIVE state, requiring reactivation for access.

Backup Capabilities

AWS HealthOmics does not explicitly integrate with AWS Backup service. Instead, it provides built-in data durability through its storage architecture with automatic ETag generation for content integrity validation. The service preserves data relationships for provenance and supports data export capabilities to Amazon S3 for external backup strategies if needed.

Disaster Recovery

AWS HealthOmics does not directly integrate with AWS Elastic Disaster Recovery. The service provides multi-region availability and supports cross-region data operations. Disaster recovery is primarily achieved through the service's distributed architecture across AWS availability zones and the ability to export data to S3 for cross-region replication strategies.

Recovery Objectives

AWS HealthOmics supports recovery objectives through its multi-region availability, automatic data integrity verification, and export capabilities to S3. The service's distributed architecture helps achieve faster recovery times, while data export features enable customers to implement cross-region backup strategies to meet specific RPO requirements based on their genomics data criticality.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/healthomics-quotas.html

FAQ: https://aws.amazon.com/healthomics/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/omics/latest/dev/index.html

User Guide: https://docs.aws.amazon.com/omics/latest/dev/what-is-healthomics.html

API Reference: https://docs.aws.amazon.com/omics/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/healthomics/pricing/

Cost Optimization

AWS HealthOmics offers unique cost optimization through predictable cost-per-sample pricing and automated storage tiering. The service provides a Free Tier with 1500 gigabase-months in active/archive storage, 275 omics.m.xlarge instance hours, and 49,000 GB-hours of run storage. Cost optimization strategies include using archive storage for infrequently accessed data, leveraging Ready2Run workflows for predictable pricing versus private workflows, utilizing run caching to avoid recomputing identical tasks, implementing proper storage lifecycle management with 30-day minimum retention, and optimizing workflow resource requests to match actual computational needs.

Savings Considerations

Primary cost savings opportunities include leveraging the generous Free Tier allowances for initial workloads, using archive storage class for long-term data retention at reduced costs, implementing run caching strategies to eliminate redundant computations, choosing appropriate instance types and storage capacities based on actual workflow requirements, utilizing Ready2Run workflows which offer predictable per-run pricing versus variable private workflow costs, optimizing data formats and compression to reduce storage costs, and implementing proper data lifecycle policies to automatically transition data to lower-cost storage tiers while maintaining compliance with minimum retention requirements.