Service Name

Amazon Neptune

Service Overview

Amazon Neptune is a fast, reliable, fully managed graph database service that makes it easy to build and run applications that work with highly connected datasets. The core of Neptune is a purpose-built, high-performance graph database engine optimized for storing billions of relationships and querying the graph with milliseconds latency. Neptune supports popular graph query languages including Apache TinkerPop Gremlin, Neo4j's openCypher, and W3C's RDF query language SPARQL. It is highly available with read replicas, point-in-time recovery, continuous backup to Amazon S3, and replication across Availability Zones. Neptune eliminates database management tasks like hardware provisioning, software patching, setup, configuration, and backups.

Key Use Cases

- Recommendation engines: Building personalized content and product recommendation systems using graph relationships
- Fraud detection: Identifying fraudulent patterns and activities through complex relationship analysis in real-time
- Knowledge graphs: Creating structured repositories of information with defined concepts, properties, and relationships
- Identity graphs: Building unified customer views across multiple devices and identifiers for personalization and advertising
- Network security: Analyzing network relationships and dependencies to detect threats and vulnerabilities
- Drug discovery: Modeling molecular relationships and interactions for pharmaceutical research
- Social networking: Managing complex social relationships and connections between users

Features

- **Multi-model Graph Support**: Supports both Property Graph and W3C RDF graph models with query languages Gremlin, openCypher, and SPARQL
- **High Performance**: Optimized for storing billions of relationships with millisecond latency queries and up to 100,000 queries per second
- **Neptune Serverless**: Automatically provisions and scales database capacity based on workload requirements with pay-per-use pricing
- **Global Database**: Spans multiple AWS Regions with storage-based replication, sub-second latency, and disaster recovery capabilities

- **Multi-AZ Deployment**: Provides high availability with read replicas across Availability Zones and automatic failover
- **Neptune ML**: Integrates machine learning using graph neural networks for node classification, edge prediction, and link prediction
- **Continuous Backup**: Automatic, continuous, incremental backups with point-in-time recovery and database snapshots
- **Advanced Security**: Network isolation with VPC, encryption at rest and in transit, IAM integration, and compliance certifications
- **Neptune Analytics**: Memory-optimized analytics engine for analyzing large graph datasets with built-in algorithms
- **GraphRAG**: Combines graph databases with generative AI for retrieval-augmented generation capabilities

Value Proposition

Amazon Neptune offers compelling advantages as a fully managed graph database service. It eliminates operational overhead by handling database management tasks like provisioning, patching, and backups automatically:

- Neptune provides superior performance with millisecond latency queries and the ability to handle billions of relationships
- The service offers flexibility with support for multiple graph models and query languages (Gremlin, openCypher, SPARQL), allowing customers to choose the best approach for their use cases
- Neptune's serverless option enables cost optimization by automatically scaling resources and charging only for actual usage
- The service integrates seamlessly with the AWS ecosystem and includes advanced features like machine learning capabilities, global database support for multi-region deployments, and built-in security features
- This combination of managed operations, high performance, flexibility, and AWS integration makes Neptune an ideal choice over self-managed graph databases or alternative solutions

Service Integration

Neptune integrates extensively with core AWS services to provide a comprehensive graph database solution. It seamlessly connects with Amazon S3 for data storage, backup, and bulk data loading operations:

- Integration with AWS Lambda enables serverless functions to interact with Neptune for real-time processing and event-driven architectures
- Amazon SageMaker AI integration powers Neptune ML capabilities for graph machine learning workflows

- Neptune works with Amazon VPC for network isolation and security, AWS IAM for access management, and AWS KMS for encryption key management
- The service integrates with Amazon CloudWatch for monitoring and metrics, AWS CloudTrail for auditing, and AWS Backup for automated backup management
- Data migration capabilities are provided through AWS Database Migration Service (DMS), while Amazon Kinesis enables real-time data streaming into Neptune
- Additional integrations include AWS Glue for data preparation, Amazon Athena for querying via connectors, and AWS API Gateway for building graph-based APIs

Onboarding and Offboarding

Getting started with Neptune is straightforward through multiple pathways. Customers can begin with the free trial offering 750 hours of db.t3.medium or db.t4g.medium instance usage, 10 million I/O requests, and 1 GB each of storage and backup storage for eligible new users:

- The AWS Management Console provides an intuitive interface to create Neptune clusters, while AWS CLI and APIs offer programmatic access
- Neptune Serverless allows immediate scaling without capacity planning, making it ideal for getting started quickly
- The service includes comprehensive documentation, getting started guides, and AWS Graph Notebook for learning and development
- Customers can use sample datasets and reference architectures from the AWS GitHub repository
- Training resources include AWS Skill Builder courses, workshops, and video tutorials
- The Neptune Workbench provides Jupyter notebooks for interactive development and visualization, enabling users to learn graph concepts and query languages without production environment costs

Getting Started Guide: https://docs.aws.amazon.com/neptune/latest/userguide/graph-get-started.html

Data Removal

Neptune provides comprehensive data removal options for offboarding. Customers can delete individual DB clusters which removes all data, automated backups, and cluster resources, though manual snapshots are preserved unless explicitly deleted. Point-in-time recovery data is retained for the backup retention period (1-35 days) before automatic deletion. For complete offboarding, customers should delete manual snapshots, remove any Neptune Analytics graphs, clean up associated AWS resources like VPC endpoints and Lambda functions, and ensure proper cleanup of integrated services. The service supports selective data deletion through standard graph query operations before cluster

termination. Cross-region replicas and global database configurations require separate deletion steps in each region.

Backup/Restore and Disaster Recovery

Neptune provides robust backup and disaster recovery capabilities through multiple mechanisms. Automated continuous incremental backups occur with configurable retention periods from 1 to 35 days, enabling point-in-time recovery to any second within the retention window. Manual database snapshots can be created and retained indefinitely for long-term backup needs. Neptune Global Database offers disaster recovery across AWS Regions with RPO of 1 second and RTO under 1 minute. Cross-region snapshot copying enables additional disaster recovery options. The service supports Neptune-to-Neptune replication using streams for real-time backup cluster synchronization. All backup operations occur without performance impact on production workloads, and storage-based replication ensures data durability across multiple Availability Zones.

Backup Capabilities

Neptune includes comprehensive built-in backup capabilities with automatic continuous incremental backups enabled by default. The service maintains backup retention periods from 1 to 35 days and supports manual snapshots for extended retention:

- Neptune integrates with AWS Backup for centralized backup management across multiple AWS services, allowing customers to create backup plans, automate backup schedules, and manage retention policies through a unified console
- Cross-region backup copying is supported for disaster recovery scenarios, and all backup operations are designed to have no performance impact on running database workloads

Disaster Recovery

Neptune supports disaster recovery through multiple approaches but does not directly integrate with AWS Elastic Disaster Recovery service. The primary disaster recovery mechanisms include Neptune Global Database for cross-region replication with sub-second RPO and sub-minute RTO, Neptune streams-based replication to maintain synchronized backup clusters in different regions, and cross-region snapshot copying for recovery scenarios:

- Managed planned failover capabilities allow controlled region switching for maintenance operations, while unplanned failover supports emergency recovery situations
- These native Neptune disaster recovery features provide comprehensive protection without requiring AWS Elastic Disaster Recovery integration

Recovery Objectives

Neptune helps customers achieve aggressive RPO and RTO objectives through its disaster recovery features. Neptune Global Database delivers RPO of 1 second and RTO of less than 1 minute for cross-region scenarios using storage-based replication. For local high availability, Multi-AZ deployments with read replicas provide automatic failover typically within minutes. Point-in-time recovery enables restoration to any second within the backup retention period. Neptune streams replication can achieve RPO and RTO measured in minutes for custom disaster recovery implementations. The combination of continuous backups, automated failover, and global database capabilities allows customers to design solutions meeting strict availability and recovery requirements for mission-critical graph database workloads.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/neptune/latest/userguide/limits.html

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

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/neptune/

User Guide: https://docs.aws.amazon.com/neptune/latest/userguide/

API Reference: https://docs.aws.amazon.com/neptune/latest/apiref/

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/neptune/pricing/

Cost Optimization

Neptune offers several unique cost optimization options tailored for graph database workloads. Neptune Serverless provides automatic scaling with pay-per-use pricing, potentially reducing costs by up to 90% for variable workloads by charging only for active capacity:

- Customers can choose between Neptune Standard for typical I/O usage or Neptune I/O-Optimized for predictable pricing on I/O-intensive applications
- The service supports both on-demand and reserved instance pricing through Database Savings Plans with 1-year commitments
- AWS Graviton4 R8g instances offer up to 4.7 times better price-performance for write workloads compared to previous generations

- Neptune Analytics allows pausing and resuming graph workloads with reduced compute pricing when paused
- Query optimization through proper indexing and graph model design can significantly reduce I/O costs, while strategic use of read replicas can distribute load and optimize resource utilization

Savings Considerations

Key cost savings strategies for Neptune include leveraging Neptune Serverless for variable workloads to avoid over-provisioning and paying only for actual usage. Database Savings Plans provide significant discounts for predictable workloads with 1-year commitments:

- Upgrading to newer instance generations like Graviton4 R8g instances can deliver substantial price-performance improvements - up to 4.7 times better for write operations and 3.7 times better for read operations
- Right-sizing instances based on actual performance metrics like MainRequestQueuePendingRequests and BufferCacheHitRatio prevents over-provisioning
- Optimizing graph data models and queries reduces I/O transfer costs, while storing large objects outside Neptune in S3 can minimize storage expenses
- Strategic use of read replicas for load distribution and choosing appropriate backup retention periods help manage overall costs
- For development environments, utilizing the free trial and properly scheduling instance start/stop operations can provide additional savings