# Service Name

Amazon Managed Service for Prometheus

## Service Overview

Amazon Managed Service for Prometheus is a serverless, Prometheus-compatible monitoring service that makes it easier to securely monitor container environments at scale. It automatically scales the ingestion, storage, and querying of operational metrics as workloads scale up and down, while integrating with AWS security services to enable fast and secure access to data. Built on the open-source Prometheus data model and powered by Cortex, it provides the same PromQL query language and data model used with traditional Prometheus, but without the operational overhead of managing underlying infrastructure. The service is designed for high availability using multi-AZ deployments and replicates data across three Availability Zones.

### Key Use Cases

- Container monitoring: Monitor performance and operational metrics from Amazon EKS, Amazon ECS, AWS Fargate, and self-managed Kubernetes clusters at scale
- Application observability: Collect, store, and analyze time-series metrics from containerized applications using familiar Prometheus tools and PromQL queries
- Multi-environment monitoring: Unified monitoring solution across AWS, on-premises, hybrid, and multi-cloud container environments with centralized metric storage and querying

## Features

- **Serverless monitoring**: Fully managed service with automatic scaling for ingestion, storage, and querying without infrastructure management
- **Prometheus compatibility**: Native support for Prometheus data model, PromQL query language, and over 150+ Prometheus exporters
- **AWS managed collectors**: Agentless metric collection that automatically discovers and pulls metrics from Kubernetes clusters through ENIs
- **Alert Manager**: Built-in alerting with deduplication, grouping, routing to SNS topics, and integration with PagerDuty
- **High availability**: Multi-AZ deployments with data replication across three Availability Zones and automatic failover
- **Security integration**: Encryption in transit and at rest, IAM access control, AWS KMS integration, and VPC endpoint support
- **Query insights and control**: Query logging to CloudWatch, cost control with QSP thresholds, and expensive query monitoring

## Value Proposition

Amazon Managed Service for Prometheus eliminates the operational overhead of managing Prometheus infrastructure while providing enterprise-grade scalability, security, and availability. Customers benefit from automatic scaling that handles up to 1 billion active series per workspace, native integration with AWS security services, and seamless compatibility with existing Prometheus tooling:

- The service offers significant cost advantages through efficient resource utilization and pay-as-you-go pricing, while enabling teams to focus on application development rather than monitoring infrastructure maintenance
- Integration with Amazon Managed Grafana provides a complete managed observability solution, and the service's multi-AZ architecture ensures high availability without additional configuration effort

### Service Integration

Amazon Managed Service for Prometheus deeply integrates with Amazon EKS and Amazon ECS for container metrics collection, Amazon Managed Grafana for visualization and dashboards, and AWS Distro for OpenTelemetry (ADOT) for standardized metric collection. It leverages Amazon CloudWatch for usage metrics monitoring and alerting, Amazon SNS for alert notifications, AWS IAM for access control, and AWS KMS for encryption key management:

- The service supports VPC endpoints for private network access and integrates with AWS CloudTrail for API call logging
- Additional integrations include Amazon MSK for Kafka monitoring, AWS Lambda for custom processing, and AWS Cost and Usage Reports for billing analysis

## Onboarding and Offboarding

Getting started requires three key steps: creating a workspace to store metrics, ingesting metrics either manually or through automatic scraping, and querying the data using PromQL. Customers can create workspaces through the AWS Management Console, AWS CLI, or SDKs, then configure metric ingestion using AWS managed collectors for EKS clusters or custom Prometheus agents:

- The service supports both agentless collection through managed scrapers and traditional Prometheus remote write configurations
- Integration with Amazon Managed Grafana provides immediate visualization capabilities, while alert rules can be configured for proactive monitoring
- AWS provides comprehensive documentation, getting started guides, and example configurations to streamline the onboarding process

Getting Started Guide: https://docs.aws.amazon.com/prometheus/latest/userguide/AMP-getting-started.html

### Data Removal

Data removal occurs through workspace deletion, which permanently removes all stored metrics and metadata. Customers must delete any active managed scrapers before deleting workspaces to prevent continued metric ingestion and charges. Data is automatically deleted after the retention period (150 days by default, configurable up to 1095 days). The service provides configurable retention policies per workspace, and customers can monitor usage through CloudWatch metrics to ensure complete data removal after offboarding.

## Backup/Restore and Disaster Recovery

Amazon Managed Service for Prometheus provides built-in high availability through multi-AZ deployments with automatic data replication across three Availability Zones. The service includes automatic patch management and infrastructure maintenance. For disaster recovery, customers can implement cross-region data replication by configuring multiple workspaces in different regions and using Prometheus remote write to send metrics to secondary workspaces, ensuring data availability even when primary regions are unavailable.

### Backup Capabilities

The service does not integrate directly with AWS Backup but provides data resilience through automatic replication across multiple Availability Zones within a region. Data durability is ensured through the underlying AWS infrastructure, and customers can implement backup strategies by replicating metrics to secondary workspaces in different regions or using Prometheus remote write configurations to maintain copies of critical metrics data.

### Disaster Recovery

Amazon Managed Service for Prometheus does not integrate with AWS Elastic Disaster Recovery but provides built-in disaster recovery capabilities through multi-AZ architecture and cross-region workspace replication. The service automatically handles infrastructure failures and provides high availability within regions. For comprehensive disaster recovery, customers can configure secondary workspaces in different AWS regions and implement cross-region metric replication using Prometheus remote write functionality.

### Recovery Objectives

The service helps meet RPO and RTO objectives through automatic multi-AZ replication providing near-zero RPO within regions, and sub-minute RTO for automatic failover scenarios. Cross-region workspace configurations can achieve broader disaster recovery objectives with RPO dependent on metric ingestion frequency and RTO based on DNS failover or application-level switching between regions. The serverless architecture enables rapid recovery without infrastructure provisioning delays.

## Service Constraints

Service Quotas: https://docs.aws.amazon.com/prometheus/latest/userguide/AMP_quotas.html

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

## Technical Requirements

Service Documentation: https://docs.aws.amazon.com/prometheus/latest/userguide/what-is-Amazon-Managed-Service-Prometheus.html

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

API Reference: https://docs.aws.amazon.com/prometheus/latest/APIReference/

## 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/prometheus/pricing/

### Cost Optimization

Amazon Managed Service for Prometheus offers several unique cost optimization features including label-based active series limits to control metric volume by producer, query insights and control with QSP thresholds to manage expensive queries, and tiered pricing that scales with usage. The service provides CloudWatch usage metrics for monitoring ingestion rates and active series, enabling proactive cost management:

- Customers can optimize costs by reducing collection frequency, limiting active series through metric filtering, using native alerting instead of external systems, and implementing query optimization techniques
- The pay-as-you-go model eliminates infrastructure overhead costs while automatic scaling prevents over-provisioning

### Savings Considerations

Primary cost savings come from eliminating self-managed Prometheus infrastructure, reducing operational overhead, and leveraging automatic scaling to match actual usage. The service's tiered pricing model provides cost efficiency at scale, while managed collectors reduce agent maintenance costs:

- Organizations can achieve significant savings through reduced engineering time for monitoring infrastructure, elimination of storage and compute costs for self-hosted solutions, and optimized query performance

- AWS Free Tier provides 40 million samples for ingestion, storage, and querying monthly, enabling cost-effective evaluation and small-scale deployments
- Cross-region replication costs should be considered for disaster recovery implementations, balanced against the operational savings from managed infrastructure