

# Service Name

AWS Lake Formation

## Service Overview

AWS Lake Formation is a fully managed service that simplifies building, securing, and managing data lakes on AWS. It provides centralized governance, security, and sharing capabilities for analytics and machine learning workloads by allowing fine-grained access control for data stored in Amazon S3 and metadata in AWS Glue Data Catalog. Lake Formation uses its own permissions model that complements IAM permissions, enabling secure access to data lakes and external data sources. The service automates complex data lake setup tasks including data ingestion, cataloging, transformation, and security management while providing comprehensive audit logging and compliance capabilities.

## Key Use Cases

- **Data Lake Governance:** Centrally govern and secure enterprise data lakes with fine-grained access control at database, table, column, row, and cell levels
- **Cross-Account Data Sharing:** Securely share data internally and externally across AWS accounts with tag-based and named resource access controls
- **Analytics and Machine Learning:** Enable secure data access for analytics services like Amazon Athena, Amazon Redshift, Amazon EMR, and Amazon SageMaker
- **Compliance and Audit:** Implement comprehensive data access monitoring, audit logging, and regulatory compliance for sensitive data
- **Data Cataloging:** Automatically discover, catalog, and label data using AWS Glue crawlers with metadata management capabilities

## Features

- **Fine-grained Access Control:** Database, table, column, row, and cell-level permissions with tag-based and attribute-based access control
- **Cross-Account Data Sharing:** Secure data sharing across AWS accounts and external organizations with centralized governance
- **Data Catalog Management:** Centralized metadata management using AWS Glue Data Catalog with automatic schema discovery
- **Audit and Compliance:** Comprehensive data access logging, monitoring, and audit capabilities for regulatory compliance
- **Blueprint Workflows:** Pre-built ETL workflows for database snapshots, incremental updates, and log file processing
- **Hybrid Access Mode:** Selective onboarding of Lake Formation permissions alongside existing IAM controls

## Value Proposition

AWS Lake Formation eliminates the complexity of building and managing data lakes by providing a fully managed service with centralized governance. Unlike traditional approaches requiring manual setup of permissions, cataloging, and security, Lake Formation automates these processes while providing enterprise-grade security features:

- Key advantages include fine-grained access control without complex IAM policies, seamless integration with AWS analytics services, simplified cross-account data sharing, and comprehensive audit capabilities
- The service reduces operational overhead, accelerates time-to-value for data initiatives, and ensures consistent security policies across the organization while maintaining compatibility with existing AWS services and third-party tools

## Service Integration

Lake Formation deeply integrates with AWS Glue and AWS Glue Data Catalog as its foundation for metadata management and ETL processing. It seamlessly works with analytics services including Amazon Athena, Amazon Redshift, Amazon EMR, and Amazon SageMaker for data access and processing:

- The service integrates with Amazon S3 for data storage and uses IAM for identity management
- Lake Formation also connects with Amazon Kinesis Data Firehose for data ingestion, AWS CloudTrail for audit logging, and supports external data sources through JDBC connections
- Additionally, it integrates with AWS Data Exchange for data sharing and supports third-party business intelligence tools and analytics platforms

## Onboarding and Offboarding

Customers begin by setting up Lake Formation through the AWS Management Console or CloudFormation templates, starting with creating a data lake administrator and configuring the default permission model. The process involves registering Amazon S3 locations as data lake storage, creating databases in the Data Catalog, and setting up user permissions:

- Customers can use pre-built blueprints to create ETL workflows for data ingestion from JDBC sources or use AWS Glue crawlers for schema discovery
- Lake Formation provides tutorials for creating data lakes from CloudTrail logs or database sources, with step-by-step guidance for configuring security policies, cross-account sharing, and integrating with analytics services

Getting Started Guide: <https://docs.aws.amazon.com/lake-formation/latest/dg/getting-started-setup.html>

## Data Removal

Lake Formation data removal involves deleting Data Catalog tables and databases, revoking all permissions granted to users and roles, and unregistering S3 locations from Lake Formation management. Customers must remove cross-account sharing configurations and delete any custom IAM roles created for Lake Formation workflows. Since Lake Formation manages metadata and permissions rather than storing actual data, the underlying S3 data remains intact unless explicitly deleted. Organizations should audit and revoke all Lake Formation permissions before decommissioning to ensure complete access removal.

## Backup/Restore and Disaster Recovery

Lake Formation relies on the underlying AWS services for backup and disaster recovery capabilities. Data stored in Amazon S3 benefits from S3's durability and availability features, while metadata in AWS Glue Data Catalog is automatically replicated within the region. Lake Formation permissions and configurations can be backed up using AWS Config or exported via APIs for disaster recovery planning. The service supports cross-region data access but does not provide native backup functionality, requiring customers to implement backup strategies for S3 data and Data Catalog metadata.

### Backup Capabilities

Lake Formation does not provide native backup capabilities and does not integrate with AWS Backup service. Backup strategies must be implemented at the underlying service level, primarily for S3 data using S3 Cross-Region Replication, versioning, or third-party solutions:

- Data Catalog metadata can be exported using AWS Glue APIs for backup purposes
- Customers should implement comprehensive backup strategies for both data and metadata components independently

### Disaster Recovery

Lake Formation does not directly support AWS Elastic Disaster Recovery service integration. Disaster recovery must be planned at the infrastructure level, focusing on S3 data replication and Data Catalog reconstruction:

- The service supports cross-region data access, enabling disaster recovery architectures where data can be accessed from multiple regions
- Organizations must design disaster recovery strategies that encompass S3 replication, IAM role recreation, and Lake Formation permission reconfiguration in recovery regions

## Recovery Objectives

Lake Formation can support customer RPO and RTO objectives through its integration with highly durable AWS services and cross-region capabilities. Since metadata is stored in AWS Glue Data Catalog with automatic replication, metadata RPO is typically measured in minutes. Data RPO depends on S3 replication strategies implemented by customers. RTO can be minimized through cross-region data access capabilities and automated infrastructure deployment using CloudFormation templates for permission and configuration recreation.

## Service Constraints

Service Quotas: [https://docs.aws.amazon.com/general/latest/gr/lake-formation.html#limits\\_lake-formation](https://docs.aws.amazon.com/general/latest/gr/lake-formation.html#limits_lake-formation)

FAQ: <https://aws.amazon.com/lake-formation/faqs/>

## Technical Requirements

Service Documentation: <https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html>

User Guide: <https://docs.aws.amazon.com/lake-formation/latest/dg/getting-started-tutorials.html>

API Reference: <https://docs.aws.amazon.com/lake-formation/latest/dg/aws-lake-formation-api.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/lake-formation/pricing/>

## Cost Optimization

Lake Formation offers several unique cost optimization opportunities including no additional charges for basic permissions management and data governance features. The service provides significant cost savings through its Storage API pricing model where customers only pay for bytes scanned, encouraging the use of columnar formats like Parquet and ORC to reduce scanning costs:

- Governed Tables feature optimizes storage layout automatically to improve query performance and reduce costs
- The service eliminates the need for separate data governance tools and reduces operational costs through automation

- Additionally, Lake Formation's ability to enable secure data sharing can reduce data duplication across organizations

## Savings Considerations

Primary cost savings include elimination of third-party data governance tool licensing fees and reduced operational overhead through automated data lake management. The service's integration with serverless AWS analytics services enables pay-per-query pricing models, avoiding fixed infrastructure costs:

- Storage optimization through automatic compaction and columnar format recommendations reduces both storage and compute costs
- Cross-account data sharing capabilities eliminate data duplication costs and reduce data transfer charges
- The unified security model reduces administrative overhead and compliance costs
- Organizations can achieve significant TCO reduction by consolidating data governance, reducing manual security management, and leveraging automated optimization features