

Service Name

AWS Entity Resolution

Service Overview

AWS Entity Resolution is a fully-managed service that helps organizations match, link, and enhance related records stored across multiple applications, channels, and data stores. The service enables users to create unified views of customer interactions, product information, and business data by using advanced matching techniques including rule-based matching, machine learning-based matching, and data service provider-led matching. It processes records containing identifiers such as names, email addresses, phone numbers, and addresses to accurately link related information without requiring expertise in entity resolution or machine learning. The service supports both batch processing for large datasets and near real-time matching for time-sensitive use cases.

Key Use Cases

- **Customer Identity Resolution:** Creating unified customer profiles by matching records across CRM systems, marketing platforms, and customer data platforms to eliminate duplicate identities and fragmented customer views
- **Healthcare Patient Matching:** Building unified patient indexes by linking patient records across multiple healthcare systems and providers to ensure accurate patient identification and comprehensive medical histories
- **Product Catalog Management:** Matching and linking product information across different systems using various codes (SKU, UPC) to maintain consistent product catalogs and inventory management
- **Fraud Detection and Financial Services:** Identifying duplicate accounts, verifying customer identities, and detecting fraudulent activities by matching customer records across financial institutions and transaction systems
- **Marketing and Advertising Optimization:** Enhancing customer targeting and personalization by linking consumer interactions across multiple touchpoints and channels for more effective cross-channel campaigns

Features

- **Flexible Data Preparation:** Supports up to 20 data inputs from AWS Glue with customizable schema mapping, data normalization capabilities, and encryption support for secure data processing
- **Rule-Based Matching:** Ready-to-use deterministic matching rules with advanced fuzzy matching algorithms including Levenshtein Distance, Cosine Similarity, and Soundex for handling data variations and typos

- **Machine Learning Matching:** Pre-configured ML models that analyze entire records together, handle missing fields, and provide confidence scores for match quality assessment
- **Data Service Provider Integration:** Integration with licensed datasets and third-party data service providers for data enrichment and enhanced matching capabilities
- **Near Real-Time Matching:** GetMatchId and GenerateMatchId APIs for synchronous lookup and matching of entities within seconds for time-sensitive applications
- **Automated Processing:** Support for both manual bulk processing and automatic incremental processing to keep matches updated as new data arrives in S3
- **Data Protection:** Built-in encryption with AWS KMS support, data regionalization, and secure data handling with hashing capabilities for PII protection

Value Proposition

AWS Entity Resolution provides a fully-managed, no-code solution that eliminates the complexity of building custom entity matching systems. Organizations choose this service because it requires no expertise in entity resolution or machine learning, offers flexible matching approaches from deterministic rules to advanced ML models, and integrates seamlessly with existing AWS data infrastructure:

- The service delivers significant time-to-value by providing pre-configured matching techniques while maintaining the flexibility to customize rules and workflows
- Key differentiators include advanced fuzzy matching algorithms for handling real-world data imperfections, near real-time matching capabilities for time-sensitive applications, built-in security and compliance features, and the ability to process datasets ranging from thousands to billions of records with automatic scaling

Service Integration

AWS Entity Resolution integrates deeply with core AWS data and analytics services to provide a comprehensive data processing ecosystem. Primary integrations include Amazon S3 for data input and output storage, AWS Glue for data cataloging and schema management, and AWS Glue Data Catalog for metadata management:

- Security integrations include AWS KMS for encryption key management, AWS IAM for access control, and AWS CloudTrail for audit logging
- Monitoring and observability are supported through Amazon CloudWatch for metrics and logging, Amazon CloudWatch Logs for workflow monitoring, and Amazon Data Firehose for log delivery
- The service also integrates with Amazon Connect Customer Profiles for unified customer views, AWS Step Functions for workflow orchestration, AWS Lambda for custom processing, and Amazon EventBridge for event-driven architectures

Onboarding and Offboarding

Getting started with AWS Entity Resolution follows a straightforward process. Customers first prepare their data in Amazon S3 and use AWS Glue crawlers to create data catalog entries:

- Next, they create schema mappings in the Entity Resolution console to define how their data fields should be interpreted for matching
- The third step involves creating matching workflows by selecting matching techniques (rule-based, ML-based, or provider service-based) and configuring output destinations
- Customers can then run workflows either manually for batch processing or set up automatic incremental processing
- The console provides guided setup with pre-built templates and ready-to-use rules
- For API access, comprehensive SDKs are available for Java, Python, JavaScript, and other languages
- AWS provides sample data, tutorials, and interactive scenarios to help users understand service capabilities before processing production data

Getting Started Guide:

<https://docs.aws.amazon.com/entityresolution/latest/userguide/what-is-service.html>

Data Removal

AWS Entity Resolution follows AWS standard data handling practices for offboarding. Customers maintain full control over their data as the service processes data from customer-controlled S3 buckets and writes results back to customer-specified locations. To remove data, customers delete their S3 input and output buckets, delete Entity Resolution workflows and schema mappings, and optionally delete associated AWS Glue Data Catalog entries. The service does not retain customer data beyond the processing period, and all intermediate processing data is automatically cleaned up after job completion. Customers can use AWS CloudTrail to audit all data access and deletion activities.

Backup/Restore and Disaster Recovery

AWS Entity Resolution does not provide built-in backup and disaster recovery capabilities as it is a processing service that operates on customer data stored in S3. Data durability and backup responsibilities lie with the underlying storage services customers use. AWS Entity Resolution workflows and configurations can be recreated using Infrastructure as Code tools like AWS CloudFormation or programmatically via APIs. Customers should implement their own backup strategies for S3 data and maintain copies of workflow configurations for disaster recovery purposes.

Backup Capabilities

AWS Entity Resolution does not have native backup capabilities and does not integrate directly with AWS Backup. The service processes data from customer S3 buckets and writes results back to customer-specified S3 locations:

- Customers are responsible for implementing backup strategies for their input data, output results, and workflow configurations
- Backup protection should be implemented at the S3 bucket level using S3 versioning, S3 replication, or AWS Backup policies applied to S3 resources

Disaster Recovery

AWS Entity Resolution does not directly integrate with AWS Elastic Disaster Recovery as it is a serverless processing service without persistent infrastructure. The service automatically runs across multiple Availability Zones within a region for high availability:

- For multi-region disaster recovery, customers must implement cross-region replication of their S3 data and recreate Entity Resolution workflows in secondary regions using Infrastructure as Code or API automation
- Workflow configurations and schema mappings can be exported and replicated across regions programmatically

Recovery Objectives

AWS Entity Resolution supports customer RPO and RTO objectives through its serverless, highly available architecture and flexible processing options. For RPO, customers can achieve low recovery point objectives by implementing cross-region S3 replication and maintaining current workflow configurations in multiple regions. For RTO, the service's serverless nature enables quick recovery as workflows can be rapidly recreated and executed without infrastructure provisioning delays. Near real-time matching capabilities support sub-second RTO requirements for critical matching operations.

Service Constraints

Service Quotas:

<https://docs.aws.amazon.com/entityresolution/latest/userguide/quotas.html>

FAQ: <https://aws.amazon.com/entity-resolution/faqs/>

Technical Requirements

Service Documentation: <https://docs.aws.amazon.com/entityresolution/latest/userguide/>

User Guide: <https://docs.aws.amazon.com/entityresolution/latest/userguide/what-is-service.html>

API Reference: <https://docs.aws.amazon.com/entityresolution/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/entity-resolution/pricing/>

Cost Optimization

AWS Entity Resolution offers several cost optimization strategies due to its pay-per-use pricing model based on records processed. Customers can optimize costs by implementing data preprocessing to reduce record volumes, using rule-based matching instead of ML-based matching when appropriate (\$0.50 vs \$0.10 per 1,000 records), and leveraging incremental processing to avoid reprocessing unchanged data:

- Additional optimizations include using data service provider-based matching for specific use cases, optimizing S3 storage classes for input and output data, implementing lifecycle policies for temporary processing data, and right-sizing processing jobs to avoid unnecessary resource consumption
- Regular monitoring of processing volumes and match rates helps identify opportunities for rule optimization and cost reduction

Savings Considerations

Primary cost savings come from the service's serverless architecture eliminating infrastructure management overhead and the pay-per-use model avoiding upfront costs. Customers save significantly compared to building custom entity resolution solutions by leveraging pre-built ML models, avoiding machine learning expertise requirements, and eliminating development and maintenance costs:

- Operational savings include reduced data engineering effort through automated schema mapping, decreased time-to-value with ready-to-use matching rules, and improved accuracy reducing downstream data cleanup costs
- Scale-based savings occur through automatic optimization of processing resources and the ability to process large datasets efficiently without manual tuning
- Organizations also benefit from reduced compliance and security implementation costs through built-in AWS security features and audit capabilities