データレイクソリューション

AWS 実装ガイド

2016 年 11 月

最終更新日 2019 年 6 月 (改訂を参照)

Copyright (c) 2019 by Amazon.com, Inc. or its affiliates.

データレイクソリューションは、https://aws.amazon.com/asl/
で閲覧可能な Amazon ソフトウェアライセンスの条項に基づいてライセンスされます

目次

このガイドについて

この実装ガイドでは、アマゾン ウェブ サービス (AWS) クラウドにデータレイクソリューションをデプロイするためのアーキテクチャ上の考慮事項と設定手順について説明します。セキュリティと可用性に関する AWS ベストプラクティスを使用して、このソリューションを AWS にデプロイするために必要な AWS のサービスを起動、設定、実行する AWS CloudFormation テンプレートへのリンクが含まれています。

このガイドは、AWS クラウドでのアーキテクチャ設計の実務経験がある IT インフラストラクチャアーキテクト、管理者、DevOps プロフェッショナルを対象としています。

概要

アマゾン ウェブ サービス (AWS) のお客様の多くは、従来のデータ管理システムと比べてさらに優れた俊敏性と柔軟性を備えたデータストレージと分析ソリューションを必要としています。データレイクでは、企業が構造化および非構造化のすべてのデータを一元化されたリポジトリに保存できるため、データを保存および分析する際の一般的な方法となってきています。効果的なデータレイクは、低コストでスケーラブルで安全なストレージを提供し、さまざまなデータ型の検索および分析機能をサポートする必要があります。

AWS クラウドは、お客様が安全で柔軟があり、コスト効率の高いデータレイクを実装するのに必要な多くの構成要素を提供します。データレイクを構築するお客様をサポートするために、AWS ではデータレイクソリューションを提供しています。データレイクソリューションは、データセットの検索やリクエストに使いやすいコンソールとともに、可用性が高くコスト効率の高いデータレイクアーキテクチャを AWS クラウドにデプロイする自動化されたリファレンスを実装します。このソリューションは、データレイクアーキテクチャの概念化、データの変換と分析に関するお客様の一般的な問題点に対処することを目的としています。このソリューションは、企業全体またはその他の外部ユーザーとデータの特定のサブセットを簡単にタグ付け、検索、共有、管理するために必要な、主要な AWS のサービスを自動的に設定します。このソリューションにより、ユーザーは最小限の労力で、新しいデータセットのカタログ化、検索可能なメタデータを使用したデータセットのアップロード、Amazon Simple Storage Service (AmazonS3) の既存のデータセットのデータプロファイルの作成を行うことができます。

データレイクソリューションは、あらゆるサイズのデータセットをネイティブ形式で安全かつ耐久性のある拡張性の高い Amazon S3 に保存および登録します。お客様は、検索可能なメタデータを使用してデータセットをアップロードし、AWS Glue や Amazon Athena と統合して、そのデータを変換および分析できます。

このソリューションでは、データソースを自動的にクロールし、データフォーマットを識別してからスキーマと変換を提案するため、時間をかけてデータフローを手作業でコーディングする必要がなくなります。さらに、ユーザー定義タグは Amazon DynamoDB に保存され、ビジネス関連のコンテキストを各データセットに追加します。このソリューションにより、企業はデータセットがデータレイクに登録された場合に特定のタグを要求するシンプルなガバナンスポリシーを作成できます。ユーザーは、利用可能なデータセットを参照したり、データセットの属性とタグを検索したりして、ビジネスニーズに関連するデータをすばやく見つけてアクセスできます。

さらに、データレイクソリューションにはフェデレーティッドテンプレートが含まれており、Microsoft Active Directory や Okta などの既存の SAML ID プロバイダーと統合できるバージョンのソリューションを起動できます。Active Directory の詳細については付録 A を、Okta の詳細については付録 B を参照してください。

コスト

データレイクソリューションの実行中に使用した AWS サービスのコストは、お客様が負担します。このソリューションを実行するための合計コストは、ロード、リクエスト、保存、処理、および提示されるデータの量によって異なります。詳細については、このソリューションで使用する各 AWS サービスの料金表ウェブページを参照してください。

アーキテクチャの概要

このソリューションをデプロイすると、AWS クラウドに以下の環境が構築されます。

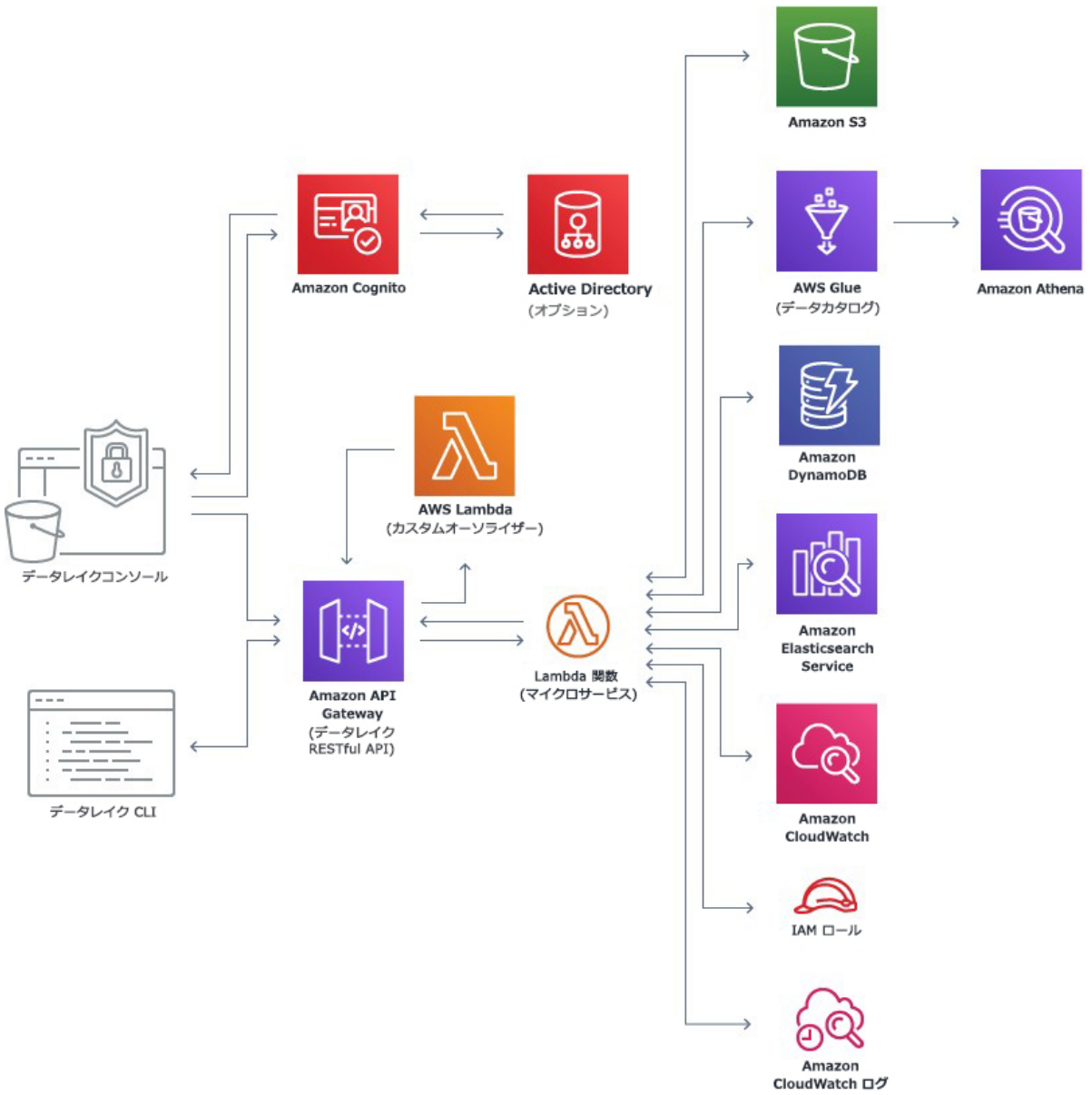


図 1: AWS でのデータレイクソリューションアーキテクチャ

このソリューションでは、AWS CloudFormation を使用して、このデータレイクリファレンス実装をサポートするインフラストラクチャコンポーネントをデプロイします。このソリューションのコアとなるのはデータレイク API の実装です。これは Amazon API Gateway を活用し、データレイクマイクロサービス (AWS Lambda 関数) へのアクセスを提供します。これらのマイクロサービスは、データパッケージの作成、データのアップロード、

既存パッケージの検索、興味深いデータのカートへの追加、データマニフェストの生成、管理機能の実行を行うためのビジネスロジックを提供します。これらのマイクロサービスは、Amazon S3、AWS Glue、Amazon Athena、Amazon DynamoDB、Amazon Elasticsearch Service (Amazon ES)、および Amazon CloudWatch Logs と連携することにより、データストレージ、管理、監査機能を提供します。

このソリューションは、データレイクコンソールを作成し、静的ウェブサイトホスティング用に構成された Amazon S3 バケットにデプロイし、ソリューションのコンソールエントリポイントとして使用される Amazon CloudFront ディストリビューションを設定します。このソリューションでは、初期設定時にデフォルトの管理者ロールも作成され、お客様が指定した E メールアドレスにアクセス招待が送信されます。フェデレーティッドスタックをデプロイする場合は、ユーザーグループと管理者グループを手動で作成する必要があることに注意してください。Active Directory についての詳細は、付録 A を参照してください。Okta についての詳細は、付録 B を参照してください。

このソリューションは、Amazon Cognito ユーザープールを使用して、コンソールおよびデータレイク API へのユーザーアクセスを管理します。各ソリューションのコンポーネントについての詳細は、付録 C を参照してください。

ソリューションの特徴

このデータレイクソリューションには、以下の機能があります。

- **データレイクリファレンス実装:** このデータレイクソリューションは、そのまま使用することもできるほか、独自のデータ管理、検索、処理のニーズに合わせてカスタマイズできるリファレンス実装として活用することもできます。
- **ユーザーインターフェイス:** このソリューションは Amazon S3 でホストされ、Amazon CloudFront から配信される直感的なウェブベースのコンソール UI を自動的に作成します。コンソールにアクセスすると、データレイクユーザー、データレイクポリシーの管理、データパッケージの追加と削除、データパッケージの検索、追加の分析のためのデータセットのマニフェストの作成を簡単に行えます。
- **コマンドラインインターフェイス (CLI):** 提供された CLI または API を使用して、データレイクのアクティビティを簡単に自動化するか、このソリューションを既存のデータオートメーションに統合して、データセットの入力、出力、および分析を行います。

- **マネージド型ストレージレイヤー:** マネージド型の Amazon S3 バケット内のデータのストレージと取得を保護および管理し、ソリューション固有の AWS Key Management Service (AWS KMS) キーを使用して保管時のデータを暗号化します。
- **データアクセスの柔軟性:** 署名済みの Amazon S3 URL を活用するか、適切な AWS Identity and Access Management (IAM) ロールを使用して、Amazon S3 のデータセットへの制御された直接的なアクセスを行います。
- **データ変換と分析:** 検索可能なメタデータを使用してデータセットをアップロードし、AWS Glue や Amazon Athena と統合して、そのデータを変換および分析します。
- **フェデレーションサインイン:** オプションで、ユーザーが Microsoft Active Directory フェデレーションサービス (AD FS) や Okta などの SAML ID プロバイダー (IdP) を介してサインインできるようにすることができます。

AWS CloudFormation テンプレート

このソリューションでは、AWS CloudFormation を使用して、AWS クラウドへのデータレイクソリューションのデプロイを自動化します。ソリューションには以下の AWS CloudFormation テンプレートが含まれており、デプロイ前にダウンロード可能です。

テンプレートを表示 **data-lake-deploy.template:** このテンプレートを使用して、データレイクソリューションとすべての関連コンポーネントを起動します。デフォルトの設定では、組み込みの認証、許可、およびユーザー/グループ管理がデプロイされます。また、特定のニーズに基づいてテンプレートをカスタマイズすることもできます。このテンプレートは、以下次のネストされたスタックを起動します。

- **data-lake-storage.template:** このテンプレートは、ソリューションの Amazon S3、Amazon ES、および Amazon DynamoDB コンポーネントをデプロイします。
- **data-lake-services.template:** このテンプレートは、AWS Lambda マイクロサービスと必要な IAM ロールおよびポリシーをデプロイします。さらに、ソリューションの AWS KMS リソースをデプロイします。
- **data-lake-api.template:** このテンプレートは、Amazon API Gateway リソースをデプロイします。

自動デプロイ

自動デプロイを開始する前に、このガイドで説明されているアーキテクチャ、設定、ネットワークセキュリティ、その他の考慮事項をよくお読みください。このセクションの手順に従って、データレイクソリューションを設定してアカウントにデプロイします。

デプロイ時間: 約 30 分

ここでカバーする内容

このアーキテクチャを AWS にデプロイする手順は、以下のステップで構成されます。詳細な手順については、各ステップのリンクを参照してください。

ステップ 1.スタックを起動する

- AWS アカウントで AWS CloudFormation テンプレートを起動します。
- 必須パラメータの値 **[Stack Name]**、[**Administrator Name**]、[**Administrator Email**]、[**Cognito Domain**] を入力します。
- その他のテンプレートのパラメータを確認し、必要に応じて調整します。

ステップ 2.データレイクコンソールにサインインする

- 管理者の E メールに送信された URL と一時パスワードを使用してサインインします。
- ソリューションのオンラインガイドを確認します。

ステップ 1.スタックを起動する

AWS CloudFormation テンプレートは、データレイクソリューションを AWS クラウドに自動的にデプロイします。

注意: このソリューションの実行中に使用した AWS サービスのコストは、お客様の負担となります。詳細については、コストセクションを参照してください。詳細については、このソリューションで使用する各 AWS サービスの料金表ウェブページを参照してください。

1. AWS マネジメントコンソールにサインインし、右側のボタンをクリックして、data-lake-deploy の AWS CloudFormation テンプレートを起動します。
   独自の実装開始点としてテンプレートをダウンロードすることもできます。

   ソリューションの起動

2. テンプレートは、デフォルトで米国東部 (バージニア北部) リージョンで起動されます。別の AWS リージョンでデータレイクソリューションを起動するには、コンソールのナビゲーションバーのリージョンセレクターを使用します。

   注意: このソリューションでは、現在特定の AWS リージョンでのみ使用可能な Amazon Cognito、Amazon Athena、および AWS Glue を使用します。したがって、これらのサービスが利用可能な AWS リージョンでこのソリューションを起動する必要があります。[1]

3. [**Select Template**] ページで正しいテンプレートを選択したことを確認し、[**Next**] を選択します。
4. [**Specify Details**] ページで、データレイクソリューションスタックに名前を割り当てます。
5. [**Parameters**] で、テンプレートのパラメータを確認し、必要に応じて変更します。このソリューションでは、以下のデフォルト値を使用します。

パラメータ	デフォルト	説明
Administrator Name	**<入力が必須>**	最初のソリューション管理者のユーザー名。ソリューションのデプロイ後、この管理者は、追加の管理者を含むその他のユーザーを作成および管理できます。
Administrator Email	**<入力が必須>**	管理者ユーザーに関連付けられた有効な E メール
Cognito Domain	**<入力が必須>**	Amazon Cognito でホストされるドメインで使用可能なドメインプレフィックスを選択します。このソリューションでは、Kibana のユーザー認証に Amazon Cognito を使用します。そのための前提条件として、ユーザープールのドメイン名を定義する必要があります。

6. [**Next**] を選択します。
7. [**Options**] ページでスタックのリソースのタグ (キーと値のペア) を指定し、追加のオプションを設定して、[**Next**] を選択します。
8. [**Review**] ページで、設定を見直して確認します。テンプレートがカスタム名を使って AWS Identity and Access Management (IAM) リソースを作成することを確認するチェックボックスを必ずオンにします。
9. [**Create**] を選択してスタックをデプロイします。

   スタックのステータスは、AWS CloudFormation コンソールの [**Status**] 列で表示できます。スタックの起動後、3 つのネストされたスタックは同じ AWS リージョンで起動されます。すべてのスタックとスタックリソースが正常に起動すると、CREATE_COMPLETE というメッセージが表示されます。これには 30 分以上かかる場合があります。

[1] 利用可能なサービスに関する AWS リージョン別の最新情報は、https://aws.amazon.com/about-aws/global-infrastructure/regional- product-services/ を参照してください。

ステップ 2.データレイクコンソールにサインインする

データレイクスタックの起動が完了すると、データレイクコンソールの URL と一時パスワードが記載された E メールが管理者に送信されます。

> **注意:** このメールは **no-reply@verificationemail.com** から送信されます。 E メール設定をチェックして、このドメインからの E メールをブロックまたはフィルタリングしないようにしてください。

1. E メールのリンクをクリックしてソリューションコンソールを開き、E メールアドレスと一時パスワードを使ってサインインします。
2. 新しいパスワードを設定するよう求められ、その後コンソールにサインインします。
3. 上部のナビゲーションバーで [**Support**] を選択し、オンラインガイドを開きます。

   具体的な手順と例については、ガイドのサブセクション (**ユーザーガイド**、**管理者ガイド**、および **CLI**) を参照してください。

セキュリティ

AWS クラウドは、スケーラブルで信頼性の高いプラットフォームを提供し、お客様がアプリケーションとデータを迅速かつ安全にデプロイできるようにします。AWS インフラストラクチャでシステムを構築する場合、お客様と AWS の間でセキュリティ上の責任が分担されるため、運用上の負担を軽減できます。AWS のセキュリティの詳細については、AWS セキュリティセンターを参照してください。

ユーザー認証

承認されたユーザーは、ソリューション生成コンソール、データレイク CLI、またはデータレイク API への直接呼び出しを使用して、データレイクにアクセスします。ユーザーは、ユーザー名 (デフォルトでは E メール) とパスワードを使用してデータレイクコンソールにサインインします。コンソールに対する認証は、Amazon Cognito のユーザープールで管理されます。

データレイク API へのリクエストは HTTPS ベースであり、ユーザーのアイデンティティを確認するために、アクセスキー (アクセスキーとシークレットアクセスキーの組み合わせ) で署名する必要があります。管理者は、個々のユーザーごとに API アクセスを付与できます。ユーザーに API アクセスが許可されている場合、データレイク API へのユーザーの呼び出しを識別するためのアクセスキーが生成されます。各ユーザーは、独自のシークレットアクセスキーを生成して、データレイク CLI を操作したり、API を直接呼び出したりできます。

追加のコンポーネントレベルのセキュリティ情報については、付録 Cを参照してください。

その他のリソース

AWS のサービス

- AWS CloudFormation
- AWS Lambda
- Amazon S3
- Amazon CloudFront
- Amazon DynamoDB
- Amazon API Gateway
- Amazon Athena
- Amazon Cognito
- Amazon Elasticsearch Service
- AWS Key Management Service
- AWS Glue
- Amazon CloudWatch

AWS のウェブページ

- データレイクとは
- AWS でのデータレイクと分析

付録 A: フェデレーティッドテンプレート - Active Directory

Microsoft Active Directory などの既存の SAML ID プロバイダと統合したいお客様のために、このデータレイクソリューションには、Active Directory (AD) のフェデレーション設定と同じワークフローをデプロイするもう一つ別の AWS CloudFormation テンプレートが含まれています。

AWS CloudFormation テンプレート

ソリューションには以下の AWS CloudFormation テンプレートが含まれており、デプロイ前にダウンロード可能です。

テンプレートを表

data-lake-deploy-federated.template: このテンプレートを使用して、Microsoft Active Directory などの既存の SAML ID プロバイダーと統合する準備ができているバージョンのソリューションを起動します。

このテンプレートは、以下次のネストされたスタックを起動します。

- **data-lake-storage.template:** このテンプレートは、ソリューションの Amazon S3、Amazon ES、および Amazon DynamoDB コンポーネントをデプロイします。
- **data-lake-services.template:** このテンプレートは、AWS Lambda マイクロサービスと必要な IAM ロールおよびポリシーをデプロイします。さらに、ソリューションの AWS KMS リソースをデプロイします。
- **data-lake-api.template:** このテンプレートは、Amazon API Gateway リソースをデプロイします。

自動デプロイ

自動デプロイを開始する前に、このガイドで説明されているアーキテクチャ、設定、ネットワークセキュリティ、その他の考慮事項をよくお読みください。このセクションの手順に従って、AD FS を使用してデータレイクソリューションを設定し、アカウントにデプロイします。

デプロイ時間: 約 30 分

ここでカバーする内容

このアーキテクチャを AWS にデプロイする手順は、以下のステップで構成されます。詳細な手順については、各ステップのリンクを参照してください。

ステップ 1.スタックを起動する

- AWS アカウントで AWS CloudFormation テンプレートを起動します。
- 必須パラメータの値 [**Stack Name**]、[**Cognito Domain**]、[**AD FS Hostname**] を入力します。

ステップ 2.AD FS を完了する

- AD FS を手動で設定します。

ステップ 1.スタックを起動する

> **注意**: このソリューションの実行中に使用した AWS サービスのコストは、お客様の負担となります。詳細については、コストセクションを参照してください。詳細については、このソリューションで使用する各 AWS サービスの料金表ウェブページを参照してください。

1 AWS マネジメントコンソールにサインインし、右側のボタンをクリックして、data-lake-deploy-federated の AWS CloudFormation テンプレートを起動します。

ソリューションの起動

独自の実装開始点としてテンプレートをダウンロードすることもできます。

2 テンプレートは、デフォルトで米国東部 (バージニア北部) リージョンで起動されます。別の AWS リージョンでデータレイクソリューションを起動するには、コンソールのナビゲーションバーのリージョンセレクターを使用します。

> **注意**: このソリューションでは、現在特定の AWS リージョンでのみ使用可能な Amazon Cognito、Amazon Athena、および AWS Glue を使用します。したがって、これらのサービスが利用可能な AWS リージョンでこのソリューションを起動する必要があります。[2]

3 [**Select Template**] ページで正しいテンプレートを選択したことを確認し、[**Next**] を選択します。

4 [**Specify Details**] ページで、データレイクソリューションスタックに名前を割り当てます。

5 [**Parameters**] で、テンプレートのパラメータを確認し、必要に応じて変更します。このソリューションでは、以下のデフォルト値を使用します。

パラメータ	デフォルト	説明
Cognito Domain	**<入力が必須>**	Amazon Cognito でホストされるドメインで使用可能なドメインプレフィックスを選択します。このソリューションでは、Kibana のユーザー認証に Amazon Cognito を使用します。そのための前提条件として、ユーザープールのドメイン名を定義する必要があります。
AD FS Hostname	**<入力が必須>**	AD FS エンドポイントのホスト名を挿入します。

6 [**Next**] を選択します。

7 [**Options**] ページでスタックのリソースのタグ (キーと値のペア) を指定し、追加のオプションを設定して、[**Next**] を選択します。

8 [**Review**] ページで、設定を見直して確認します。テンプレートがカスタム名を使って AWS Identity and Access Management (IAM) リソースを作成することを確認するチェックボックスを必ずオンにします。

9 [**Create**] を選択してスタックをデプロイします。

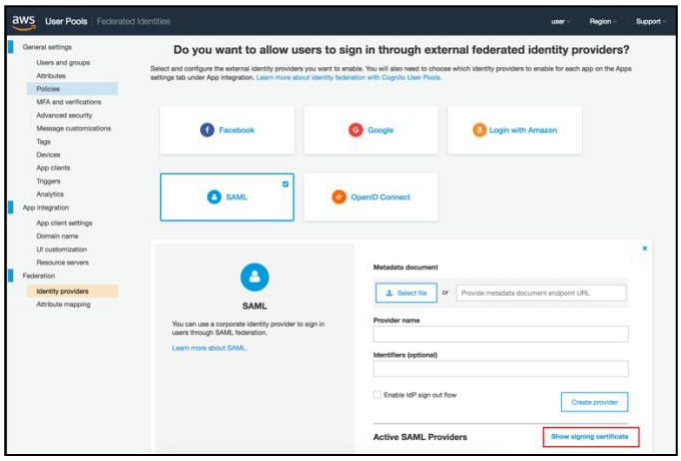
スタックのステータスは、AWS CloudFormation コンソールの [**Status**] 列で表示できます。スタックの起動後、3 つのネストされたスタックは同じ AWS リージョンで起動されます。すべてのスタックとスタックリソースが正常に起動すると、CREATE_COMPLETE というメッセージが表示されます。これには 30 分以上かかる場合があります。

[2] 利用可能なサービスに関する AWS リージョン別の最新情報は、https://aws.amazon.com/about-aws/global-infrastructure/regional- product-services/ を参照してください。

ステップ 2.AD FS を完了する

データレイクスタックの起動後、AD FS の設定を完了する必要があります。

1. AWS マネジメントコンソールにサインインし、スタックの [**Outputs**] タブに移動します。
2. [**RelyingPartyURL**]、[**RelyingPartyTrustedIdentifier**]、および [**LogoutTrustedURL**] キーの値に注意してください。
3. [**IdentityProvidersUrl**] リンクに移動します。
4. [**Federation**] コンソールの [**Identity providers**] セクションの [**Active SAML Providers**] で、[**Show signing certificate**] を選択します。

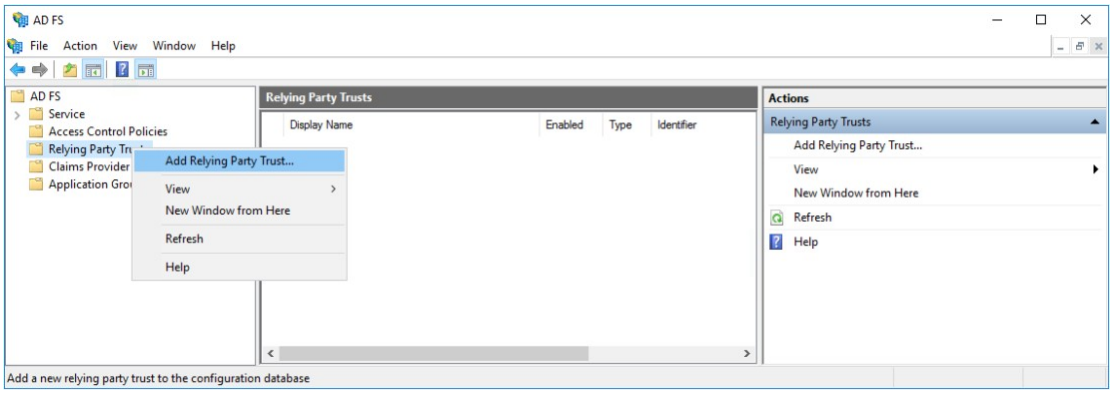


5. パブリックキーを含む証明書をコピーします。このキーは、AD FS サーバーの .cer ファイル (datalake.cer など) への署名付きログアウトリクエストを検証するために ID プロバイダーによって使用されます。

AD FS で証明書利用者として Amazon Cognito を追加する:

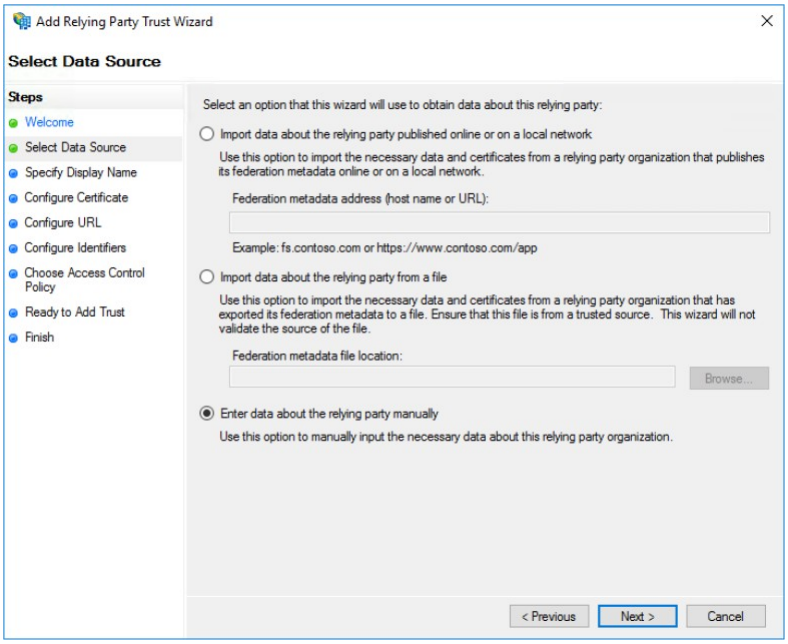
AD FS は、ID またはクレームプロバイダー (Active Directory) と証明書利用者 (Cognito) の 2 者が参加して行われます。証明書利用者は、フェデレーションサービスにおけるクレームプロバイダーの信頼によって表されるフェデレーションパートナーです。Active Directory フェデレーションサービスで新しい証明書利用者を設定するには、以下の手順に従います。

1. AD FS マネジメントコンソールで [**AD FS**] を右クリックし、[**Add Relying Party Trust**] を選択します。



2. [**Add Relying Party Trust Wizard**] で、[**Start**] を選択します。

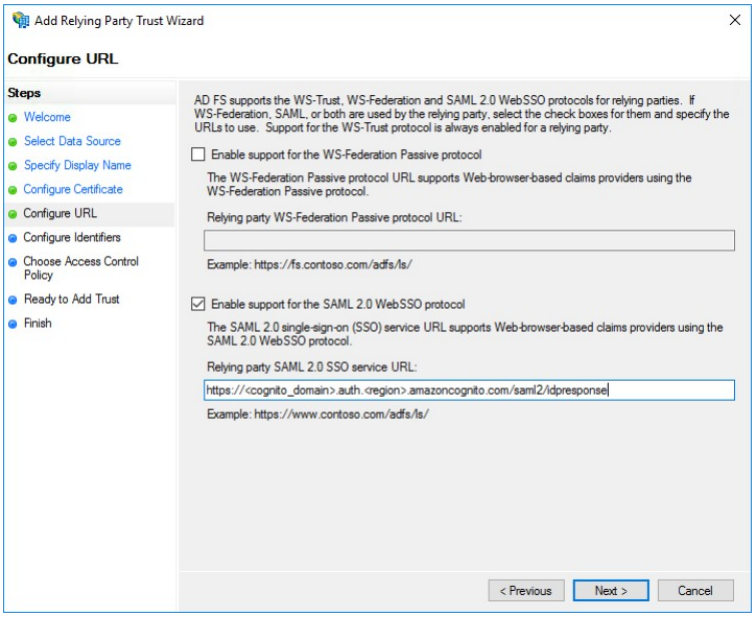
3. [**Select Data Source**] を選択します。次に、[**Enter data about the relying party manually**] ラジオボタンを選択し、[**Next**] を選択します。



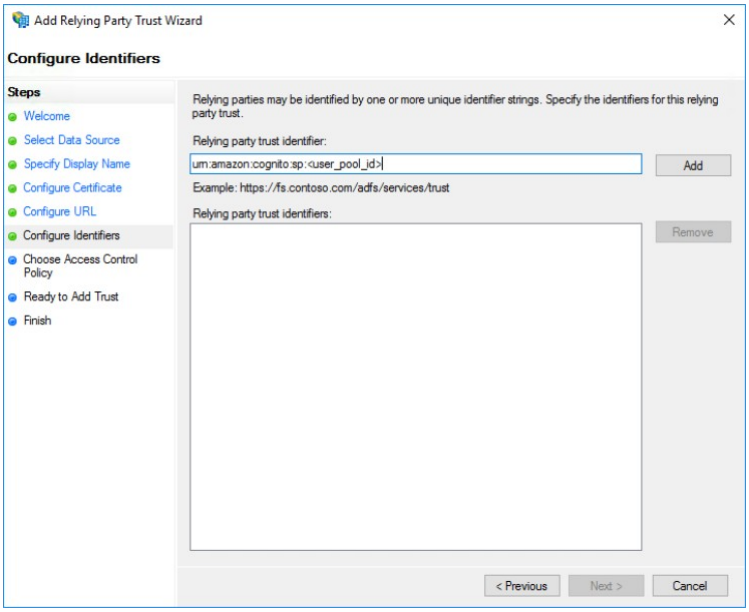
4. [**Specify Display Name**] を選択し、[Display Name] を設定します (例 :AWS でのデータレイクソリューション)。

5. [**Configure Certificate**] を選択し、[**Next**] を選択してデフォルト値を受け入れます。

6. [**Configure URL**] を選択します。次に、[**Enable support for the SAML 2.0 WebSSO protocol**] を選択し、URL 証明書利用者を設定します ([**RelyingPartyURL**] 出力パラメータで書き留めた値を使用します)。次に、[**Next**] を選択します。



7. [**Configure Identifiers**] を選択し、[**Relying party trusted identifier**] を設定します ([**RelyingPartyTrustedIdentifier**] 出力パラメーターで書き留めた値を使用します)。次に、[**Next**] を選択します。



8. ウィザードが終了するまで、[**Next**] を選択します。

サインアウトフローを有効化する

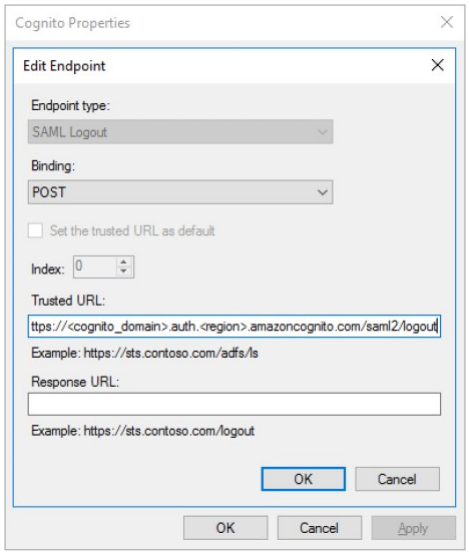
このフローを有効にすると、ログアウトが呼び出された際に、署名付きログアウトリクエストが Active Directory に送信されます。AD は署名付きログアウトリクエストを処理し、Amazon Cognito セッションからユーザーをログアウトします。

> **注意:** AD FS サーバーでは、署名付きログアウトリクエストが必要です。
> Amazon Cognito が提供する署名証明書を AD FS で設定する必要があります。

以下の手順に従って、Active Directory フェデレーションサービスからのログアウトレスポンスを使用するようにこのエンドポイントを設定します。

1. AD FS マネジメントコンソールで [**relying party**] をダブルクリックし、[**Endpoints**] タブ、[**Add SAML**] の順に選択します。

2. [**Endpoint type**] を [**SAML Logout**] に、[**Binding**] を [**POST**] に設定し、[**Trusted URL**] の値を設定します ([**LogoutTrustedURL**] 出力パラメーターで書き留めた値を使用します）。次に、[OK] を選択します。



3. [**Signature**] タブを選択し、フェデレーションコンソール (datalake.cer) からコピーした証明書を追加し、[**OK**] を選択します。

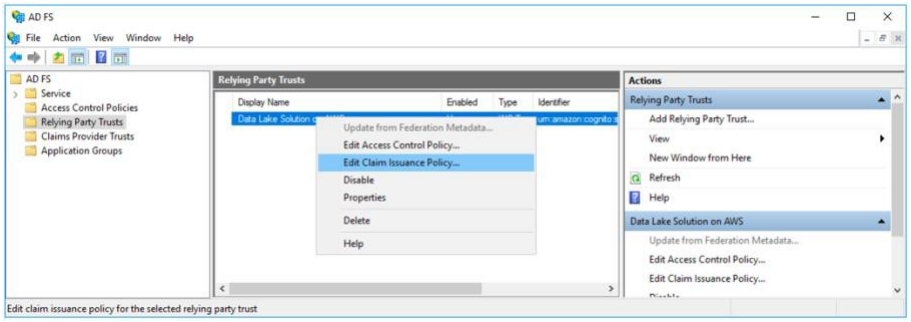
カスタムクレームルール

Microsoft AD FS では、クレームルール言語を使用して、クレームプロバイダーと証明書利用者の間でクレームを発行および変換します。クレームとは、信頼されたソースからのユーザーに関する情報です。信頼されたソースは、情報が真であり、そのソースがユーザーを認証したことをアサートしています。クレームプロバイダーは、クレームのソースです。これは、Active Directory などの属性ストアから取得された情報です。Amazon Cognito のユーザープールは、バインド後のエンドポイントで SAML 2.0 フェデレーションをサポートします。これにより、ユーザープールがユーザーエージェントを介して ID プロバイダーから SAML レスポンスを直接受信するため、アプリで SAML アサーションレスポンスを取得または解析する必要がなくなります。ユーザープールは、アプリケーションに代わってサービスプロバイダーとして機能します。

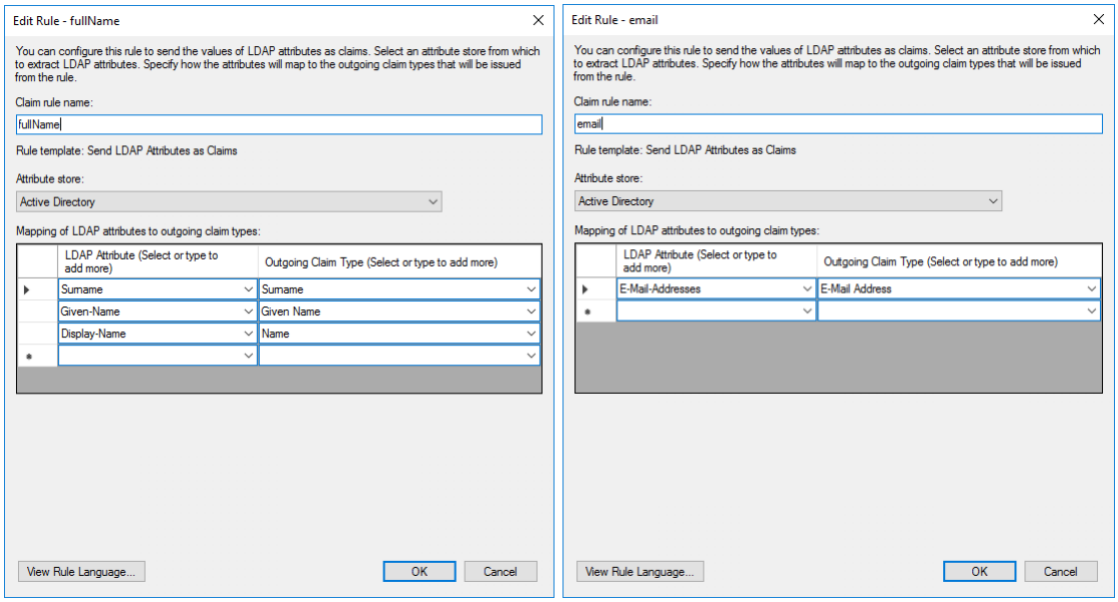
Active Directory フェデレーションサービスで新しい証明書利用者を設定するには、以下の手順に従います。

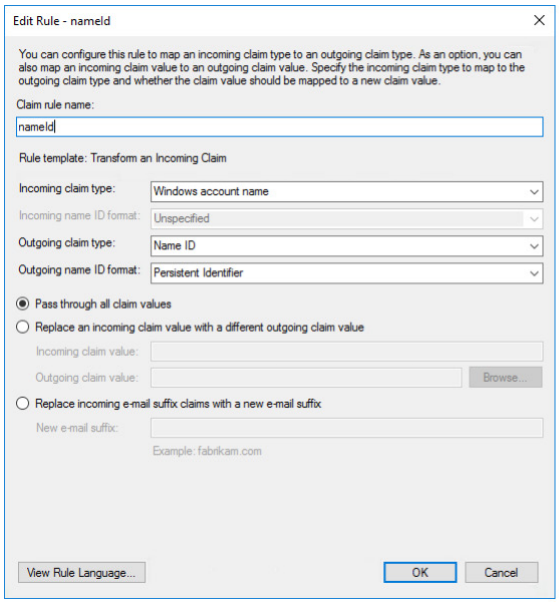
この手順では、**データレイク管理者**グループの**ロール**の送信クレームタイプのすべてのメンバーを**管理者**として設定することに注意してください。

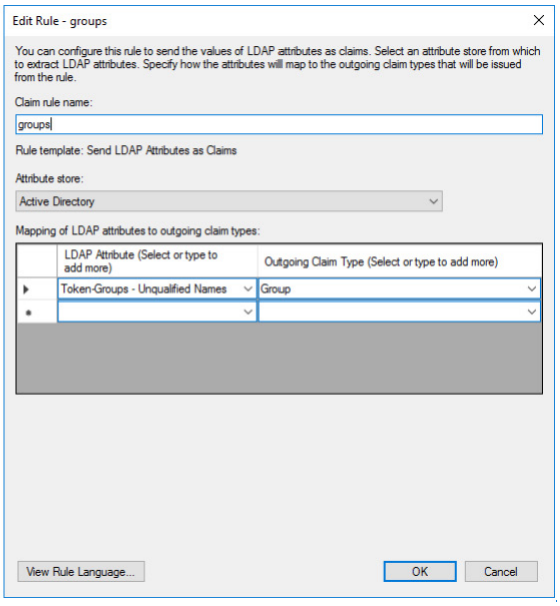
1. AD FS マネジメントコンソールで、[**relying party**] を右クリックし、[**Edit Claim Issuance Policy**] を選択します。

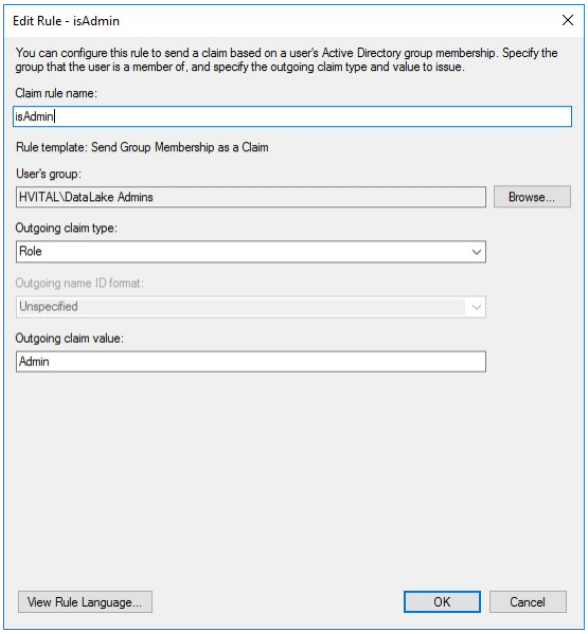


2. **クレームルール名**を指定します。
3. [**Attribute store**] を選択します。ユーザーが Active Directory にある場合、これは Active Directory になることがあります。
4. LDAP 属性 (E-Mail-Address など) を Outgoing Claim Type (E-Mail Address など) にマッピングします。
   SAML アサーションのユーザープールに必要な属性 (fullName、email、nameId、groups、isAdmin) が AD FS に設定されていることを確認してください。









5. AWS マネジメントコンソールにサインインし、スタックの [**Outputs**] タブに移動します。
6. [**ConsoleUrl**] キーの [**Value**] を選択します。AD FS マネジメントコンソールにリダイレクトされます。

付録 B: フェデレーティッドテンプレート - Okta

Okta を Security Assertion Markup Language 2.0 (SAML 2.0) ID プロバイダー (IdP) として使用する場合は、data-lake-deploy.template によってプロビジョニングされるリソースを Okta と統合するように変更できます。data-lake-deploy.template をデプロイしたら、以下の手順に従ってソリューションを Okta と統合します。

Okta アカウントを設定する

データレイクスタックの起動中に Okta ウェブサイトに移動し、Okta アプリケーションを使用してアカウントを設定できます。

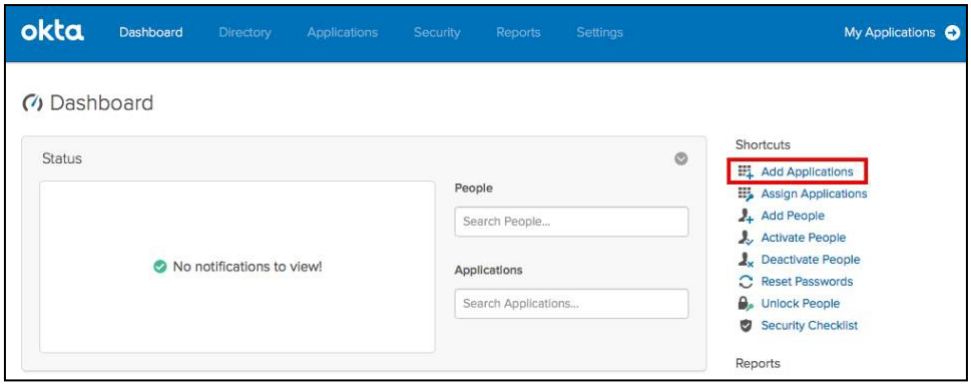
アカウントを作成するか、既存のアカウントにサインインする

> **注意：**すでに Okta アカウントをお持ちの場合は、サインインして [**Dashboard**] ページに移動します。アプリケーションを作成するまでスキップします。

1. Okta ウェブサイトに移動し、[**Sign Up Today**] を選択します。
2. [**Free Trial**] ウィンドウで連絡先情報を入力し、[**Get Started**] を選択します。提供されたアドレス宛に Okta から確認メールが送信されます。
3. 確認メールのログイン情報を使用して、アカウントにサインインします。パスワードの変更を求められます。

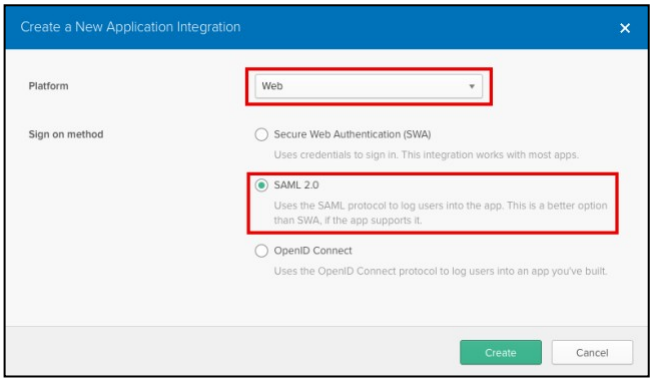
アプリケーションを作成する

1. [**Dashboard**] を選択して、管理ダッシュボードに移動します。
2. 管理ダッシュボードの [**Shortcuts**] で、[**Add Applications**] を選択します。



3. [**Create New App**] を選択します。

4. [**Create a New Application Integration**] ダイアログで、以下を選択します。
    a. [**Platform**] には [**Web**] を選択します。
    b. [**Sign on method**] には [**SAML 2.0**] を選択します。
    c. [**Create**] を選択します。



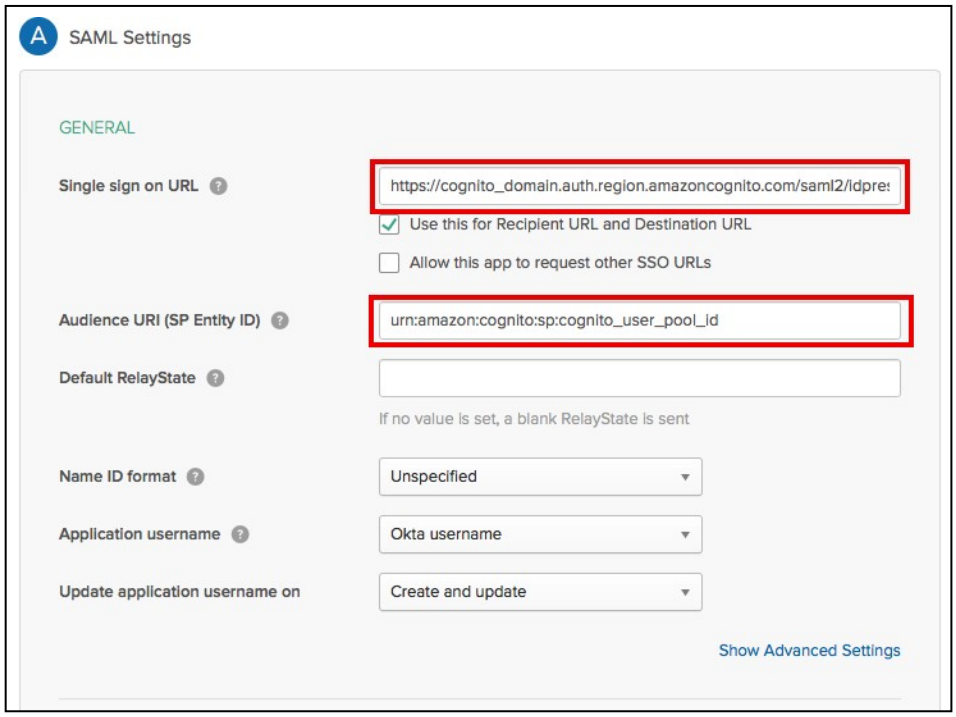
Okta アプリケーションの SAML Integration を設定する

1. [**Create SAML Integration**] ページの [**General Settings**] でアプリケーションの名前を入力し、[**Next**] を選択します。
2. [**SAML Settings**] で、以下の操作を行います。
    a. [**Single sign on URL**] に、以下のように入力します。
    https://**<cognito_domain_prefix>**.auth.**<region>**.amazoncognito.com/saml2/idpresponse

> **注意:** [**Domain prefix**] は、ユーザープールの管理ページの [**Domain name**] タブにある Amazon Cognito コンソールで確認できます。

    b. [**Audience URI (SP Entity ID)**] に、以下のように入力します。
    urn:amazon:cognito:sp:**<cognito_user_pool_id>**

> **注意:** ID は、ユーザープールの管理ページの [**General setting**] タブにある Amazon Cognito コンソールで確認できます。



3. [**Default RelayState**] は空白のままにします。
4. [**Attribute Statements**] で、以下の情報を追加します。
   - [**Name format**] が [**Unspecified**] に設定されていることを確認します。
   - [**Add Another**] を選択します。
   - 以下の表から対応するフィールドに値をコピーします。

Name	Value
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/givenname	user.lastName
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/name	user.firstName
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/emailaddress	user.email

5. [**Group Attribute Statements**] で、以下の情報を追加します。
   - [**Name format**] が [**Unspecified**] に設定されていることを確認します。
   - [**Add Another**] を選択します。
   - 以下の表から対応するフィールドに値をコピーします。
   - [**Filter**] でドロップダウン矢印を選択し、表で識別されているフィルタを選択してフィールドに値を入力します。

Name	Filter	Value
http://schemas.xmlsoap.org/claims/Group	Matches regex	.*
http://schemas.microsoft.com/ws/2008/06/identity/claims/role	Equals	admin

6. [**Next**] を選択します。
7. セクション 3 の [**Help Okta Support**] で、顧客であるかパートナーであるかを選択します。
8. [**Finish**] を選択します。

ユーザーとグループを設定する

注意: すでにユーザーとグループがある場合は、Okta アプリケーションにユーザーを割り当てるに進んでください。

1. [**Directory**] で、[**People**] を選択します。
2. [**Add Person**] を選択します。ウィンドウで、該当する情報を入力してユーザーを追加します。
3. [**Directory**] で、[**Groups**] を選択します。
4. グループを作成し、ユーザーをグループに割り当てます。
   管理者グループのメンバーには、コンソールにサインインした際に、ソリューションへの管理者アクセス権が付与されます。

Okta アプリケーションにユーザーを割り当てる

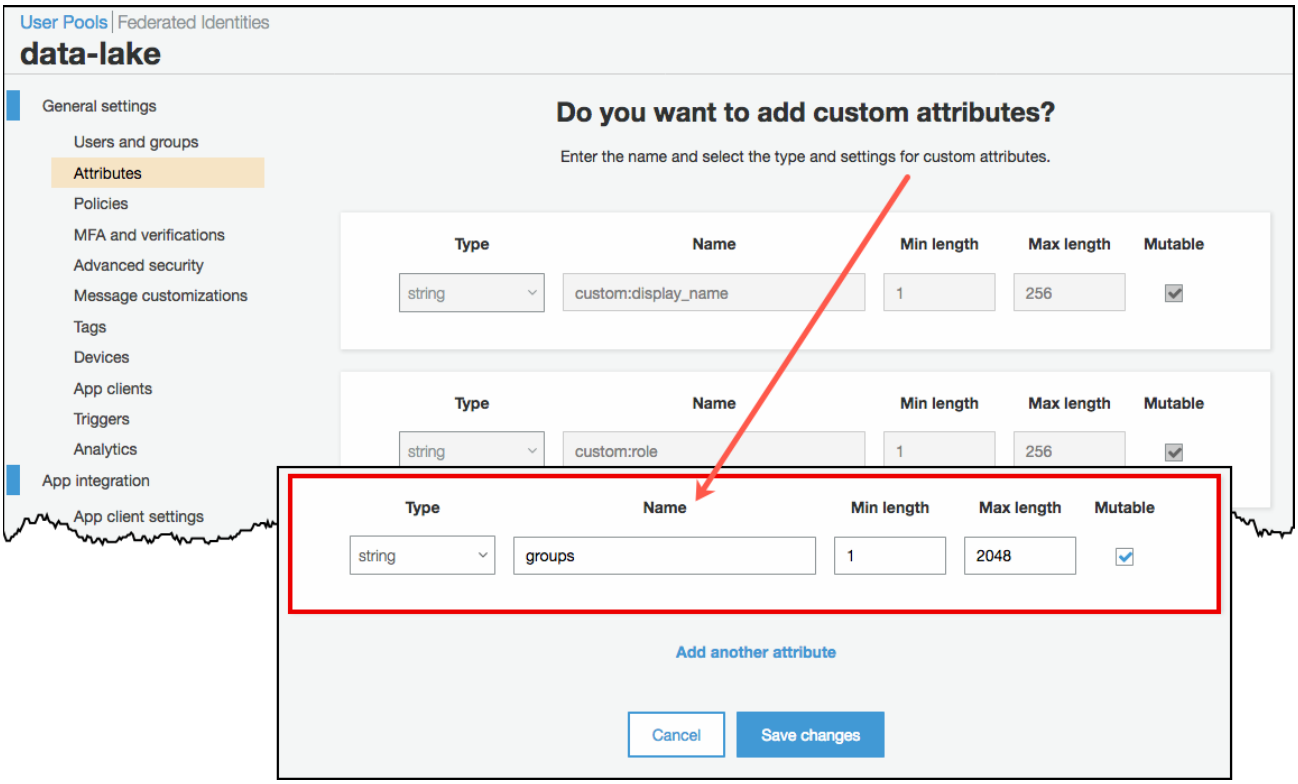
1. [**Assignments**]、[**Assign**] と進み、[**Assign to People**] を選択します。
2. 割り当てたいユーザーの横にある [**Assign**] を選択します。
3. (オプション) [**User Name**] には、ユーザー名を入力するか、ユーザーの E メールアドレスのままにします。
4. [**Save and Go Back**] を選択します。ユーザーが割り当てられます。
5. [**Done**] を選択します。

Amazon Cognito フェデレーションを完了する

データレイクスタックが起動し、Okta アプリケーションが作成されたら、Amazon Cognito フェデレーション設定を完了する必要があります。

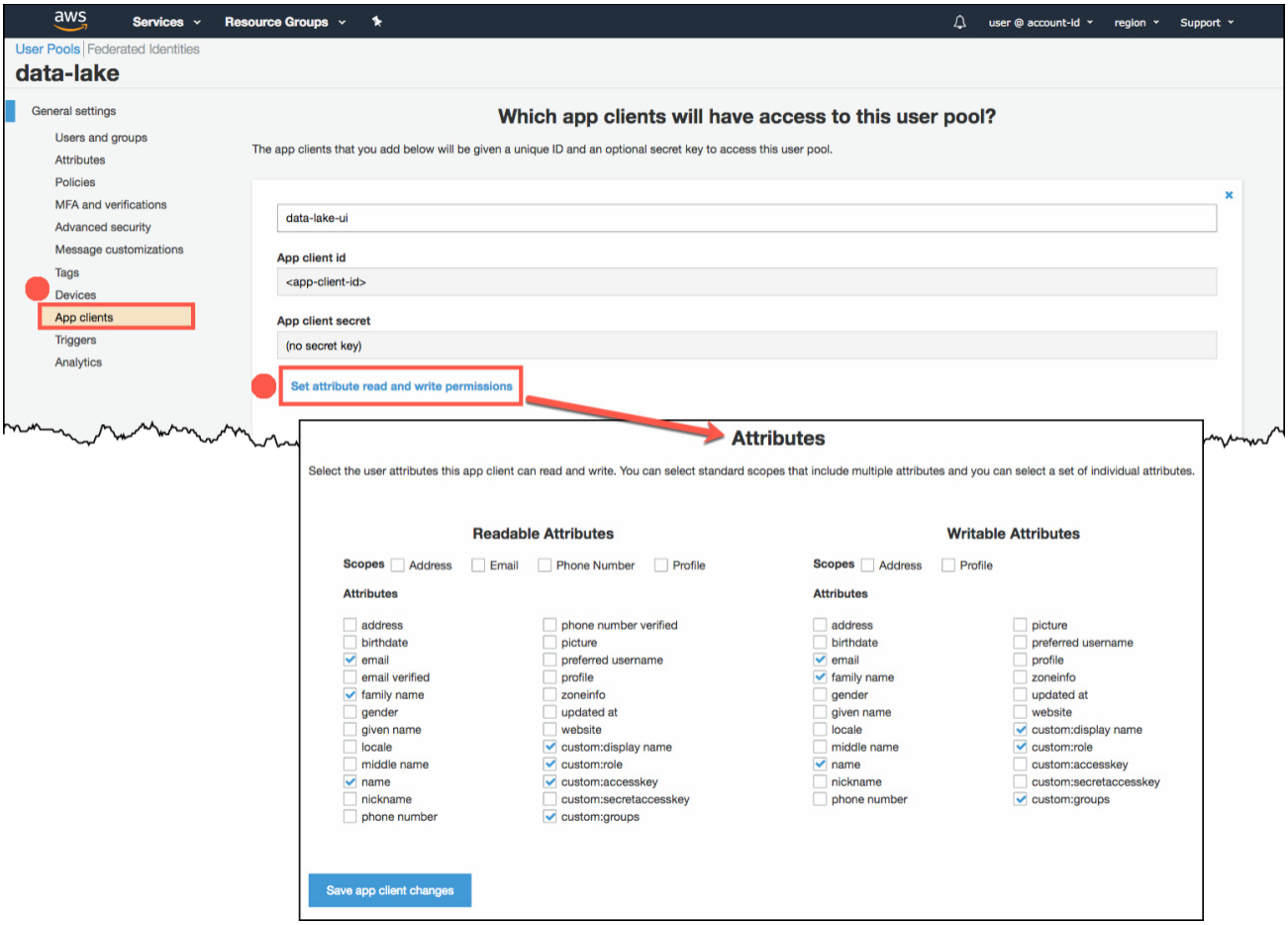
1. Cognito コンソールの [**General settings**] の [**Attributes**] で、[custom attributes] セクションまで下にスクロールし、[**Add another attribute**] を選択して、以下のカスタム属性を追加します。

Type	Name	Min Length	Max Length	Mutable
string	groups	1	2048	checked

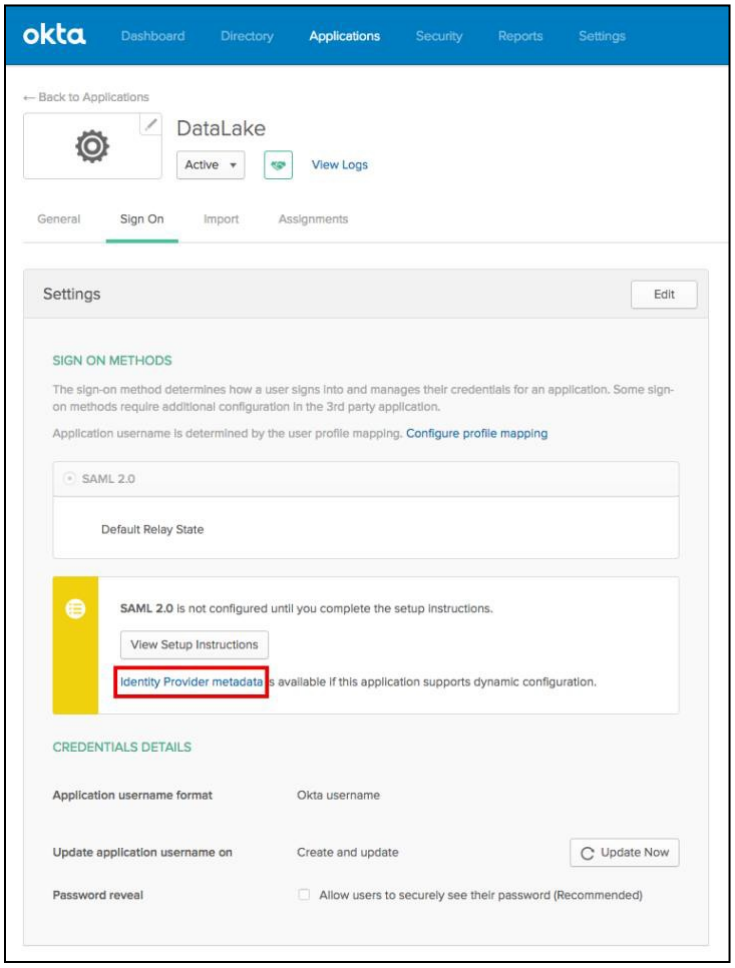


2. [**Save changes**] を選択します。
3. [**General settings**] の [**App Clients**] で **data-lake-ui** アプリクライアントを探し、[**Show Details**] を選択します。
4. [**Set attribute read and write permissions**] リンクを選択し、[**Attributes**] セクションにアクセスします。
5. 読み取り可能および書き込み可能属性が下表のとおりにチェックされていることを確認します (チェックされていない場合はチェックを入れ、異なる属性がチェックされている場合は必ずチェックを外してください)。これらの属性にチェックを入れることで、アプリクライアントはそれらにアクセスできるようになります。

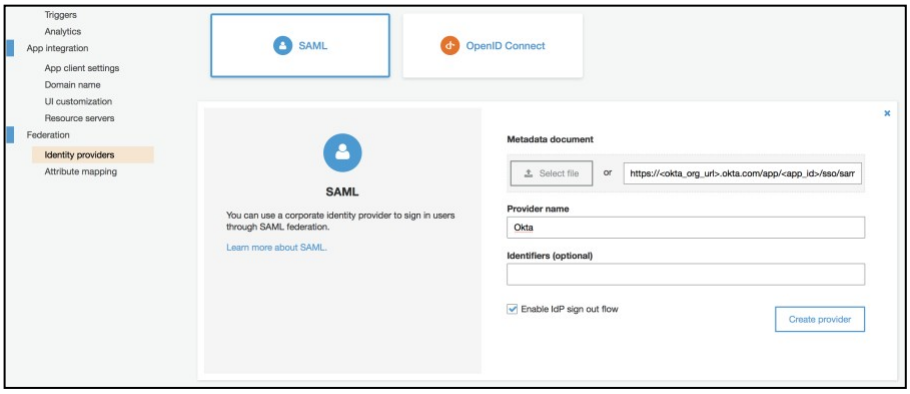
読み取り可能属性		書き込み可能属性	
• email	• custom:display name	• email	• custom:display name
• family name	• custom:role	• family name	• custom:role
• name	• custom:accesskey	• name	• custom:accesskey



6. [**Save app client changes**] を選択します。
7. [**Okta**] の [**Sign On**] で、[**Identity Provider metadata**] のリンクの場所をコピーします。



8. [**Federation**] の [**Identity providers**] で、前の手順でコピーしたリンクを使用してメタデータドキュメントを設定します。



9. [**Federation**] の [**Attribute mapping**] で、[**SAML**] タブを選択したあと [**Add SAML attribute**] を選択します。

10. Okta ID プロバイダー用に以下のマッピングを作成します。

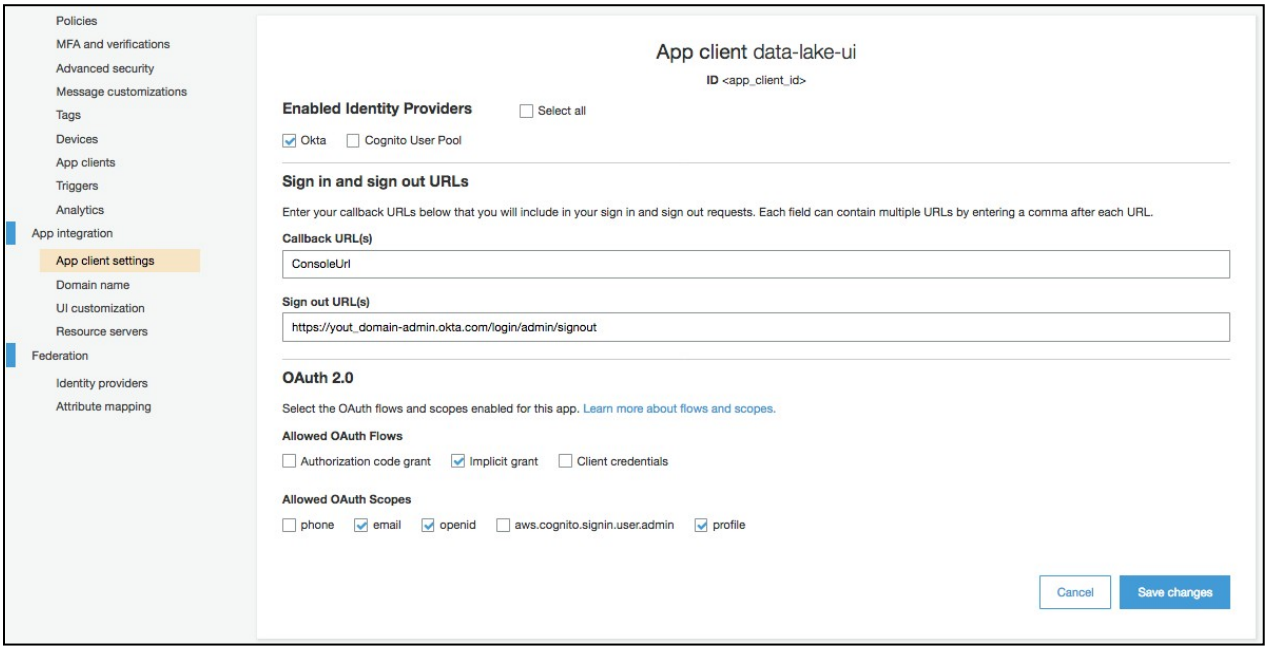
SAML 属性	ユーザープール属性
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/givenname	Name
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/displayname	custom:display_name
http://schemas.microsoft.com/ws/2008/06/identity/claims/role	custom:role
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/surname	Family Name
http://schemas.xmlsoap.org/claims/Group	custom:groups
http://schemas.xmlsoap.org/ws/2005/05/identity/claims/emailaddress	Email

11. [**Save changes**] を選択します。

12. [**App integration**] の [**App client settings**] で、[**App client data-lake-ui**] 設定を更新して新しい ID プロバイダーを使用します。

    - [**Enable Identity Providers**] で、[**Okta**] を選択します。
    - [**CallBack URL(s)**] には、AWS CloudFormation コンソールに移動し、スタックの [**Outputs**] タブを開き、[**ConsoleUrl**] キーの [**Value**] で定義された URL をコピーします。
    - [**Sign out URL(s)**] に、以下のように入力します。
      https://**<your_domain>**-admin.okta.com/login/admin/signout
    - [**Allowed OAuth Flows**] で、[**Implicit grant**] を選択します。
    - [**Allowed OAuth Scopes**] で、[**email**]、[**openid**]、および [**profile**] を選択します。

13. 適切に設定されていることを確認します。



14. [**Save changes**] を選択します。

15. [**Federation**] の [**Identity providers**] で、[**SAML**] ボタンを選択します。

16. [**Active SAML Providers**] セクションで、[**Show signing certificate**] を選択します。

17. パブリックキーを含む証明書をコピーします。このキーは、Okta サーバーの .cer ファイル (datalake.cer など) への署名付きログアウトリクエストを検証するために ID プロバイダーによって使用されます。証明書に開始タグと終了タグを含めることを忘れないでください。結果は以下のようになります。

```
-----BEGIN CERTIFICATE-----
(ここに署名証明書情報を挿入します)
-----END CERTIFICATE-----
```

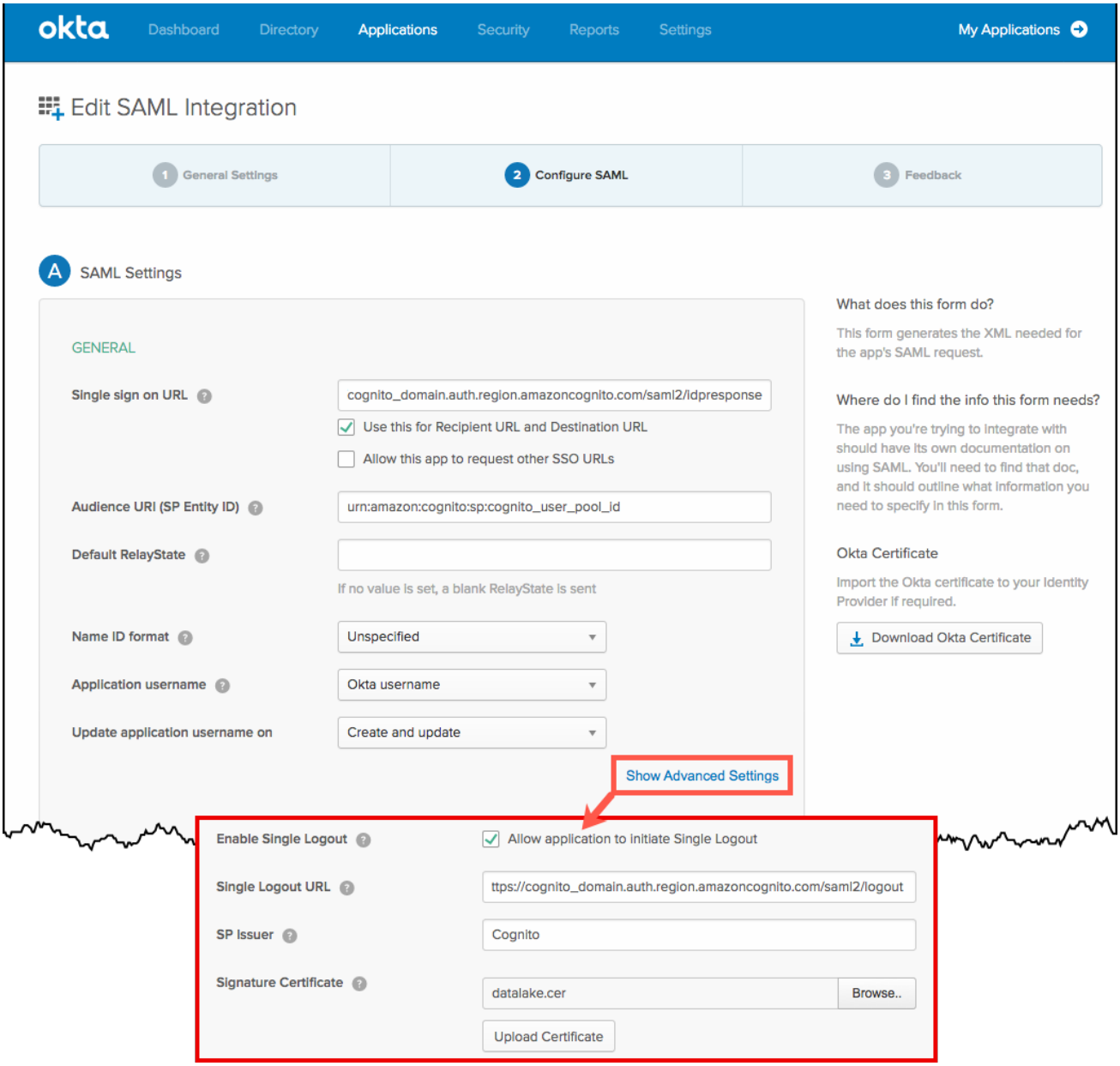
サインアウトフローを有効化する

サインアウトフローを有効にすると、ログアウトが呼び出された際に、署名付きログアウトリクエストが Okta に送信されます。Okta は署名付きログアウトリクエストを処理し、Amazon Cognito セッションからユーザーをログアウトします。

> **注意:** Okta は署名付きログアウトリクエストを想定しているため、Okta アプリケーションで Amazon Cognito が提供する署名証明書を設定する必要があります。

以下の手順に従って、Okta アプリケーションからのログアウトレスポンスを使用するようにこのエンドポイントを設定します。

1. Okta ウェブサイトで [**Applications**] を選択し、アクティブなアプリケーションダッシュボードに移動します。
2. Okta アカウントを設定するで作成したデータレイクアプリケーションを選択します。
3. [**General**] タブを選択し、[**SAML Settings**] までスクロールして [**Edit**] を選択します。
4. [**General Settings**] で、[**Next**] を選択します。
5. [**SAML Settings**] で [**Show Advanced Settings**] を選択し、ログアウトパラメータを設定します。
    a. 追加のフィールドを表示するには、[**Enable Single Logout**] オプションを選択します。
    b. [**Single Logout URL**] に、以下のように入力します。
    https://**<cognito_domain>**.auth.**<region>**.amazoncognito.com/saml2/logout
    c. [**SP Issuer**] に「Cognito」と入力します。
    d. [**Signature Certificate**] で、コピーした証明書ファイル (datalake.cer など) を参照し、[**Upload Certificate**] を選択します。



6. [**Next**]、[**Finish**] の順に選択します。

Amazon S3 のデータレイクコンソール設定ファイルを更新する

すべての Amazon Cognito および Okta パラメータを設定したらデータレイクコンソール (Amazon S3 バケット内) に戻り、その設定を更新してログインフェデレーションをアクティブ化します。

1. AWS マネジメントコンソールで、Amazon S3 に移動します。
2. データレイクバケット (datalake-**<region>**-**<account_id>**) を選択します。

3. [**lib**] ファイルフォルダに移動します。
4. app-variables.js 設定ファイルをダウンロードします。
5. 設定ファイルを開き、パラメータを更新します。

```
var FEDERATED_LOGIN = true;

var LOGIN_URL =
"https://<yourDomainPrefix>.auth.<region>.amazoncognito.com/login?re
sponse_type=token&client_id=<yourClientId>&redirect_uri=<ConsoleUrl> ";

var LOGOUT_URL =
"https://<yourDomainPrefix>.auth.<region>.amazoncognito.com/
logout?response_type=token&client_id=<yourClientId>&redirect_uri=
<ConsoleUrl>&logout_uri=<ConsoleUrl>";
```

Amazon Cognito コンソールで特定のパラメーターを見つけます。

- **<yourDomainPrefix>** および **<region>** は、[**App integration**] の [**Domain name**] で確認できます。
- **<yourClientId>** は、[**General settings**] の [**App clients**] で確認できます。[**data-lake-ui App client**] に帰属する [**App client ID**] を使用します。
- **<ConsoleUrl>** は、AWS CloudFormation コンソールで確認できます。スタックの [**Outputs**] タブを開き、[**ConsoleUrl**] パラメータで定義された URL をコピーします。

6. 変更したファイルを Amazon S3 にアップロードします。
7. データレイクコンソールにアクセスするには、以下のリンクを使用します。

   https://**<yourDomainPrefix>**.auth.<region>.amazoncognito.com/login?respons
   e_type=token&client_id=**<yourClientId>**&redirect_uri=**<ConsoleUrl>**

付録 C: ソリューションコンポーネント

AWS KMS キー

データレイク AWS KMS キー (エイリアス: **datalake**) は、ソリューション所有で、Amazon S3 に保存されるすべてのデータセットオブジェクトを暗号化するために作成されます。さらに、AWS KMS キーはデータレイクへの API アクセス用に各ユーザーの Amazon Cognito ユーザープールレコードのシークレットアクセスキーを暗号化するために使用されます。

Amazon CloudFront

このソリューションは、Amazon CloudFront ディストリビューションを設定して、データレイクコンソールの HTTPS リクエストを処理します。

Amazon S3

このソリューションは、デフォルトの Amazon S3 バケットを使用して、ユーザーがデータレイクにアップロードするパッケージに関連付けられたデータセットとマニフェストファイルを保存します。さらに、バケットには、ユーザーがカート (パッケージのコレクション) をチェックアウトするときに生成されたマニフェストファイルが保存されます。このバケットへのすべてのアクセス (パッケージおよびマニフェストマイクロサービスからの get および put アクション) は、署名付き URL を介して制御されます。このバケットに保存されるすべてのオブジェクトは、データレイク AWS KMS キーを使用して暗号化されます。

2 番目の Amazon S3 バケットは、データレイクコンソールをホストします。このコンソールは、ユーザー認証に Amazon Cognito を使用する静的なウェブサイトです。エンドユーザーは S3 エンドポイントに直接アクセスできません。すべてのアクセスは、Amazon CloudFront ディストリビューションを介して行う必要があります。

Amazon Athena と AWS Glue

このソリューションでは、各データパッケージ内で AWS Glue クローラーが自動的に設定され、毎日のスキャンをスケジュールして変更を追跡します。クローラーはデータセットをクロールし、その一部を検査してデータスキーマを推測し、出力を AWS Glue データカタログで定義された 1 つ以上のメタデータテーブルとして保持します。

このカタログは一度作成されると、さまざまなデータソースと形式にわたる統合メタデータリポジトリを提供し、Amazon Athena および Amazon Redshift Spectrum と統合して、データレイク内のデータを直接インタラクティブにクエリおよび分析します。また、Amazon EMR、AWS Glue 抽出、変換、読み込み (ETL) ジョブ、Apache Hive データウェアハウスと互換性のある任意のアプリケーションを使用するため、データの分類、クリーンアップ、エンリッチ化、移動ができるようになります。

> **注意:** Athena または Amazon Redshift Spectrum を使用してデータベースとテーブルを以前に作成した場合は、Athena を AWS Glue データカタログにアップグレードする必要があります。Athena を初めて使用する場合は、何も変更する必要はありません。詳細については、AWS Glue データカタログへのステップバイステップのアップグレードを参照してください。

Amazon Cognito のユーザープール

データレイクコンソールは、Amazon Cognito を使用したユーザーアクセスに対して保護されており、Amazon Cognito のユーザープールとの統合を通じてデータレイクユーザーを管理するための管理インターフェイスを提供します。管理者のみがユーザーとグループを作成できます。ユーザーが作成されると、ソリューションはデータレイクに参加するための招待を自動的にユーザーに送信します。フェデレーティッドテンプレートを使用する場合、すべての管理タスクは AD サーバーで実行する必要があることに注意してください。

管理者が新しいユーザーを作成すると、そのユーザーに、関連付けられたアクセス許可と共に、以下のいずれかのロールが割り当てられます。

- **メンバー:** メンバーロールは、データレイク内で管理者業務ではないアクションを実行できます。これらのアクションには、以下が含まれます。
    - 所有者または表示可能なパッケージがメンバーグループにある場合、パッケージを表示および検索する
    - データレイク内のすべてのパッケージを表示および検索する
    - カート内のパッケージのマニフェストを追加、削除、および生成する
    - 作成したパッケージを作成、更新、および削除する
    - 作成したパッケージのメタデータを作成および更新する
    - 作成したパッケージに対してデータセットを追加および削除する
    - データレイクプロファイルおよび API アクセス情報を表示する
    - 管理者が API アクセスを許可した場合にシークレットアクセスキーを生成する

- **管理者:** 管理者ロールには、データレイクへのフルアクセス権があります。管理者ロールは、メンバーロールのアクションに加えて、以下のアクションを実行できます。
    - ユーザーの招待を作成し、ユーザーを 1 つ以上のグループに割り当てる
    - グループの作成、更新、および削除
    - データレイクユーザーの更新、無効化、有効化、および削除
    - グループへのユーザーの割り当て、削除、および再割り当て
    - ユーザーの API アクセスの作成、取り消し、有効化、および無効化
    - データレイク設定の更新
    - ガバナンス設定の作成、更新、および削除

データレイク API とマイクロサービス

データレイク API は HTTPS 経由でリクエストを受け取ります。API リクエストが行われると、Amazon API Gateway はカスタムオーソライザー (AWS Lambda 関数) を活用して、すべてのリクエストが承認されるようにします。

データレイクマイクロサービスは、すべてのデータレイクオペレーションのビジネスロジックとデータアクセスレイヤーを提供する一連の AWS Lambda 関数です。各 AWS Lambda 関数は、指定された機能を実行するために、最小権限 (最低限必要なアクセス権限) を持つ IAM ロールを引き受けます。以下のセクションでは、各データレイクマイクロサービスの概要を示します。

管理マイクロサービス

data-lake-admin-service は、/admin/* エンドポイントに送信されたデータレイク API リクエストを処理する AWS Lambda 関数です。管理マイクロサービスは、ユーザーとグループの管理、全般設定、ガバナンス設定、API キー、およびデータレイク内のすべてのオペレーションのロール認証を含む、すべての管理サービスを処理します。

カートマイクロサービス

data-lake-cart-service は、/cart/* エンドポイントに送信されたデータレイク API リクエストを処理する AWS Lambda 関数です。カートマイクロサービスは、アイテムリスト、アイテムの追加、アイテムの削除、およびユーザーカートのマニフェストの生成を含む、すべてのカートオペレーションを処理します。

マニフェストマイクロサービス

data-lake-manifest-service は、マニフェストファイルのインポートとエクスポートを管理する AWS Lambda 関数です。マニフェストマイクロサービスは、インポートマニフェストファイルをアップロードします。これにより、既存の Amazon S3 コンテンツをパッケージに一括インポートできます。また、チェックアウト時にユーザーのカート内の各パッケージのエクスポートマニフェストファイルを生成します。

パッケージマイクロサービス

data-lake-package-service は、/packages/* エンドポイントに送信されたデータレイク API リクエストを処理する AWS Lambda 関数です。パッケージマイクロサービスは、リスト、パッケージの追加、パッケージの削除、パッケージの更新、メタデータの一覧表示、メタデータの追加、メタデータの更新、データセットの一覧表示、データセットの追加、データセットの削除、マニフェストの処理、AWS Glue オンデマンドクローラーの実行、AWS Glue テーブルの一覧表示とアクセス、および Amazon Athena でのデータセットの表示を含む、すべてのパッケージ操作を処理します。

検索マイクロサービス

data-lake-search-service は、/search/* エンドポイントに送信されたデータレイク API リクエストを処理する AWS Lambda 関数です。検索マイクロサービスは、クエリ、インデックスドキュメント、およびインデックスドキュメントの削除を含む、すべての検索オペレーションを処理します。

プロファイルマイクロサービス

data-lake-profile-service は、/profile/* エンドポイントに送信されたデータレイク API リクエストを処理する AWS Lambda 関数です。プロファイルマイクロサービスは、シークレットアクセスキーの取得と生成を含む、データレイクユーザーのすべてのプロファイルオペレーションを処理します。

ログ記録マイクロサービス

data-lake-logging-service は、データレイクのマイクロサービスと Amazon CloudWatch Logs の間をインターフェイスする AWS Lambda 関数です。各マイクロサービスは、オペレーションイベントとアクセスイベントをログ記録サービスに送信し、ログ記録サービスによって Amazon CloudWatch Logs にイベントが記録されます。CloudWatch コンソールでこのログ (datalake/audit-log) にアクセスできます。

Amazon DynamoDB テーブル

データレイクソリューションでは、Amazon DynamoDB テーブルを使用して、データパッケージ、設定、およびユーザーカート項目のメタデータを保持します。以下のテーブルは、デプロイ中にプロビジョニングされ、データレイクマイクロサービスを介してのみアクセスされます。

- **data-lake-packages:** データパッケージのタイトルと説明の永続的保存、およびパッケージにアクセスできるグループのリスト
- **data-lake-metadata:** パッケージに関連付けられたメタデータタグ値の永続的保存
- **data-lake-datasets:** Amazon S3 オブジェクトへのデータセットポインタの永続的保存
- **data-lake-cart:** ユーザーカートアイテムの永続的保存
- **data-lake-keys:** ユーザーアクセスキー ID 参照の永続的保存
- **data-lake-settings:** データレイク設定とガバナンス設定の永続的保存

Amazon Elasticsearch Service クラスター

このソリューションでは、Amazon Elasticsearch Service クラスターを使用して、検索用のデータレイクパッケージデータのインデックスを作成します。クラスターには、検索マイクロサービスおよび Amazon Cognito 認証を介してのみアクセスできます。

付録 D: 運用メトリクスの収集

このソリューションには、匿名の運用メトリクスを AWS に送信するオプションが含まれています。当社はこのデータを使用して、お客様によるこのソリューションの使用状況をよりよく理解し、提供するサービスや製品の改善に役立てます。有効にすると、以下の情報が収集され、AWS に送信されます。

- **Solution ID:** AWS ソリューション識別子
- **Unique ID (UUID):** データレイクソリューションのデプロイごとにランダムに生成された一意の識別子
- **Timestamp:** データ収集タイムスタンプ
- **Cluster Size:** ソリューションがデプロイする Amazon Elasticsearch クラスターのサイズ

このオプションで収集されたデータは AWS に帰属します。データ収集には、AWS プライバシーポリシーが適用されます。この機能をオプトアウトするには、以下のように AWS CloudFormation テンプレートマッピングセクションを変更します。

```
Solution:
    Data:
        SendAnonymousUsageData: "Yes"
```

"Yes" を"No"に変更

```
Solution:
    Data:
        SendAnonymousUsageData: "No"
```

ソースコード

GitHub リポジトリにアクセスして、このソリューションのテンプレートとスクリプトをダウンロードし、カスタマイズを他のユーザーと共有できます。

ドキュメントの改訂

日付	変更	セクション
2016 年 11 月	初回リリース	
2018 年 6 月	詳細なアクセス権限、AWS Glue と Amazon Athena の統合	概要、コスト、ソリューションの特徴、付録 C、付録 D
2018 年 9 月	Active Directory フェデレーション	概要、アーキテクチャの概要、ソリューションの特徴、付録 A
2019 年 6 月	Okta フェデレーション	付録 B

注意

お客様は、この文書に記載されている情報を独自に評価する責任を負うものとします。このドキュメントは、(a) 情報提供のみを目的としており、(b) AWS の現行製品とプラクティスを表したものであり、予告なしに変更されることがあり、(c) AWS およびその関連会社、サプライヤー、またはライセンサーからの契約義務や確約を意味するものではありません。AWS の製品やサービスは、明示または暗示を問わず、いかなる保証、表明、条件を伴うことなく「現状のまま」提供されます。お客様に対する AWS の責任は、AWS 契約により規定されます。本書は、AWS とお客様の間で行われるいかなる契約の一部でもなく、そのような契約の内容を変更するものでもありません。

データレイクソリューションは、https://aws.amazon.com/asl/ で閲覧可能な Amazon ソフトウェアライセンスの条項に基づいてライセンスされます。

© 2019, Amazon Web Services, Inc. or its affiliates.All rights reserved.