数据湖解决方案

部署手册

顾明

版本：v2.3

最后更新时间: 2020 年 10 月

目 录

背景介绍

很多 Amazon Web Services 客户需要能够提供比传统数据管理系统更高的敏捷性和灵活性的数据存储和分析解决方案。数据湖是存储和分析数据的一种新方法，越来越受欢迎，因为它可使公司管理来自各种不同来源的多种数据类型，并将结构化和非结构化的此数据存储在集中存储库中。

Amazon Web Services 云提供帮助客户实施安全、灵活且经济高效的数据湖所需的许多构建块。这其中包括帮助摄取、存储、查找、处理和分析结构化和非结构化数据的 Amazon managed Services。为支持我们的客户构建自己的数据湖，Amazon Web Services 提供了数据湖解决方案，它是一种自动化参考实施方法，可在 Amazon Web Services 云上部署高度可用且经济高效的数据湖架构，同时提供了一个用于数据集搜索和请求的用户友好型控制台。

Amazon Web Services 解决方案实施概览

Amazon Web Services 提供了一个数据湖解决方案，该解决方案可以自动配置所需的核心 Amazon Web Services 服务，以便轻松标记、搜索、共享、转换、分析和管理公司内部或其他外部用户的特定数据子集。该解决方案部署了一个控制台，用户可以通过访问该控制台搜索和浏览满足其业务需求的可用数据集。该解决方案还包含一个联合模板，让您能够启动可与 Microsoft Active Directory 集成的解决方案版本。
部署此解决方案将在由西云数据运营的 Amazon Web Services（宁夏）区域或由光环新网运营的 Amazon Web Services（北京）区域部署中构建以下环境。

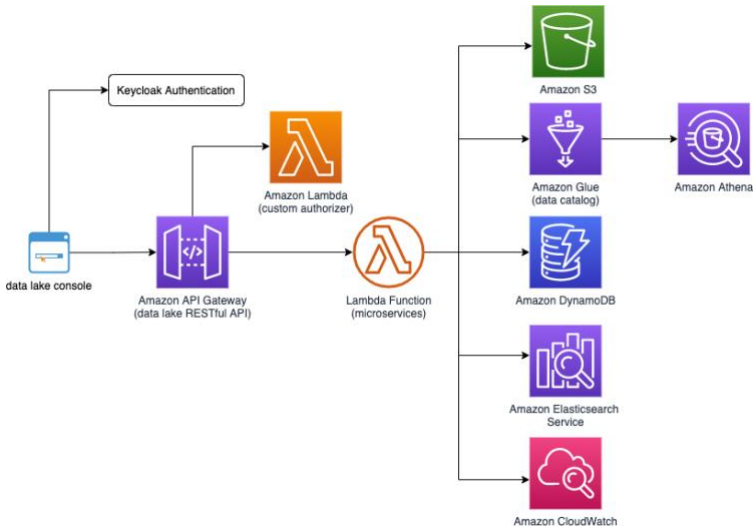


图 1：Amazon Web Services 上的 Data Lake 解决方案架构

Amazon CloudFormation 模板可配置解决方案的核心 Amazon Web Services 服务，其中包括一系列 Amazon Lambda 微服务（函数）、用于可靠的搜索功能的 Amazon Elasticsearch、用于数据转型的 Amazon Glue 和用于分析的 Amazon Athena。

该解决方案利用 Amazon S3 的安全性、持久性和可扩展性来管理组织数据集的持久性目录，并利用 Amazon DynamoDB 来管理相应的元数据。编制好数据集的目录后，

其属性和描述性标签将可供搜索。用户可以搜索并浏览解决方案控制台中的可用数据集，并且可以创建他们需要访问的数据列表。

该解决方案将跟踪用户选择的数据集，并在用户签出时生成一个其中包含至所需内容的安全访问链接的清单文件。

解决方案功能

该数据湖解决方案提供以下功能：

- **数据湖参考实施**：即用型的数据湖解决方案，或作为可定制的参考实现，以满足独特的数据管理，搜索和处理需求。
- **用户界面**：该解决方案自动创建一个托管在 Amazon S3 上并由 Amazon CloudFront 交付的基于 Web 的直观的控制台 UI。访问控制台可轻松管理数据湖策略，添加或删除数据包，搜索数据包以及创建数据集清单以进行其他分析。
- **托管存储层**：保护和管理托管 Amazon S3 存储桶中数据的存储和检索，并使用特定于解决方案的 Amazon Web Services 密钥管理服务（KMS）密钥对静态数据进行加密。
- **数据访问的灵活性**：利用预签名的 Amazon S3 URL，或使用适当的 Amazon Identity and Access Management（IAM）角色来控制但直接访问 Amazon S3 中的数据集。
- **数据转换和分析**：上传具有可搜索元数据的数据集，该数据集可与 Amazon Glue 和 Amazon Athena 集成以转换和分析数据。
- **Keycloak 用户登录**：使用第三方的 Keycloak 进行用户注册和登录认证服务进行登录和验证

快速部署 data lake 方案

本文的步骤只针对在运行在由西云数据运营的 Amazon Web Services（宁夏）区域或由光环新网运营的 Amazon Web Services（北京）区域中，您可以使用以下链接快速启动一个 CloudFormation 堆栈来部署和管理整个方案：

部署前提

1. 经过 ICP 备案的域名，因为中国区的 CloudFront 地址不能直接访问，这个域名会作为 CNAME 指向 data lake 部署完生成的 CloudFront 的地址，data lake 的 Web UI console 通过在这个域名来访问。

2. ICP 备案域名对应的证书，可以在 ACM 中创建该域名的证书，并通过域名验证，或者使用用户自己申请的域名对应的第三方证书。
3. 提前配置好一个可用的 Keycloak 的身份和访问控制系统，可以使用 Amazon Web Services 的 Keycloak 解决方案来一键式的部署 Keycloak，参考…。需要在 Keycloak 系统中配置好认证 data lake 所对应的 Realm，client，同时需要在 Keycloak 中注册好用户，后面会详细列出 Keycloak 的设置步骤。

创建 ACM 证书

Data lake 的 Web UI Console 使用 HTTPS 服务，这也就是说必须使用 ACM 证书或者第三方的证书。具体使用方式，请参考：

如何上传 SSL 证书并将其导入 Amazon Identity and Access Management (IAM)

本部署指南以使用 Amazon Certificate Manager (ACM) 为例进行说明。关于 ACM 更多信息，请参考 https://aws.amazon.com/cn/certificate-manager/

首先，登录 ACM 控制台。

点击 【请求证书】按钮。

选择 【请求公有证书】单选框。

然后点击【请求证书】按钮。

在【添加域名】中的【域名】输入框，输入 data lake 服务所需要使用的域名。

点击【下一步】。

选择【DNS 方式验证】。

点击【审核】按钮。

请记录下 【名称】【类型】及【值】三个字段的值，这三个值需要在后面的 Route53 中添加记录集使用。点击【继续】按钮。

认证域名

在 Route 53 中添加 CNAME 记录，以认证该域名为您所有并可以使用。首先，进入 Route 53 的管理页面，点击你的域名。

如果没有创建托管区域，请参考 https://aws.amazon.com/cn/route53/

点击【创建记录集】，在右侧的窗口【名称】输入框输入在 ACM 验证步骤中记录的【名称】【类型】及【值】三个字段。

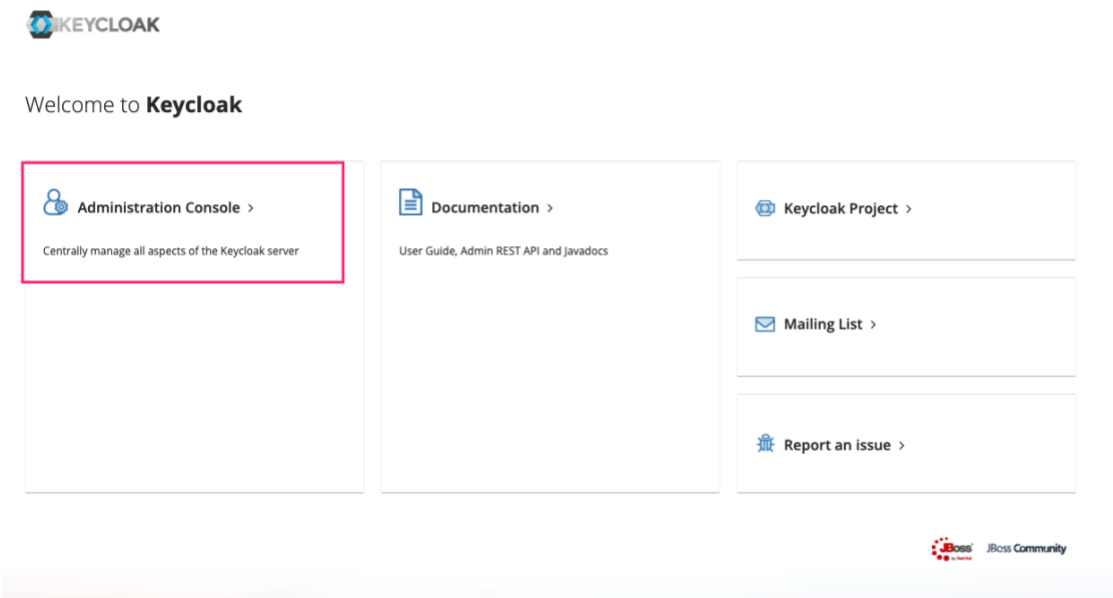
添加完成后。回到 ACM 界面等待大概 5 分钟。点击刷新按钮。等待 ACM 证书的状态变为 【已颁发】。

到此为止，ACM 证书已经申请完毕。

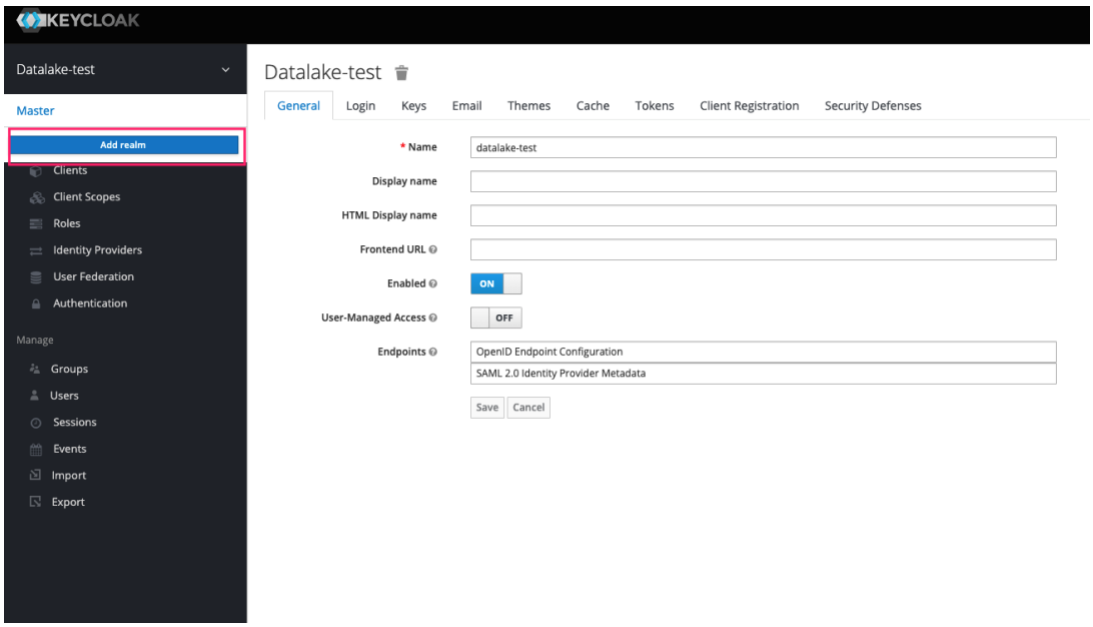
请记下上图橙色方框所标记的 ARN 字符串。该值将用于在 CloudFront 的设置中选择证书

Keycloak 设置

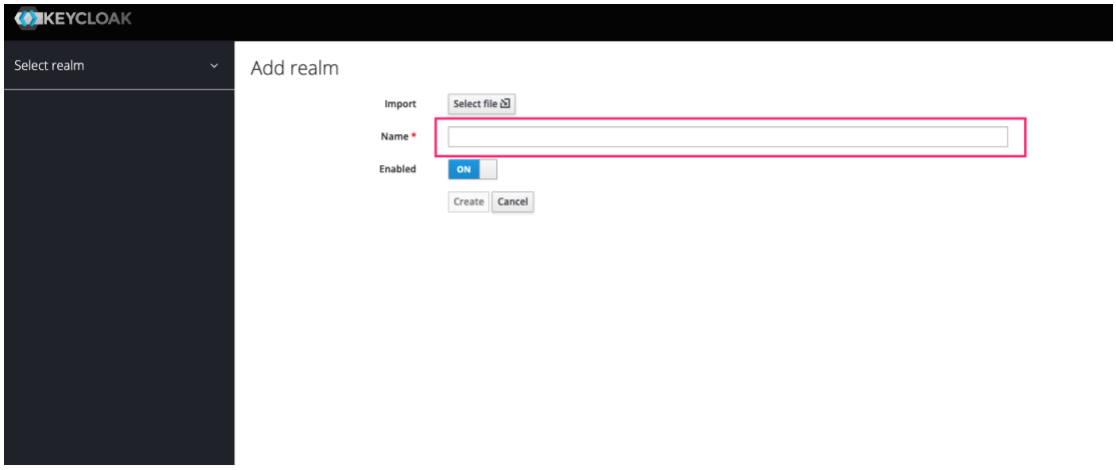
Keycloak 的设置需要管理员的权限登录到 Keycloak 系统。



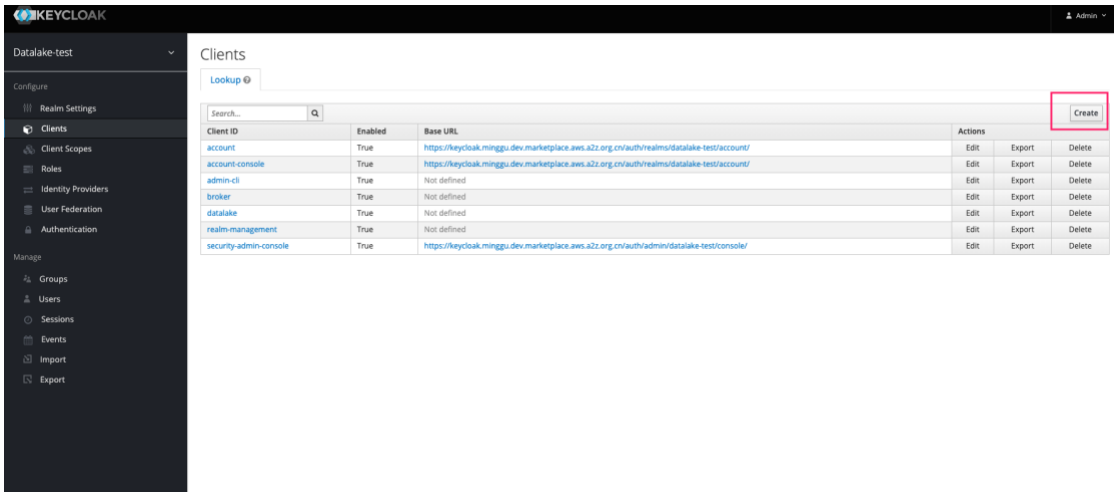
点击 Administration Console,输入 Admin 的 用户名和密码登录。



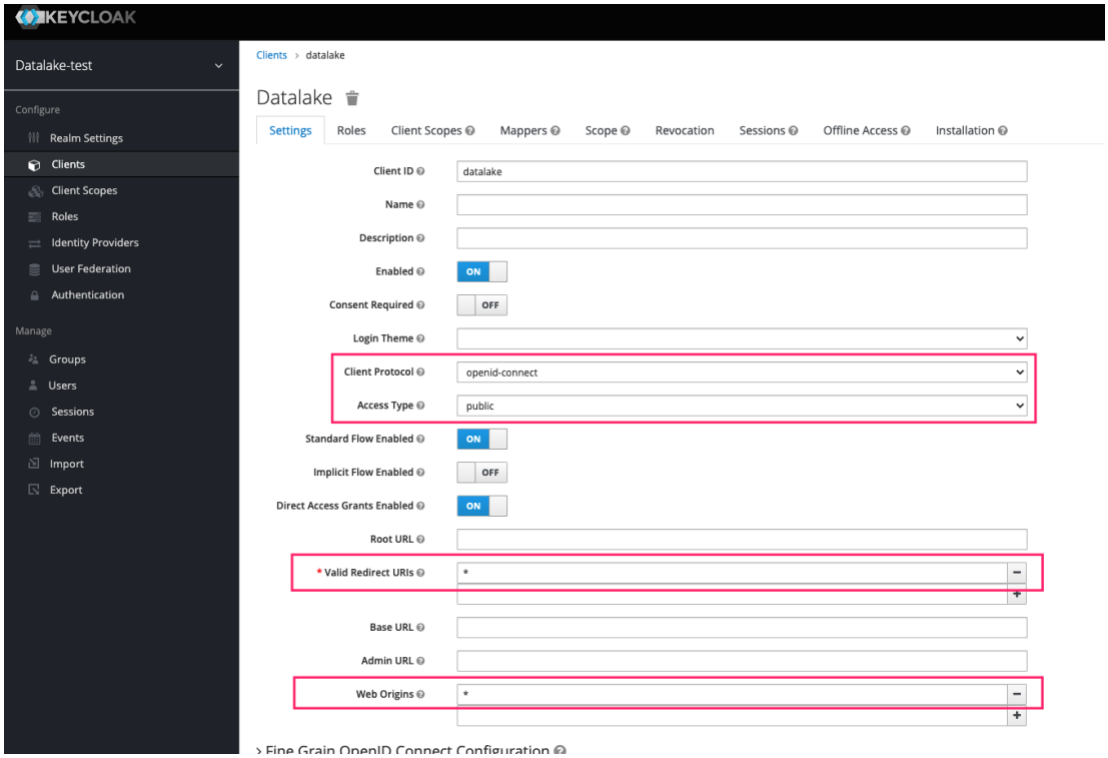
点击 Add realm



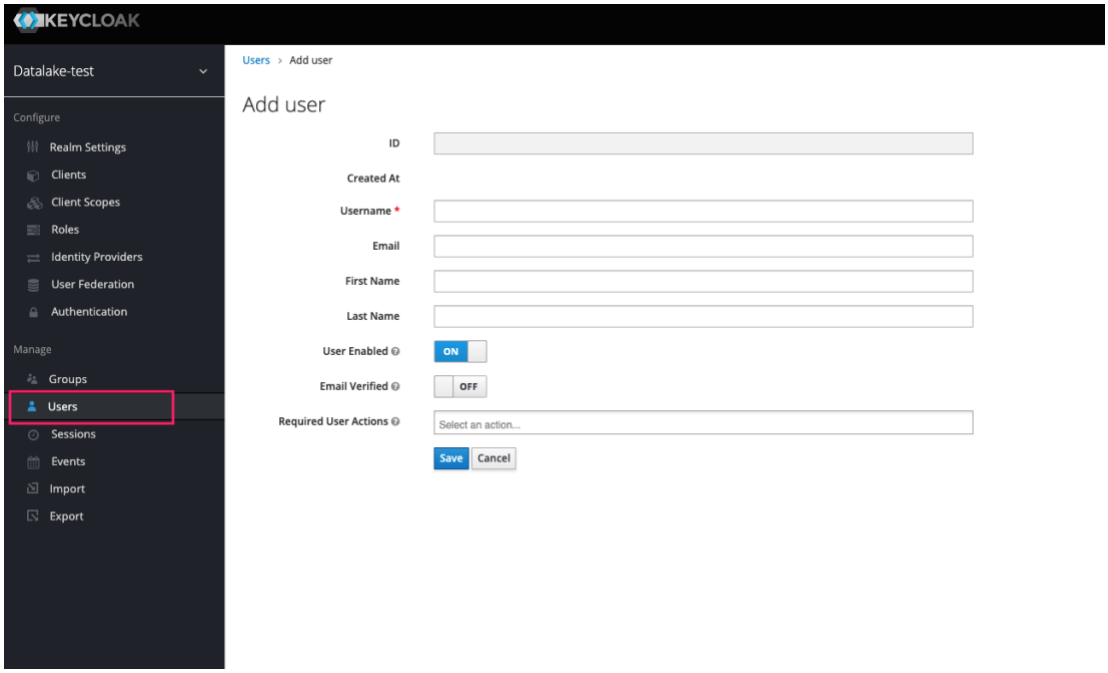
输入 Realm 的名字，比如 datalake-test



在 Keycloak 的 Clients 页面中创建一个新的 Client，点击图中的 Create 按钮



在 Client 对应的设置页面设置，注意 Valid Redirect URIs 这个值设置为*，Web Origins 设置为*.



点击 Users 页面在当前的 Realm 中创建用户，后续用户使用当前 Realm 中创建的用户来登录 data lake 的 Web UI Console。

部署 Data lake 的 CloudFormation 模版

Launch Stack 部署到北京区域

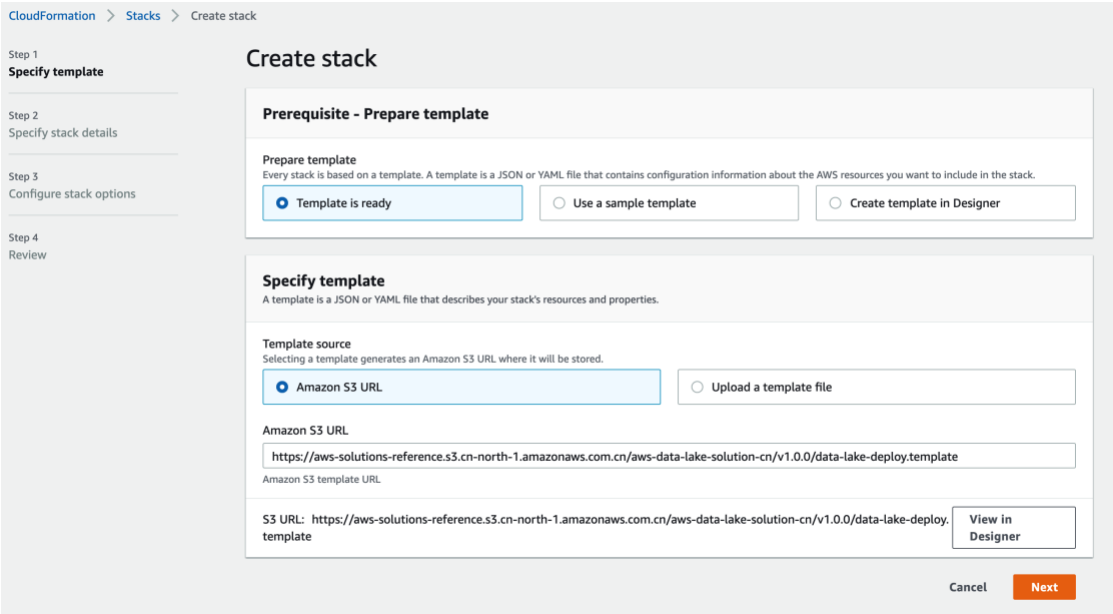
Launch Stack 部署到宁夏区域

如果您在 Amazon Web Services 海外区域部署 data lake 方案，请参考这个页面。

以及参考 Amazon Web Services 全球网站的使用说明。

1. 开始启动堆栈

点击宁夏和北京区域链接，打开 Amazon Web Services 管理控制台（如果还没登录会先跳转到登录页面，登录后进入模板启动页面）。默认情况下，此模板在宁夏区域启动，您同时可以使用控制台右上方的区域选择链接在其他的区域启动此方案。



2. 配置启动参数

点击下一步，进入详细参数的配置页面，Data lake 模版部署需要输入下面三个参数，具体的参数说明见下面的表格:

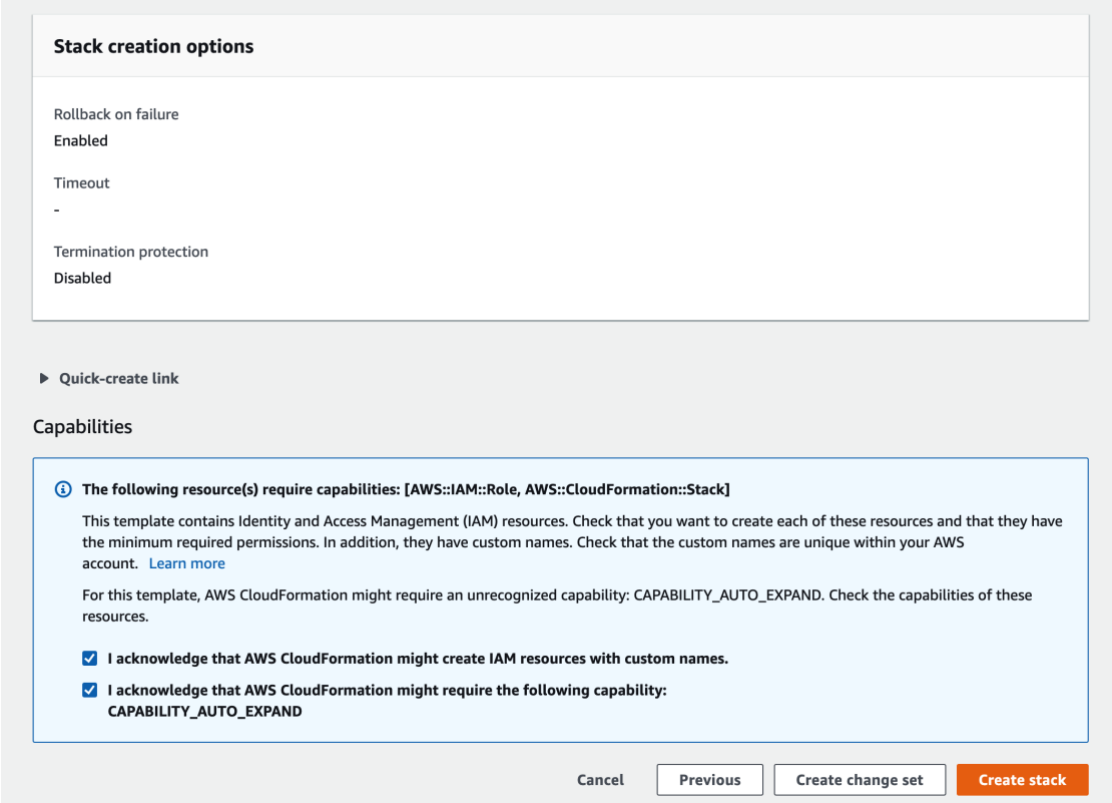
	参数	默认	描述
1	KeycloakClientId	No	Data Lake 需要作为 client 来访问对应的 Keycloak 认证 Server，需要先在 Keycloak server 端创建 clientid，然后把 clientid 的值拷贝到这个参数

2	KeycloakDomain	No	设置 Keycloak 认证 Server 对应的域名，例如: https://www.KeycloakServer.org, 注意需参数要带“https://”开头，域名的末尾不要添加”/“符号
3	KeycloakRealm	No	Keycloak 认证服务中对应的 Realm 的名字，一个 Keycloak 的 Server 可以包含多个 Realm，需要确保输入的 Realm 的名字是对应的用来认证 data lake 的 Realm。

参数设置完成后，点击下一步，并在下一步的配置堆栈选项保留默认，直接点击下一步。

3. 授权 IAM 并创建堆栈

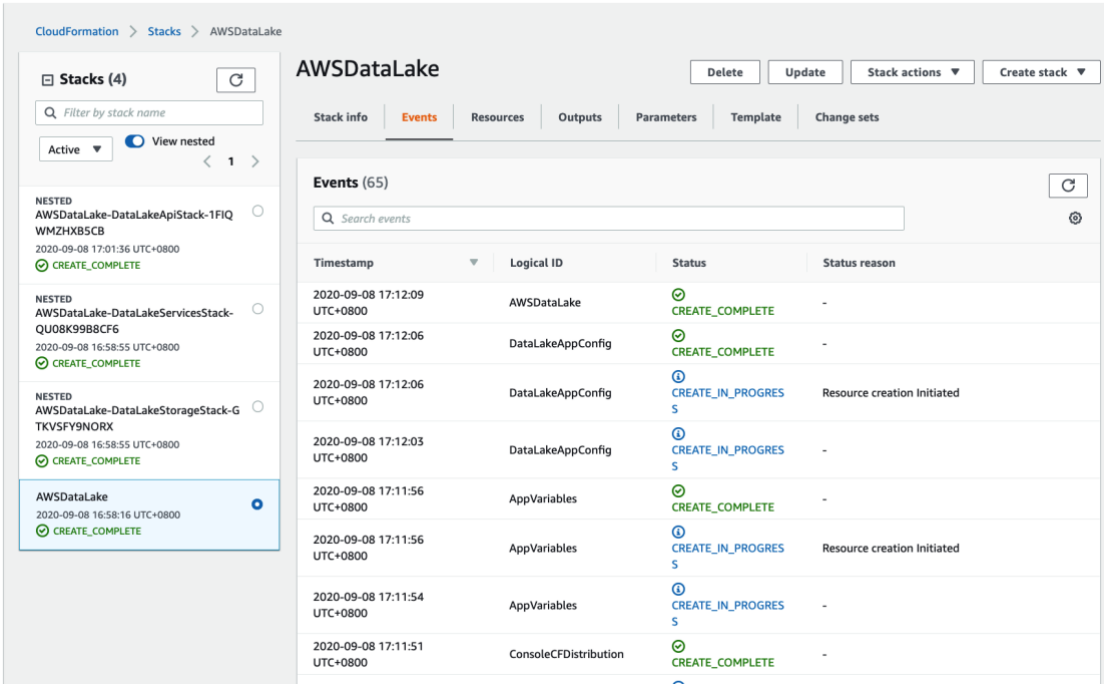
在最后的审核页面中，勾选最下方的【我确认，Amazon CloudFormation 可能创建 IAM 资源。】选择框，接下来点击【创建堆栈】，开始堆栈的创建。



4. 等待堆栈创建成功

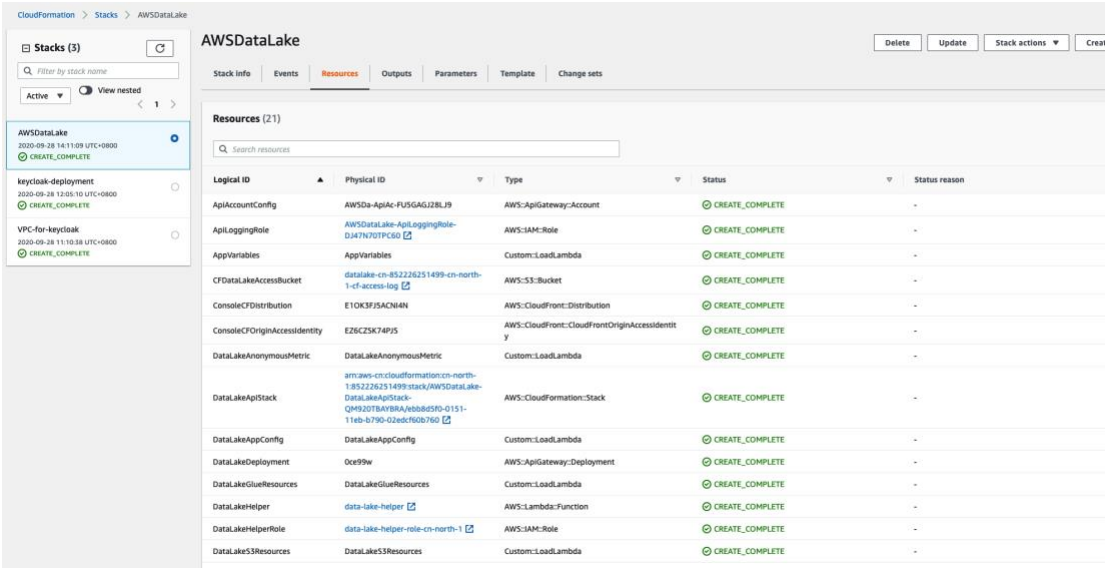
您可以在 Amazon CloudFormation 控制台的【状态】列中查看堆栈的状态，并点击右上方的刷新按钮更新状态。

大约三十分钟内，您可以看到堆栈状态变为 CREATE_COMPLETE，此时堆栈创建成功。

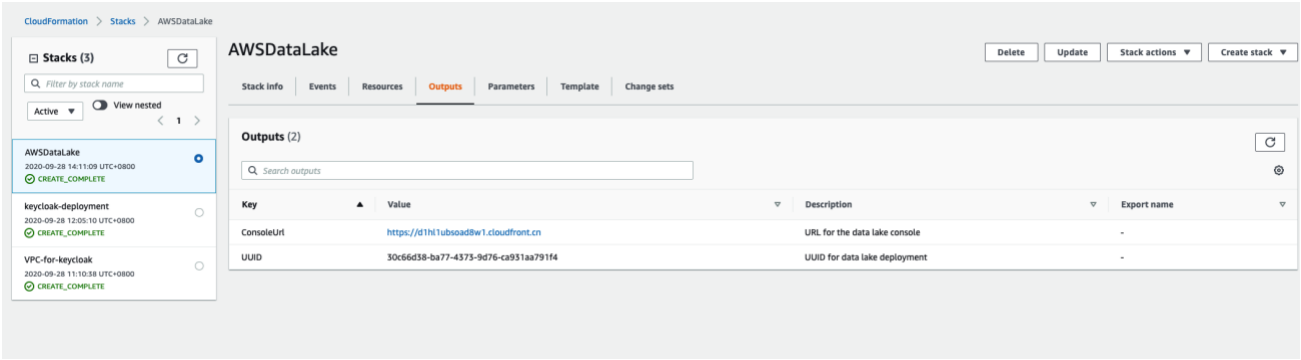


5. 查看堆栈资源

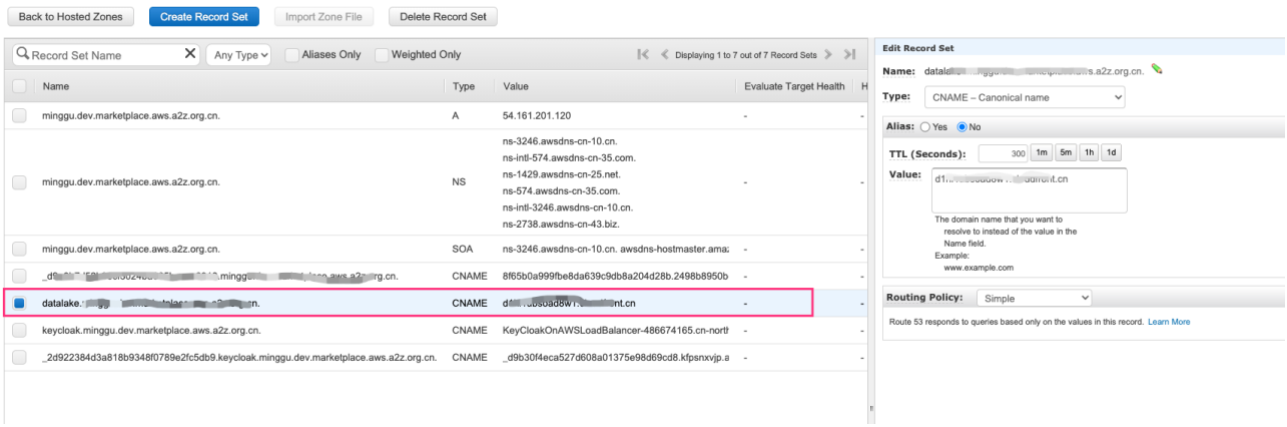
选中堆栈后短则【资源】标签，可以看到这个堆栈启动的所有资源，如果您感兴趣可以一一点进去查看，了解每一部分的功能。



部署完成后，Cloudformation 的 Output 窗口会输出生成的 data lake 的 web UI console 所对应的 CloudFront 的 URL。

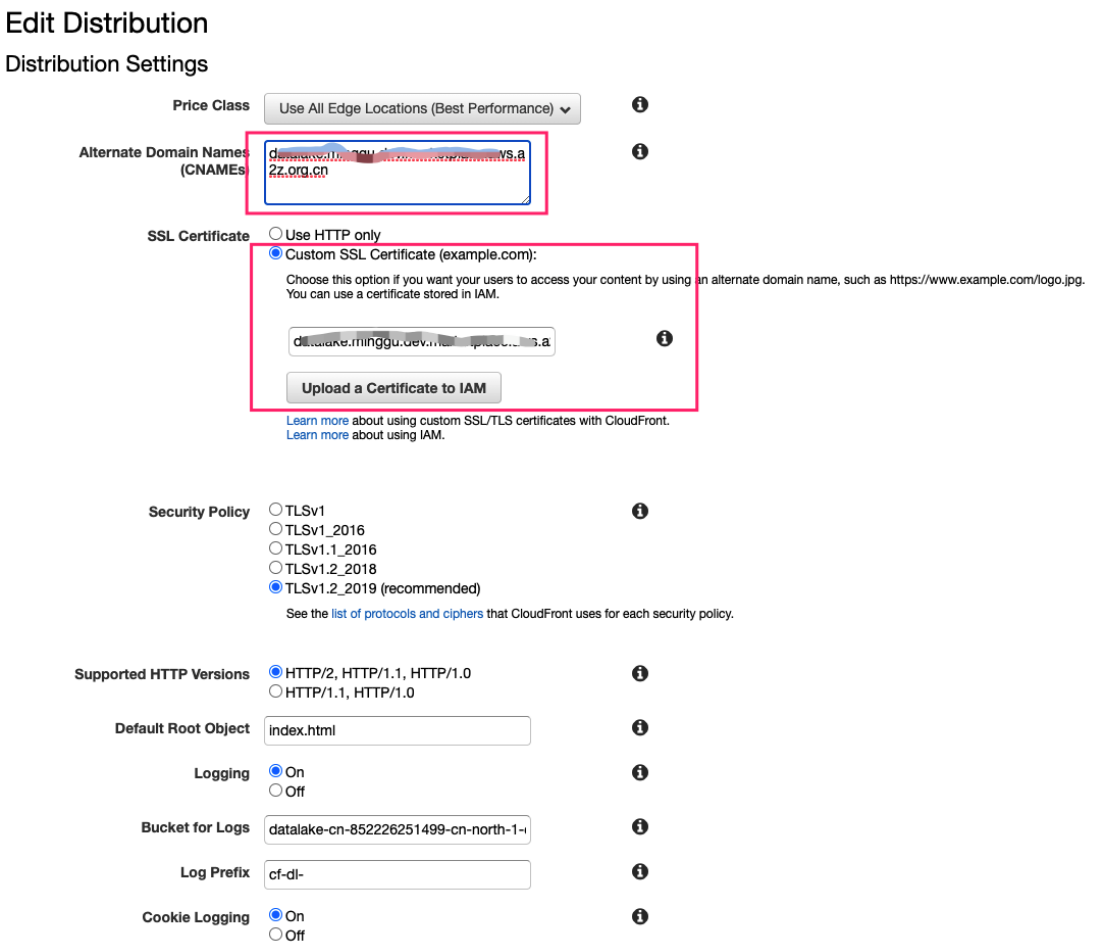


进入 Route53 的 Console，在 Hosted Zone 下面创建一条 CNAME 的 Record Set，Value 的值为 CloudFront 对应的 URL，注意 CloudFront 的地址不要带 https://。

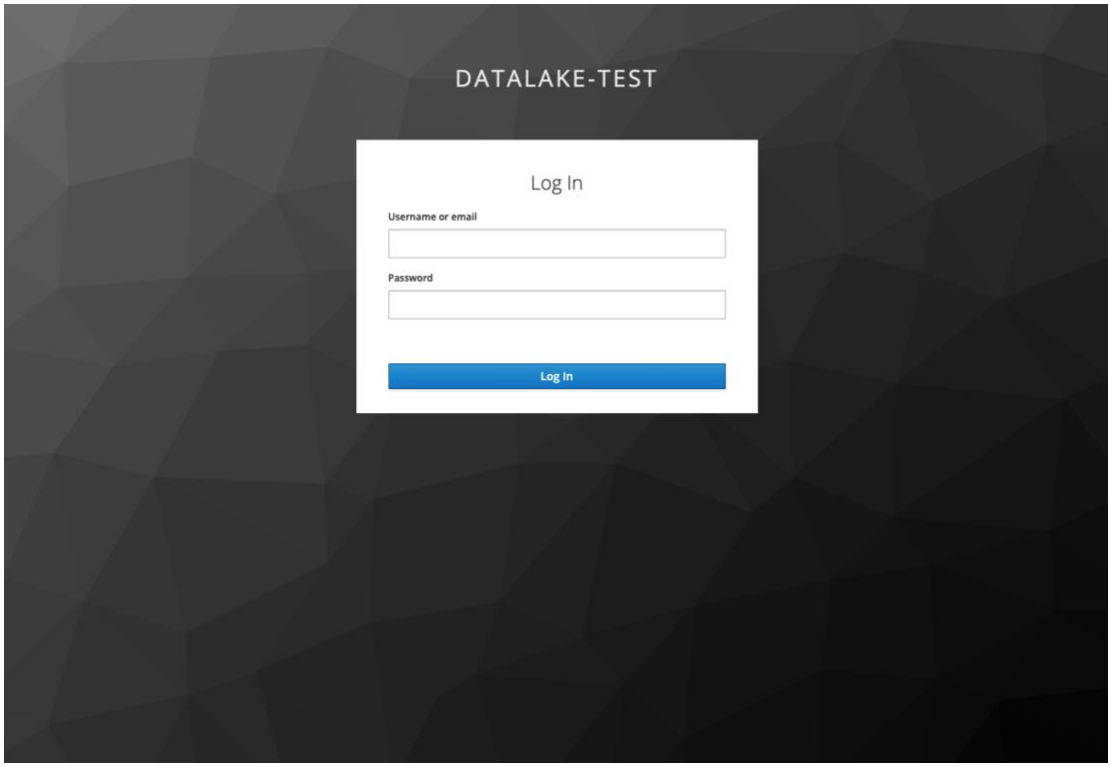


进入 CloudFront 的 Console，点击 data lake 对应的 CloudFront 的记录，点击 Edit，将 ICP 备案的域名填入“Alternate Domain Names”，在 SSL Certificate 选项中选择 ICP 对应的证书或者点击”Upload a Certificate to IAM”来上传第三方的证书。

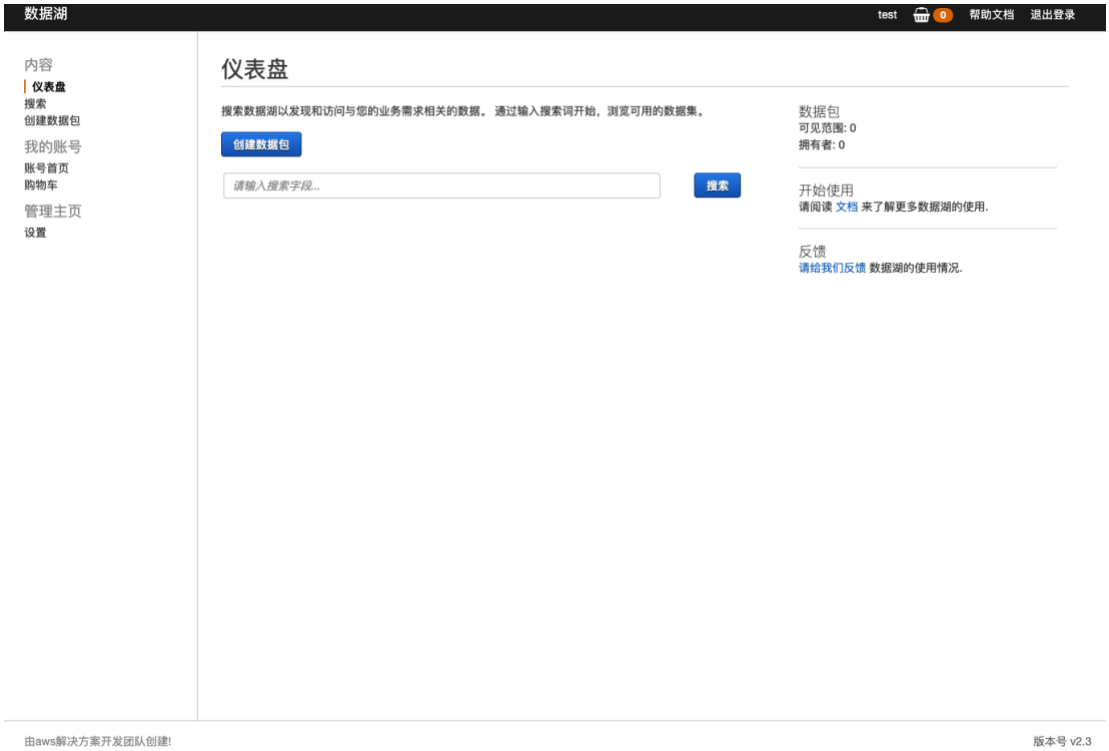
(注意上传第三方的证书到 IAM 之后需要等五分钟左右才能在 Custom SSL Certificate 列表中显示)。



到此为止，Data lake 的部署和设置就完成了，可以通过访问 ICP 对应的域名来访问 data lake 的 Web UI Console，浏览器会自动跳转到对应的 Keycloak 的 domain 来进行用户登录：



用户输入用户名和密码后会显示 data lake 的 Web Console 如下：



到此说明 Data Lake 的部署和设置已经全部完成，可以使用 Web UI Console 来使用 data lake 提供的功能了。