Salesforce.org Easy Datalake User Guide

Prepared for Salesforce by the Envision Engineering Team
May 2022

EnvisionEngineering

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Contents

1 Introduction

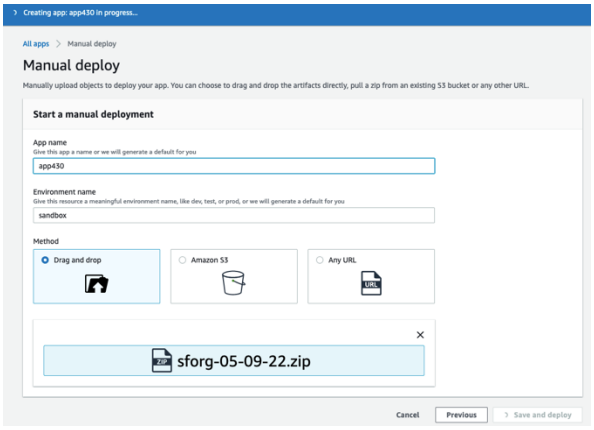
This document provides guidance for setting up a datalake with the EZ Datalake App. Refer to the EZ Datalake Handover document for instructions on building and deploying the Easy Datalake PoC code in an AWS account.

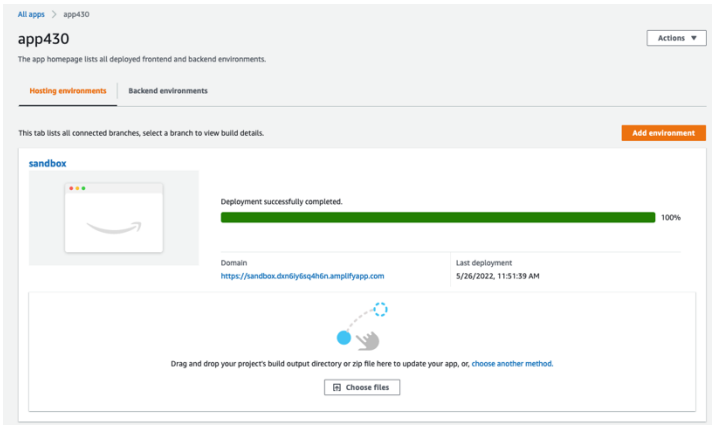
2 Setting up the App

The EZ Datalake is self-hosted within the NPO AWS Account. For this, the app needs to be first deployed as a static website using AWS Amplify.

Sign in to the AWS Management Console and open the Amplify console. Manually deploy the app following the steps here: https://docs.aws.amazon.com/amplify/latest/userguide/manual-deploys.html#drag-and-drop

The zip file to upload with the manual deployment should be generated with the build process outlined in the EZ Datalake Handover document. Alternatively, you can refer to instructions in the code on generating the build for manual deployment. This build should be made available to users in GitHub to download.





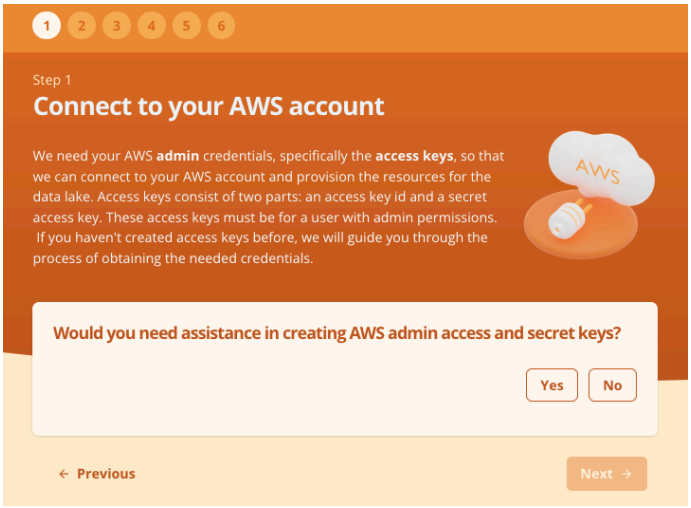
It typically takes a few minutes to deploy the app after which a URL to the EZ Datalake app is presented. Click on the URL to load the EZ Datalake app and set up the data lake and import data in 6 steps.

The image below shows the URL that was generated in this instance and the welcome screen showing the user what to expect with an outline of the 6 steps to be completed through the wizard like UI.

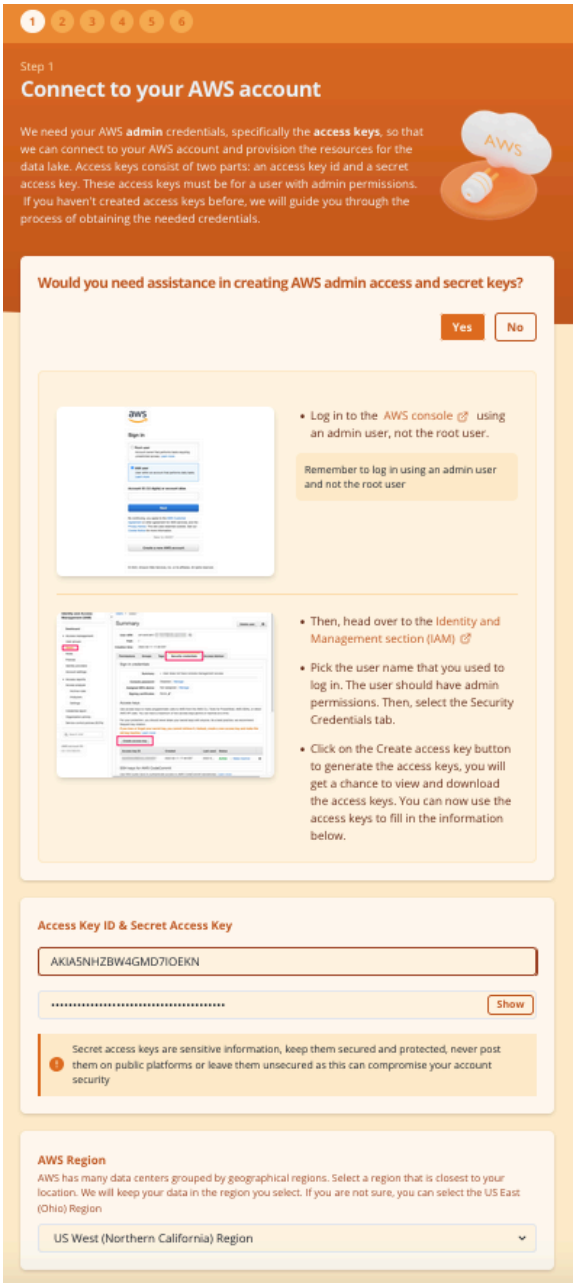
3 Datalake Setup in 6 Steps

3.1 Step 1

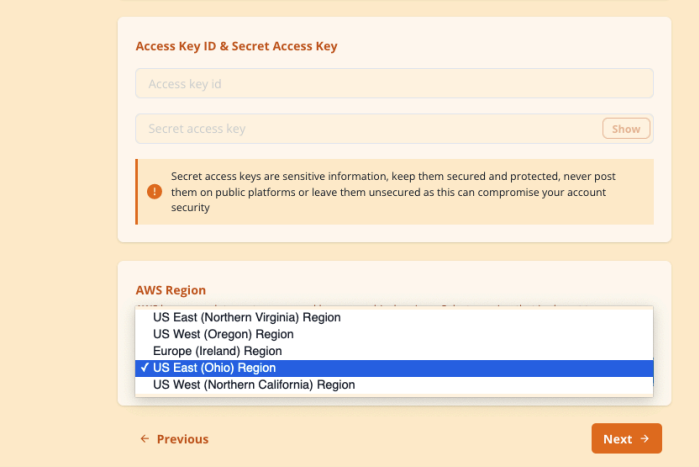
In Step 1, we connect to the NPO User's AWS Account. The EZ Datalake App requires Access Key ID & Secret Access key for an IAM Admin user.



If the user has never created these, the app shows guidance on how to login to AWS Console and use Identity and Access Management (IAM) to pick the IAM user with admin permissions. The UI then guides the user to create and download the access keys. Once created, the access key information is used to complete Step 1.



At the bottom of Step 1, there is a drop-down box to select the region where the data lake should be provisioned.

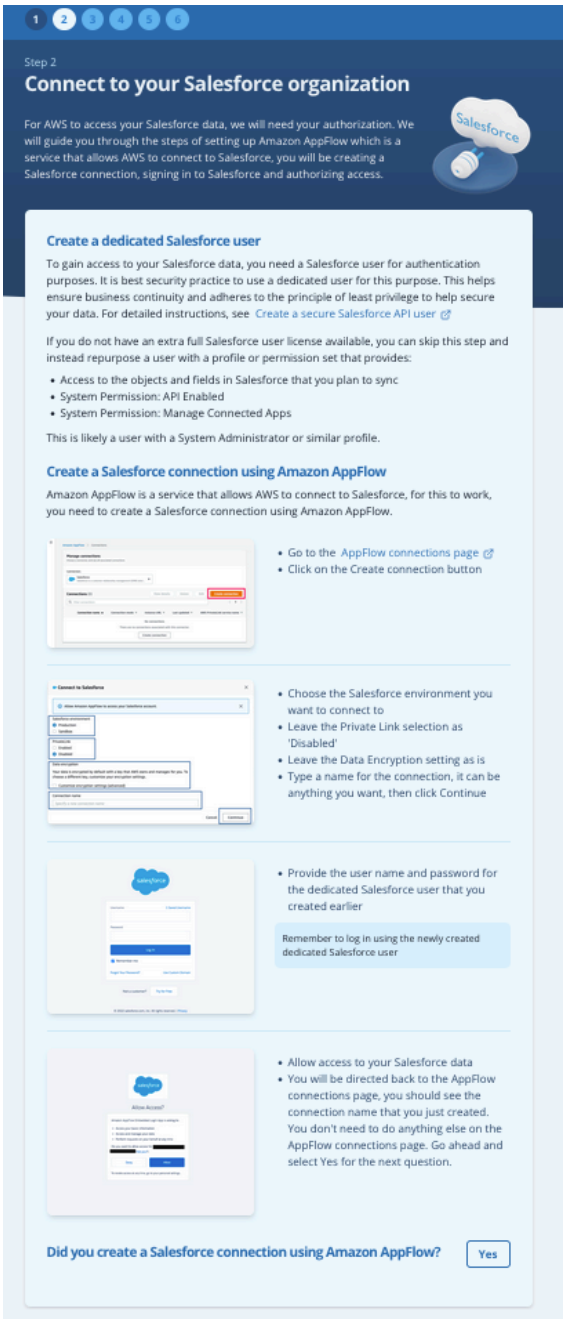


3.2 Step 2

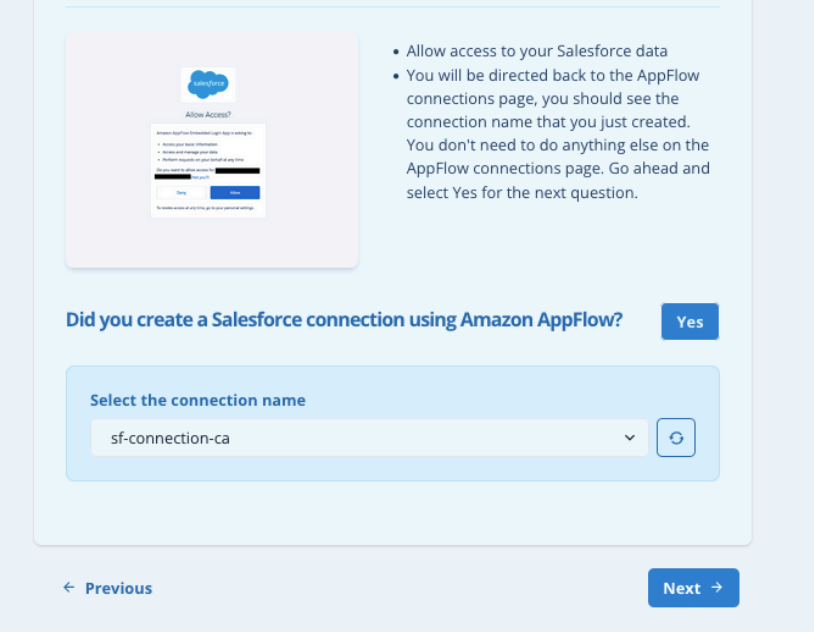
In Step 2, the EZ Datalake app will gather credentials to connect to the NPO Salesforce Organization. For AWS to access the NPO's Salesforce data, we will need user authorization. For this, the app gives guidance to the user to create a dedicated Salesforce user with the necessary profile or permission set that provides:

- Access to the objects and fields in Salesforce that you plan to sync
- System Permission: API Enabled
- System Permission: Manage Connected Apps

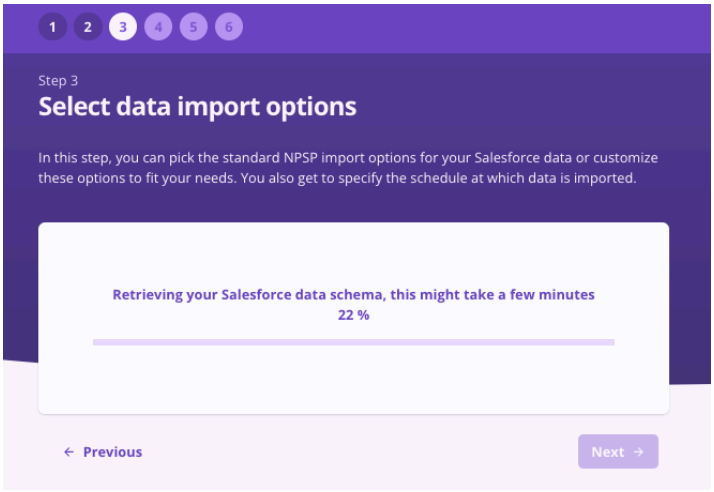
This is likely a user with a System Administrator or similar profile. Additionally, the app guides the user to setup Amazon AppFlow which is a service that allows AWS to connect to Salesforce. The user is guided to create a Salesforce connection, sign in to Salesforce and authorize access. The image below shows this in-app guidance.



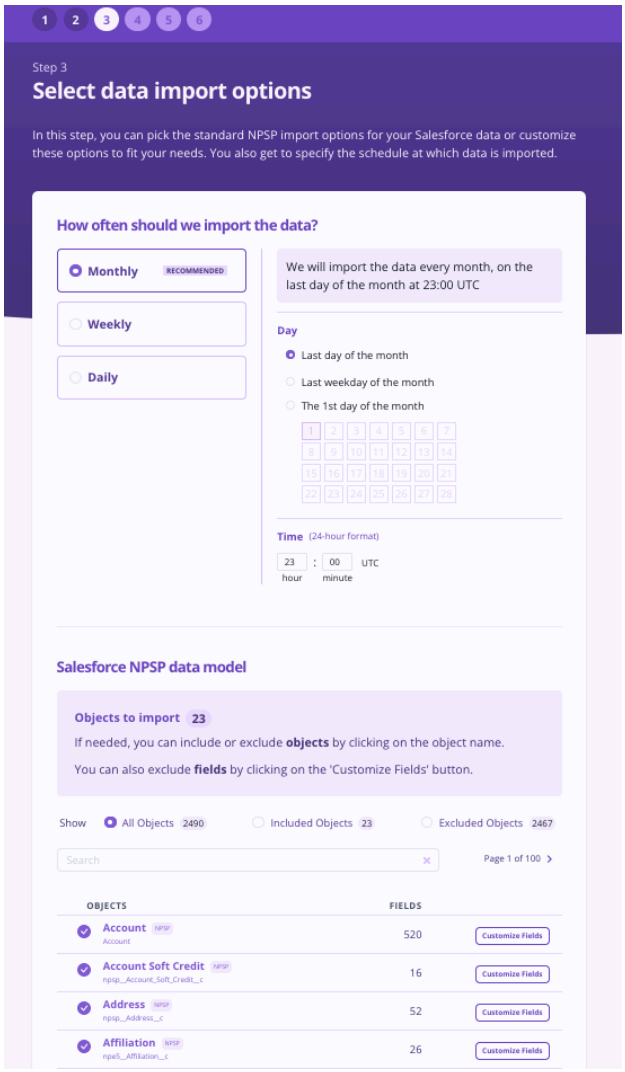
Once the Salesforce connection is created through Amazon AppFlow, a drop-down box will populate the list of connections that can be selected for use. Connections are region-specific and the list will show only the connections that are created in the selected region. Once a connection is selected from the drop-down box, the NPO User can proceed to Step 3.

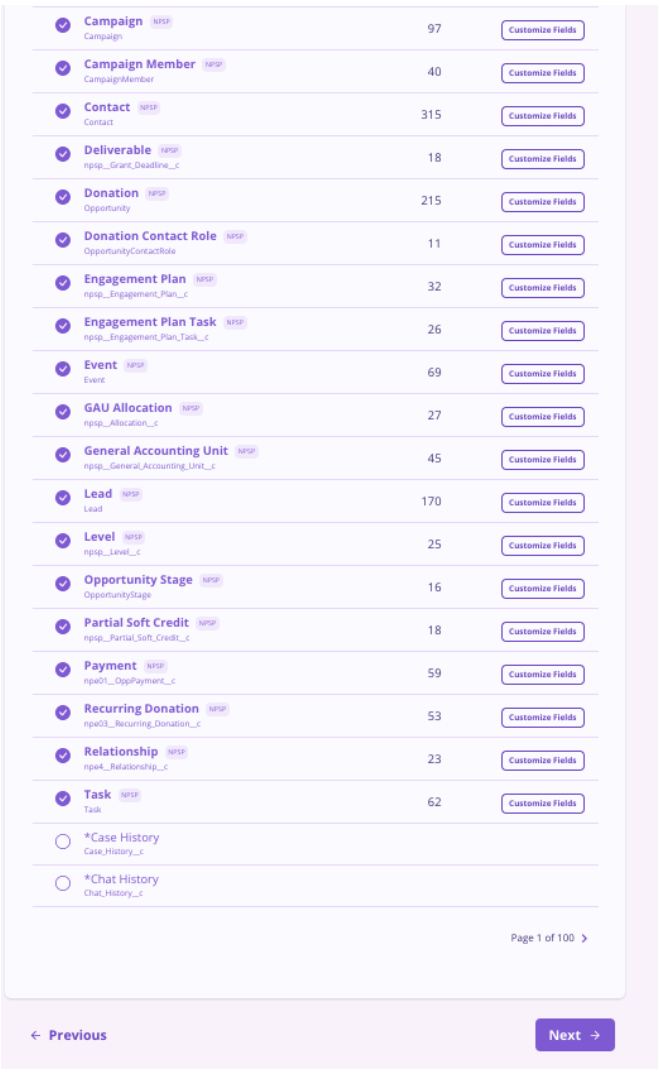


3.3 Step 3

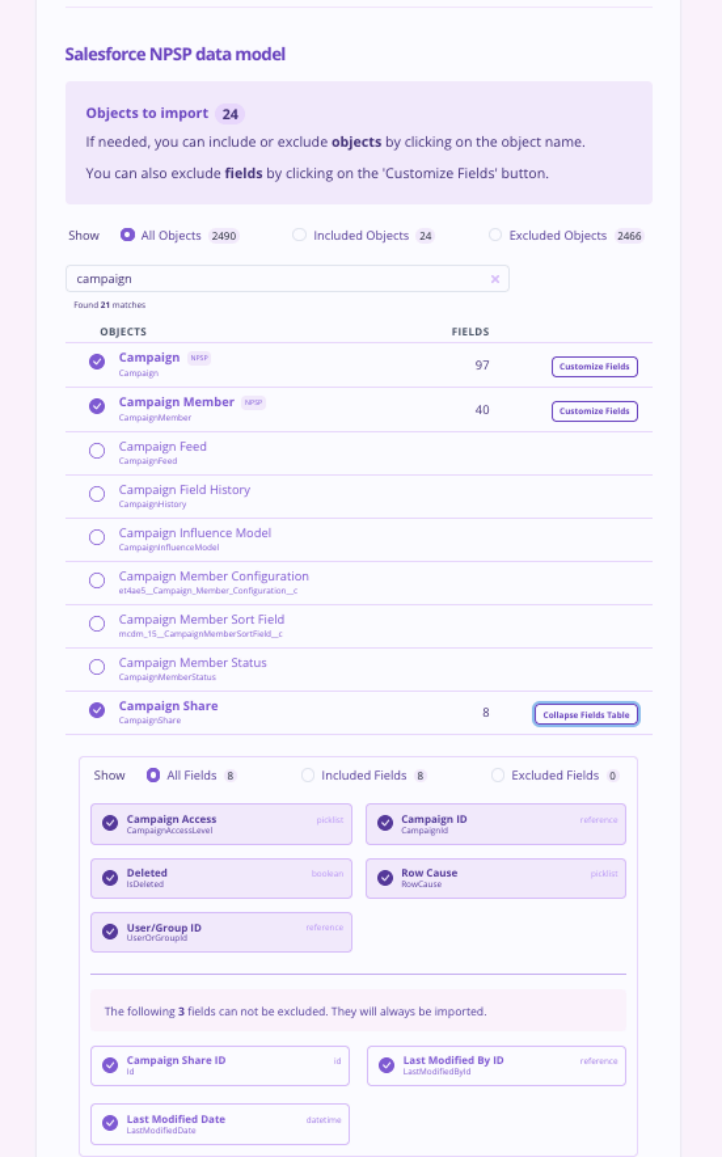


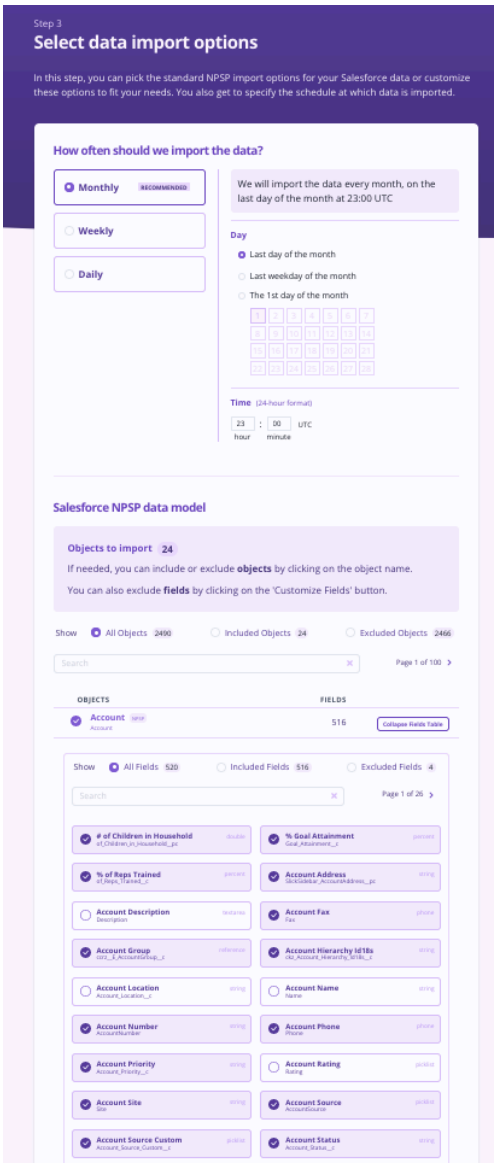
During step 3, Salesforce data schema is retrieved to show the NPSP data model. An NPO organization can have standard objects, core NPSP objects as well as custom objects. The retrieved schema is used to show the list of objects to the NPO user to be included or excluded as desired.





In Step 3, the NPO User can click on default options to setup the datalake and periodic data import. A typical NPO User is expected to click on all default options to proceed to Step 4. However, an advanced NPO User can customize the objects and fields to be included or excluded in the data import. The image on the left below shows a custom object being added for import and the image on the right, shows fields being excluded from the Account object.

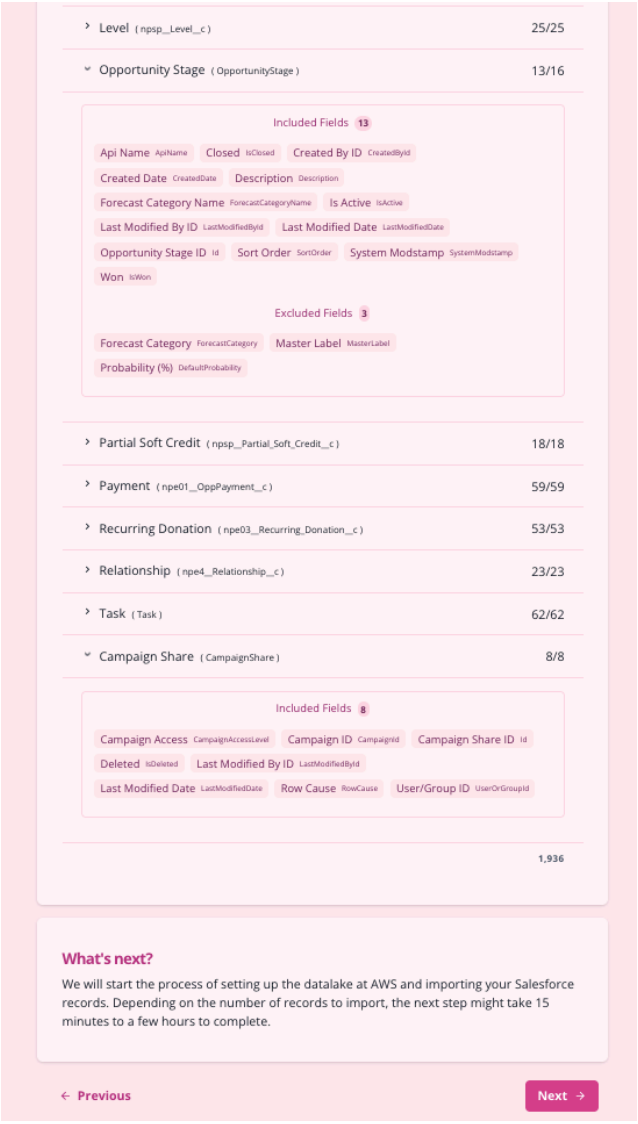




A default monthly sync is recommended. However, the user can select a weekly or even a daily sync and choose the date and the time of the sync. Note that we don't recommend a daily sync option for a large NPO Organization as this could incur additional costs.

3.4 Step 4

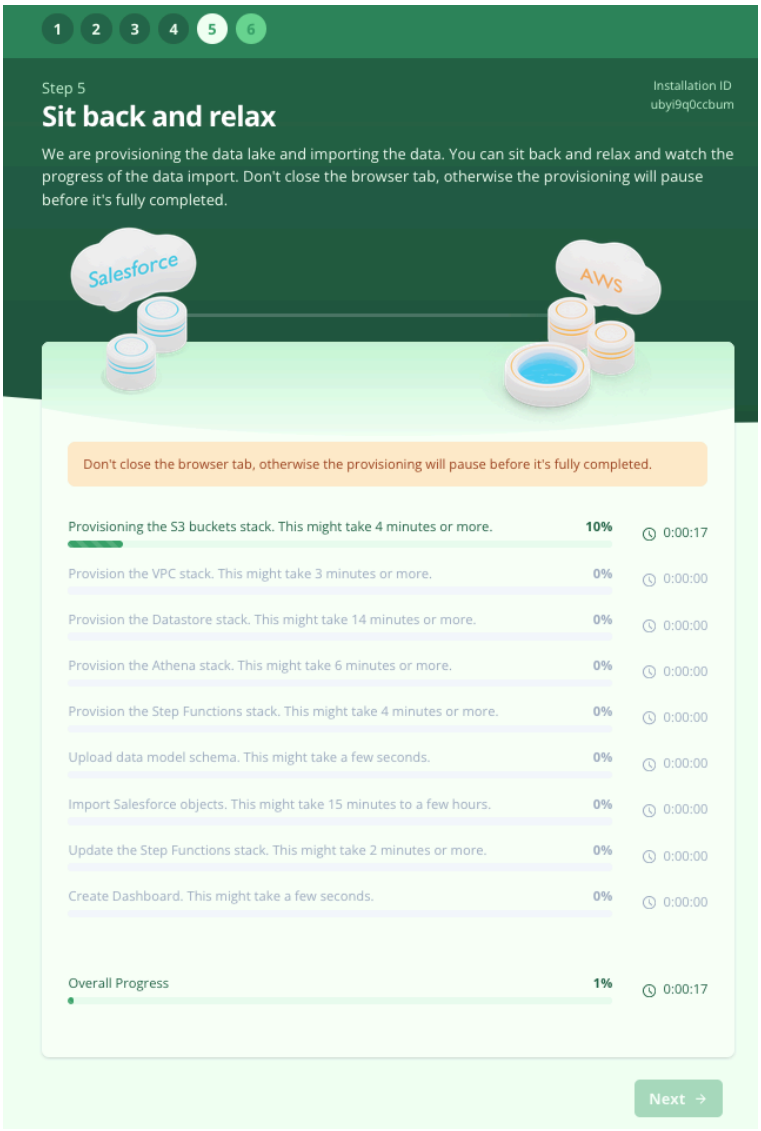
In Step 4, the user can review the data lake setup and data import options and confirm the changes they've done (if any). This step is where the NPO User is allowed to go back to previous step to make any other needed changes before confirming the setup of resources for the data lake.



Once review is complete, the NPO User can click “Next” to start provisioning the resources in Step 5.

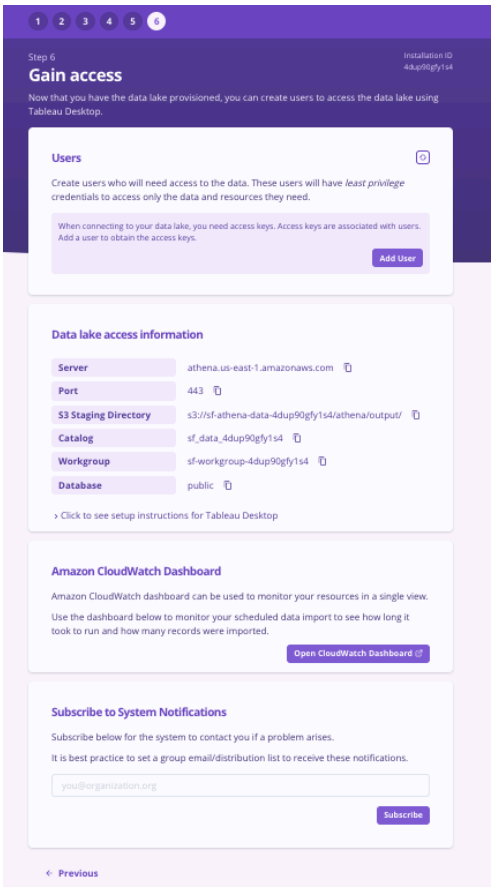
3.5 Step 5

Step 5 is where the data lake is provisioned and the data is imported at the end. This can take several minutes to several hours depending on the size of the data in the NPO Organization. We caution the user to leave the EZ Datalake app running in the browser while the data lake is being provisioned. However, there is a resume feature described in Section 4.1 where a user can resume the data lake provisioning if the browser is accidentally closed.

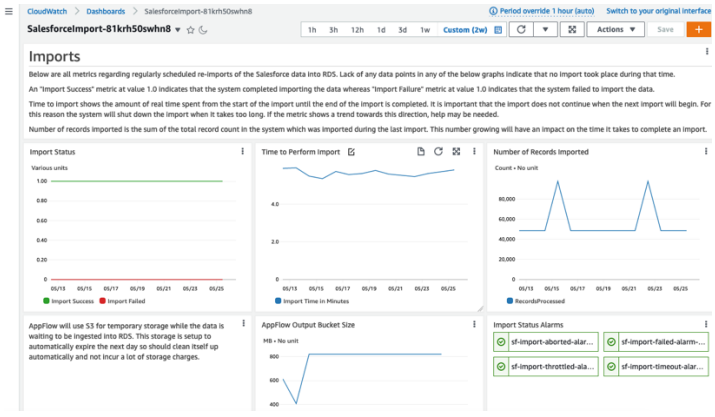


3.6 Step 6

Once the data lake is setup, in Step 6, the NPO Users can create and manage IAM users to grant them access to Tableau Desktop. These users will have *least privilege* credentials to access only the data and resources they need. In Step 6, the user can also view the Data Lake access information and instructions for Tableau Desktop installation and access for Windows and Mac, access CloudWatch Dashboard and subscribe to systems notifications.



Amazon CloudWatch dashboard can be used to monitor your resources in a single view. The screenshot below shows metrics for the user to monitor the scheduled data import to see how long it took to run and how many records were imported.



Additionally, users can subscribe to system notifications to be alerted if a problem arises.

4 Resume Feature

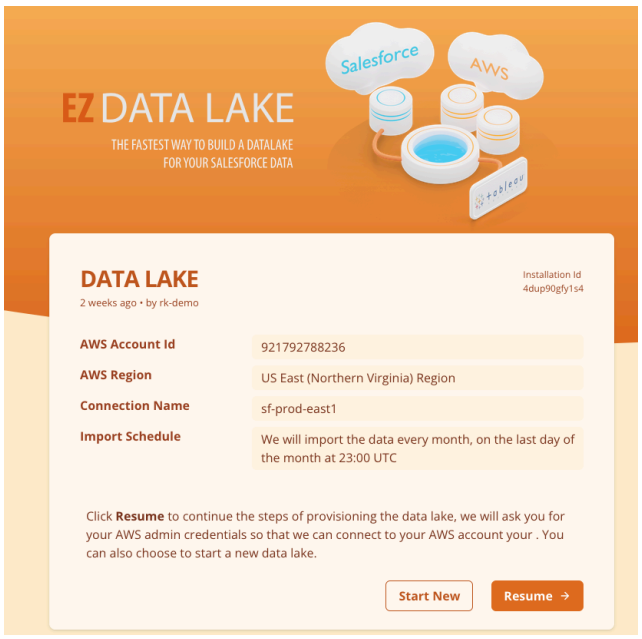
The resume feature allows an NPO System Administrator to continue deployment if the app was closed during any of the five initial steps. Additionally, the resume feature can also be used to retrieve the Data Lake provisioning information after deployment and also a start a new datalake deployment if that is desired.

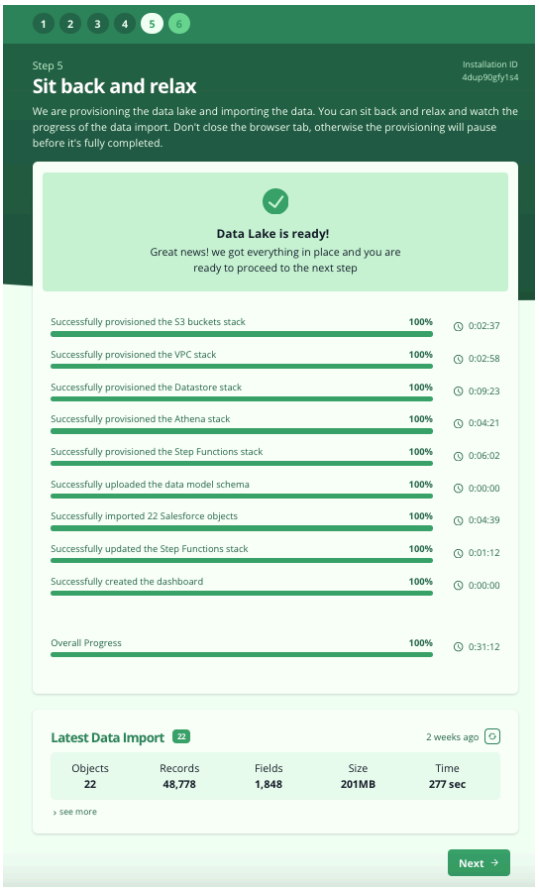
4.1 Resume while deployment is in progress

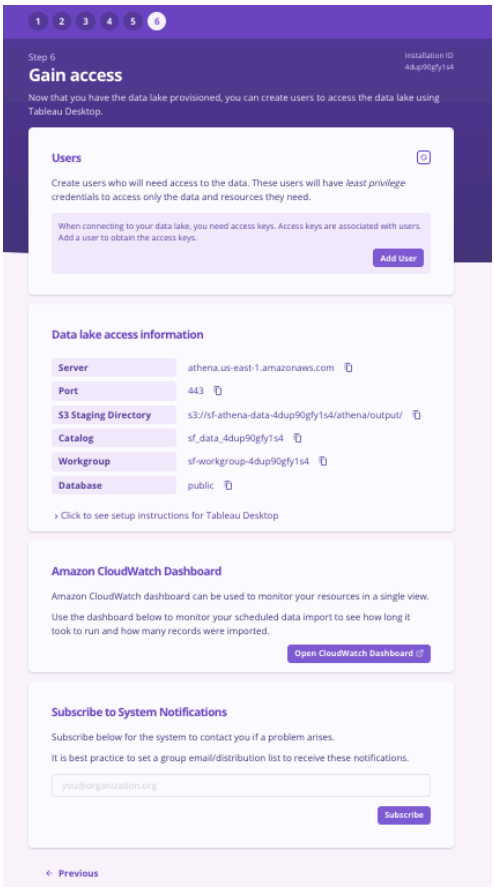
During any of the six steps with the EZ Datalake App, it is recommended that the NPO User keep the EZ Datalake app in the browser running till the datalake is deployed. However, if the user does accidentally close the browser, the “resume” feature can help pick up where the deployment was left off. Note that depending on which step the user is in when the browser is closed, the app will resume in the same step.

4.2 Resume after deployment

Once the datalake is deployed, the “resume” feature can be used for the NPO User to retrieve details on the datalake that was setup earlier. Additionally, the NPO System Administrator can use the app to create and manage IAM users to grant them access to Tableau Desktop. These users will have *least privilege* credentials to access only the data and resources they need. The screenshots below show how the NPO System Administrator can enter their IAM Admin Key ID and Secret Access Key to retrieve the details of the datalake from Step 6 and manage users for Tableau Desktop access, view Data Lake access information and retrieve instructions for Tableau Desktop, access CloudWatch Dashboard and subscribe to systems notifications in Step 6.







4.3 Resume another datalake

Typically, an NPO User is expected to setup a single datalake. If for any reason, a new datalake needs to be setup, the same app can be utilized to deploy another datalake. This can be useful perhaps if a datalake for a sandbox needs to be setup first before moving to production organization. The “resume” feature can then be used to view and manage the multiple datalake as show below in the screenshots.

5 Data Consistency / Data Deletion

Data can be deleted in Salesforce in multiple ways triggering soft deletes and hard deletes. The EZ Datalake app ensures data consistency with the data in the Salesforce by performing full data import on each periodic sync operation. Refer to EZ_Datalake_Testing_Deletion_Scenarios.doc where different deletion scenarios are detailed along with testing outcomes.