**March 2018**



**Intelligent Data Lake Workshop**

*Lab 0 - Prerequisite, building a Data lake*

Table of Contents

[Overview 3](#_Toc506992481)

[Create the Data Lake: 4](#_Toc506992482)

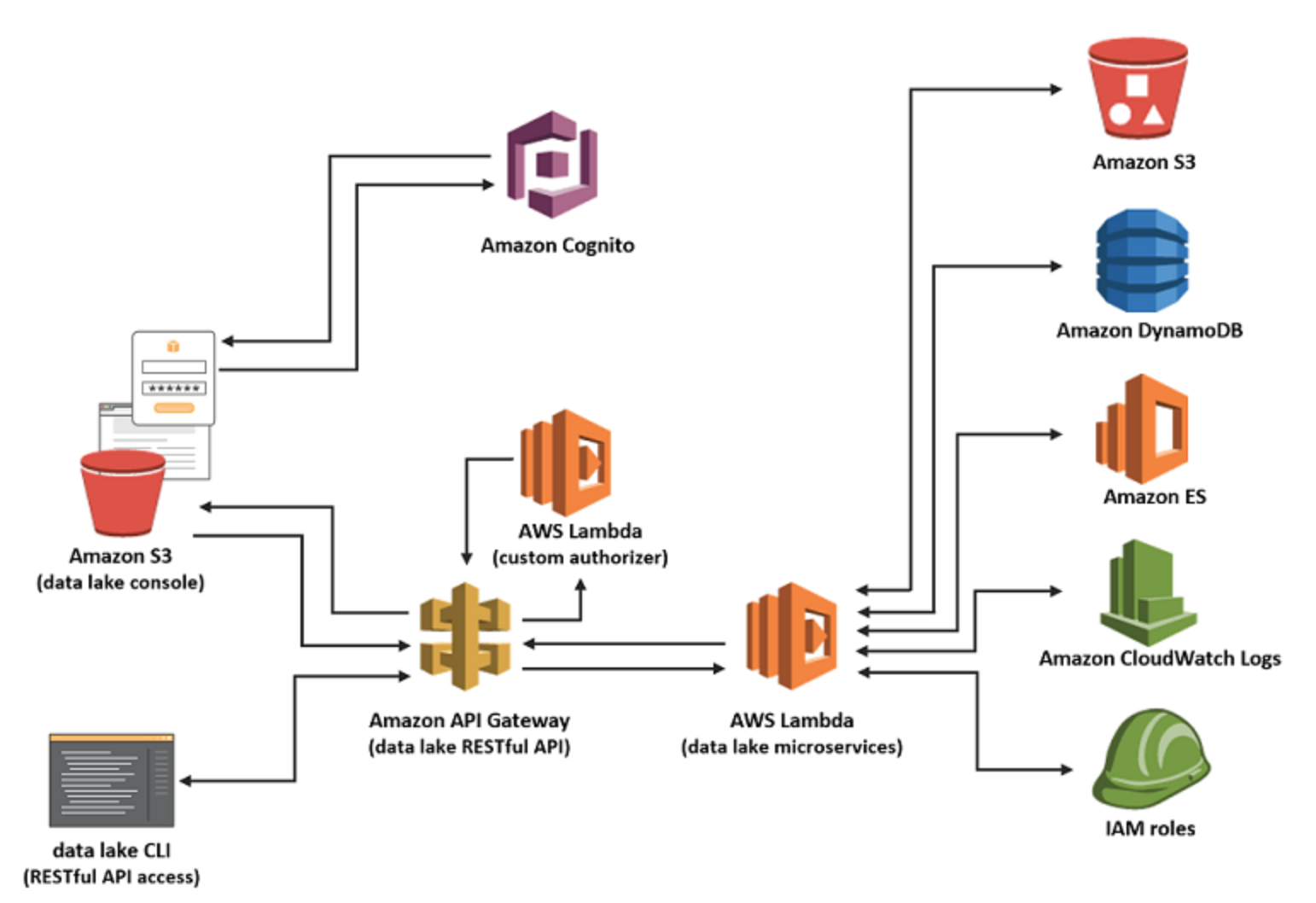
[How to use the Data lake 7](#_Toc506992483)

[Conclusion 9](#_Toc506992484)

# Overview

In this lab, we will focus on the basics of a Data Lake architecture on AWS and create one using the CloudFormation (CF) template. The Data Lake you will build in this lab will enable your users with the following functionalities –

1. Create data packages with meta tags for easy search.
2. Search packages with the keywords.
3. Download the contents of the packages from S3.
4. User management and governance.

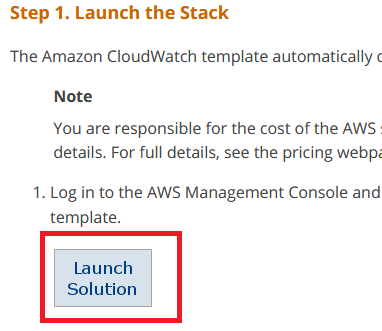


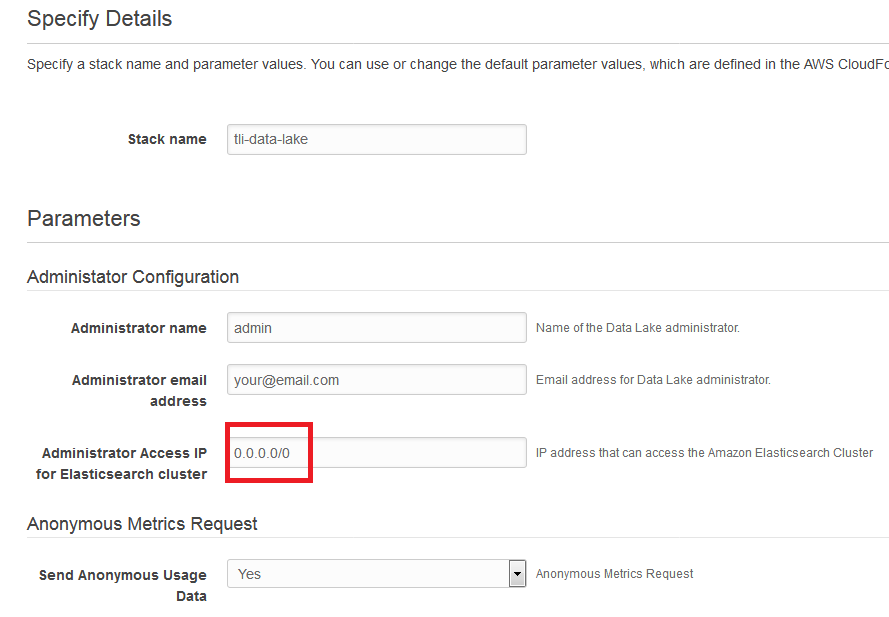
# Create the Data Lake:

1. Please use Chrome or Firefox browser to ensure smooth lab experience.
2. Sign into the AWS Management Console <https://console.aws.amazon.com/>.
3. In the upper-right corner of the AWS Management Console, confirm you are in the desired AWS region (e.g., N. Virginia).

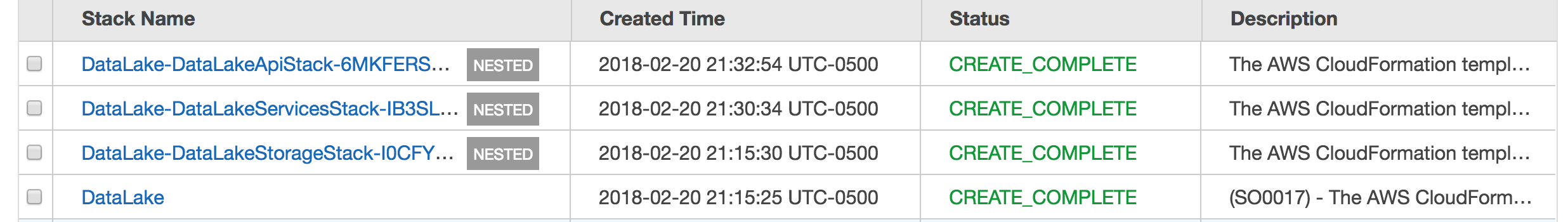
Note, be sure to select a region such as N. Virgina or Oregon that offers Amazon Kinesis Analytics. Also, make sure you are doing all the labs in the same region.

1. In a new browser window, go to the Automated Deployment page located at <https://docs.aws.amazon.com/solutions/latest/data-lake-solution/deployment.html>
2. In the section labeled ‘Step 1. Launch the Stack’, click on the ‘**Launch Solution**’ button, which will open the CloudFormation console in your AWS account.



1. Provide a Stack name such as “**YourInitials-data-lake**”. In “**Administrator Access IP for Elasticsearch cluster**” field, be sure to enter “**0.0.0.0/0**”. In “**Administrator email address**” field, be sure to enter a valid email address, CloudFormation will send the temporary administrator password to the email and make sure to check junk folder. 
2. Click **Next.**  On the **Options** page, you can specify tags (key-value pairs) for resources in your stack and set additional options, and then click **Next**.
3. On the **Review** page, review and confirm the settings. Be sure to check the box acknowledging that the template will create AWS Identity and Access Management (IAM) resources with custom names.
4. Choose **Create** to deploy the stack.

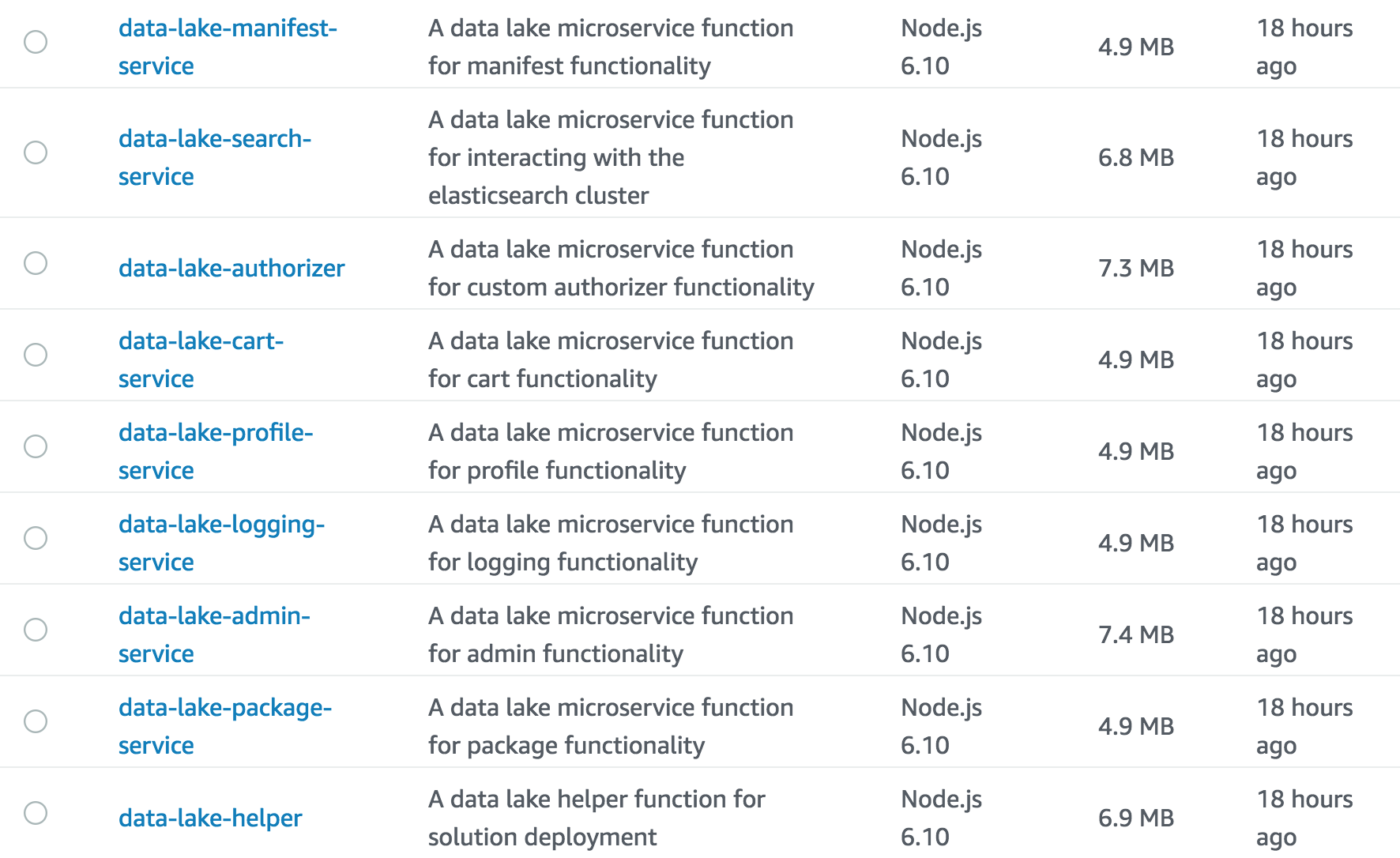
After the stack launches, there are total of **four** stacks launched in the same AWS Region. Once all of the stacks and stack resources have successfully launched, you will see the message **CREATE\_COMPLETE** all four stacks.



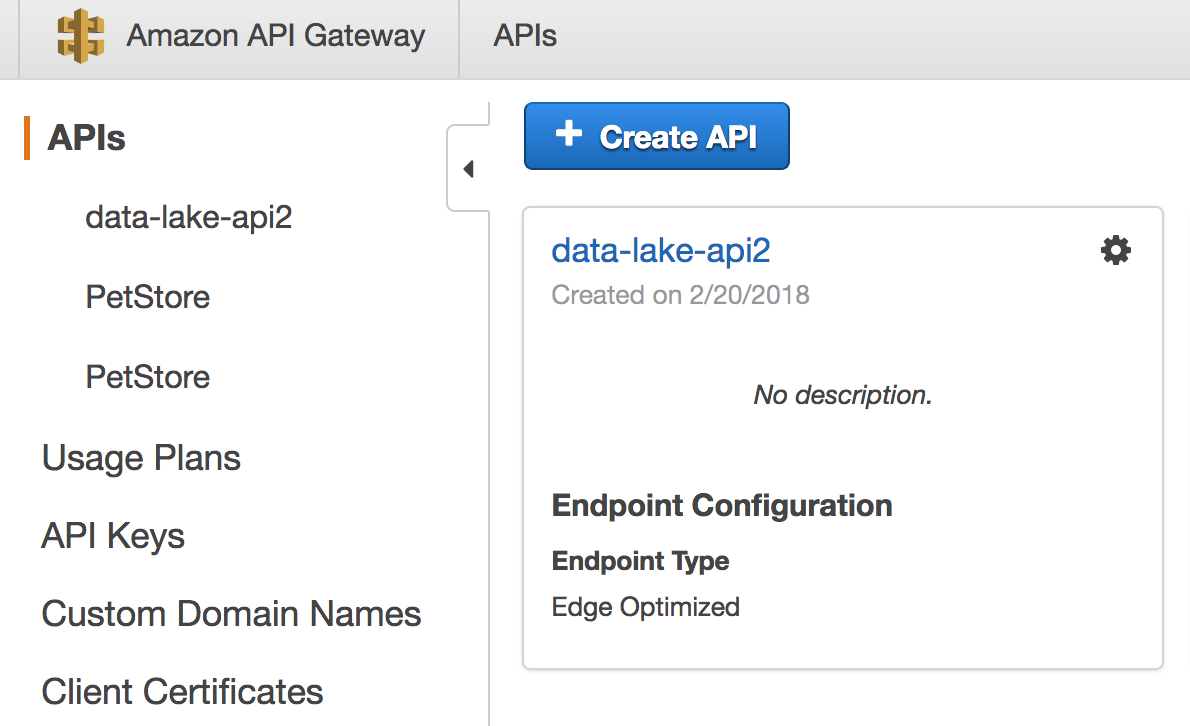
1. The creation process can take 25 minutes or longer and will create the following resources in your AWS account –

**Amazon S3 buckets –** There will be two S3 buckets created. One for the purpose of hosting the Data Lake static web application and the other for the purpose of storing the meta-data related to your packages.

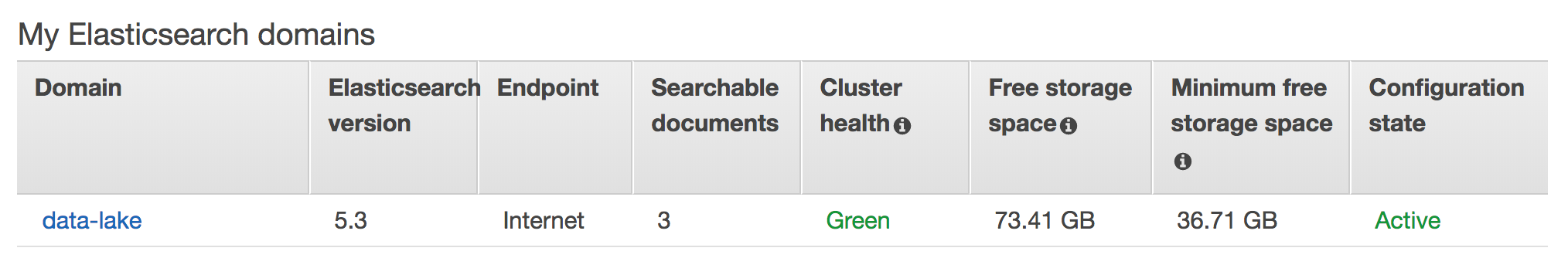
**AWS Lambda Functions –** There will be nine lambda functions created for different purposes such as creating packages, admin, performing each in Amazon Elasticsearch etc. Essentially each of these lambda functions make a microservice which are accessed by the web application via API Gateway.



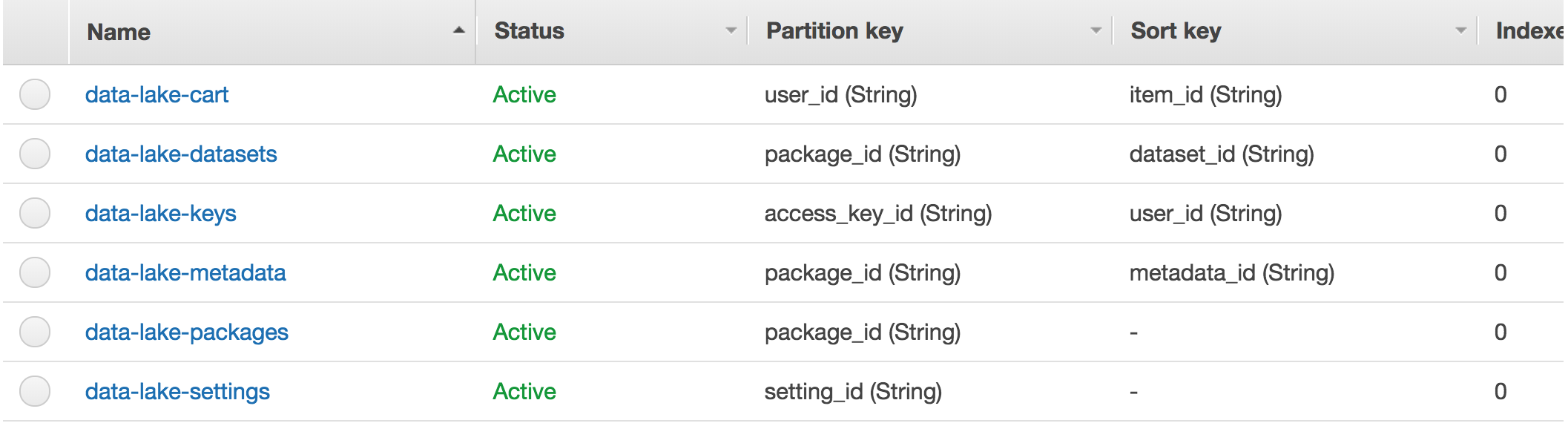
**Amazon API Gateway –** The API gateway provides a RESTful interface on top of various Lambda functions.



**Amazon ElasticSearch Cluster –** ES cluster will be used by the Data lake solution to store the metadata about your packages.



**Amazon DynamoDB Tables –** There will be six DynamoDB tables created by the CF template. These tables are used to store metadata information as well such as search tags for each package.

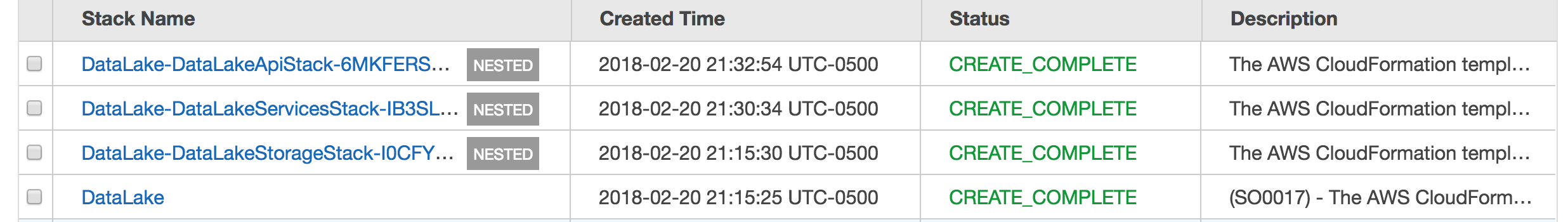


**Amazon Cognito UserPool –** The Cognito UserPool is created to manage users for the Data lake web application. You can look at all the active users for your Data Lake by going into the details view of the user pool with name ‘data-lake’ in your AWS account.

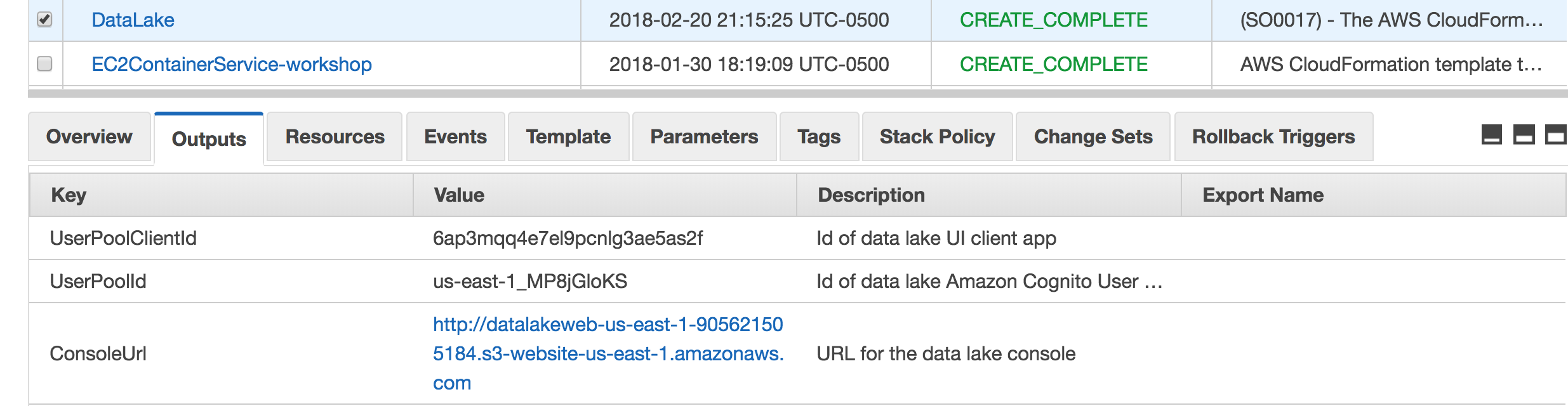
Feel free to inspect all these resources under your account and see their configurations.

# How to use the Data lake

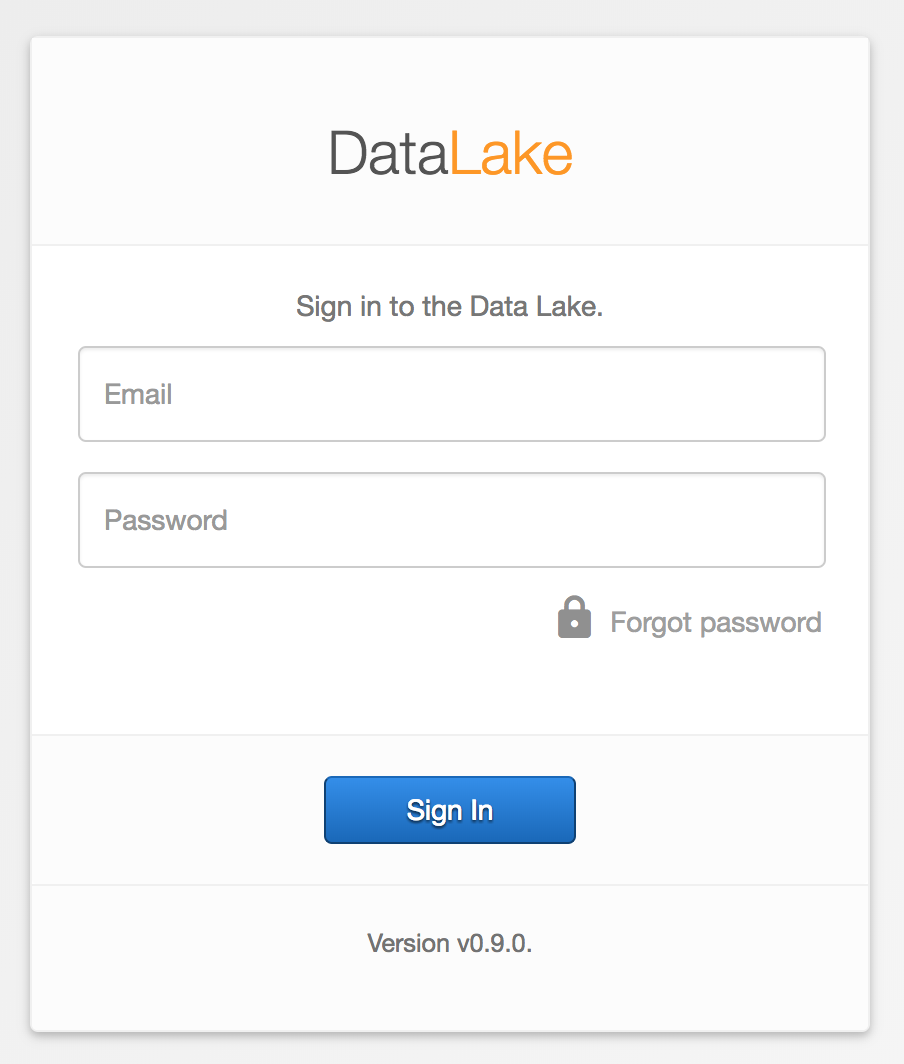
Once the CloudFormation template is in ‘CREATE\_COMPLETE’ status, your Data Lake is ready to be used. Please note you will end up having total four stacks in your console, and that is because the main template has three other nested stacks in it.

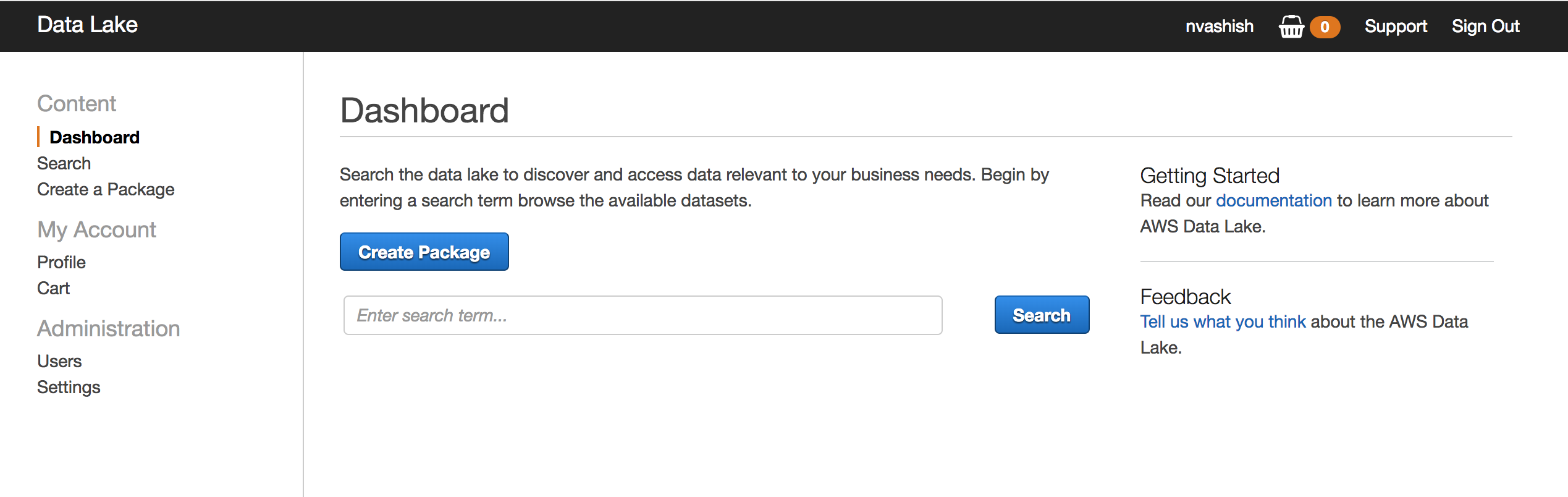


1. Check your email which you entered as an input parameter to your CF template. You should have an email with a temporary password for your user with a link to log into the Data lake console. Click on that link and it will take you to the log in screen of the Data Lake, where you will be asked to change your password the first time you login.
2. You can also log into the Data lake console using the endpoint provided by the CF template in the output tab. Click on the checkbox next to the main stack, and under the outputs tab, you will find a Console URL. This is the URL to use your Data lake.



1. Once you log into the Data lake, you will see the home page/Dashboard of the Data lake as below –





1. Primarily, you can do two things using the console –

* Create Packages - A Package is a container of your data in the Data Lake which allows you to create certain tags to associate with the data in that package. So, for example, you can create a package for streaming data coming from IoT sensors with some tags such as device name, device location, date and time.
* Search Packages – You can search the contents of the Data lake by specifying keywords and it will search the packages which are tagged with those keywords.

1. Other than the two main functionalities described above, you can also manage users of the Data lake and create some governance by requiring some tags to be specified while creation of the packages.
2. For detailed instructions on how to use the Data lake console and APIs, please visit - <http://docs.awssolutionsbuilder.com/data-lake/>

# Conclusion

In this lab you have learned how to create a Data lake in your AWS account by using the AWS Solutions Builder CloudFormation template. At the end of the next lab, you will use this Data lake to create and search a package, generated with the results of the Kinesis Analytics.