**March 2018**



**Intelligent Data Lake Workshop**

*Lab 2 – Glue, Athena, EMR, Redshift Spectrum and QuickSight*

Table of Contents

[Overview 3](#_Toc522444321)

[Query Data with Amazon Redshift Spectrum and AWS Glue Data Catalog (Optional) 9](#_Toc522444322)

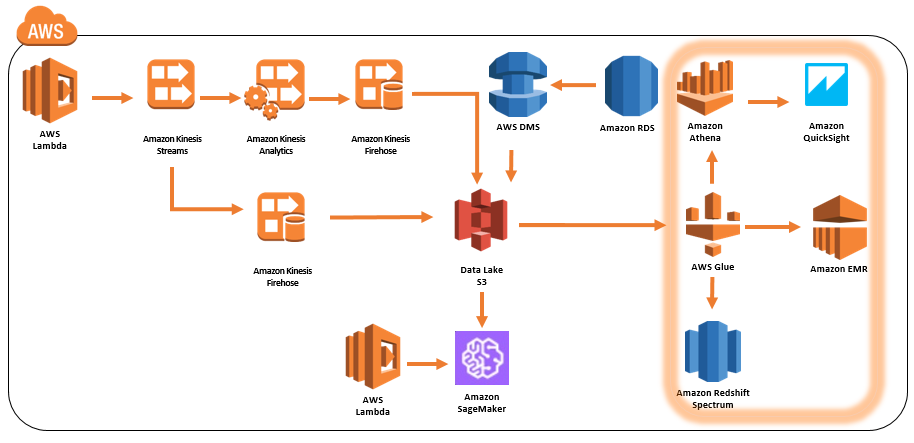
[Conclusion 15](#_Toc522444323)

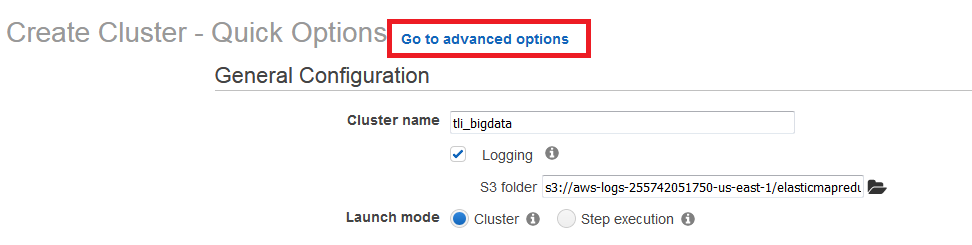
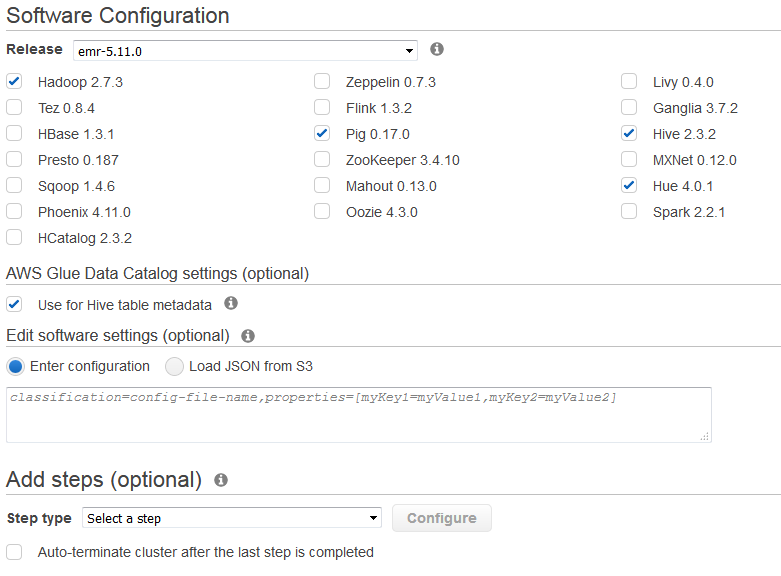
[Appendix 16](#_Toc522444324)

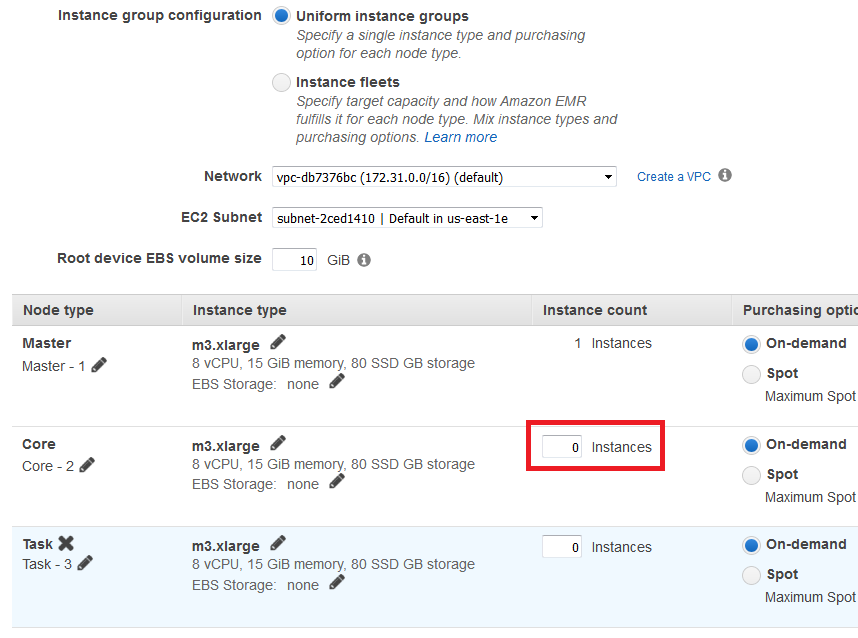
# Overview

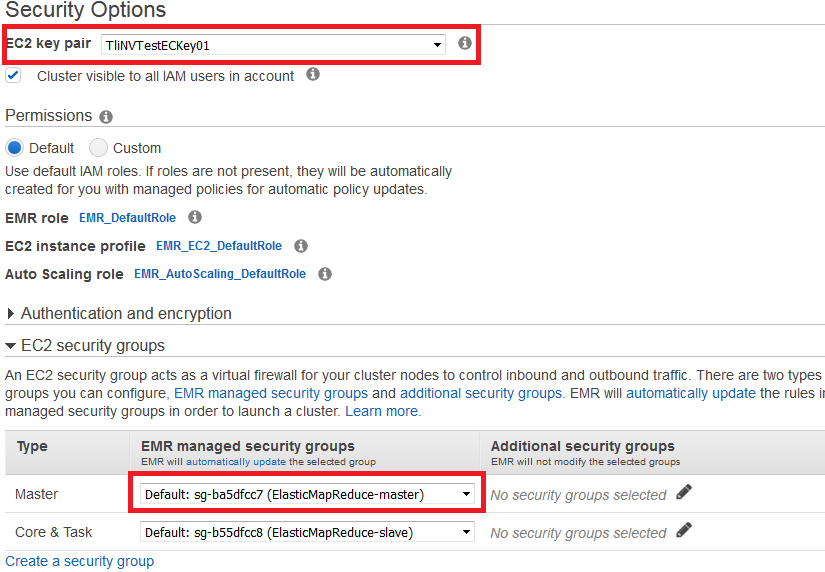
In lab 1, you used Kinesis Streams to collect and store the streaming IoT sensor data, then used Kinesis Analytics to process and analyze the streaming data continuously. You also used Amazon Kinesis Firehose to export both the raw and processed data into S3 for further analysis. In this Lab, you will use Glue Data Catalog to define schemas on the data and share it with disparate systems/services such as Amazon Athena, Amazon EMR and Amazon Redshift Spectrum. You will also explore how to use Glue ETL tools to transform raw data to Parquet format. Finally, you will use QuickSight to visualize the data stored on S3 with Athena. Diagram below depicts the complete architecture.

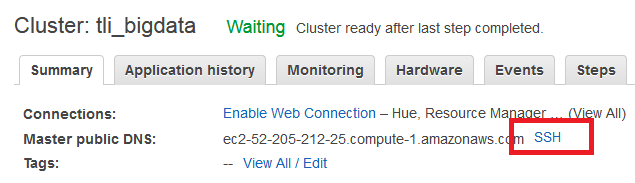
Diagram below with highlighted area depicts what you will be building in this lab.

 Query Data with Amazon EMR and AWS Glue Data Catalog (Optional)

1. The AWS Glue Data Catalog is a drop-in replacement for the Apache Hive Meta-store, so it is also easy to run Amazon EMR (Elastic MapReduce) with Hive and query the data on S3 with Glue Data Catalog. In this section, we will create an EMR cluster to demonstrate how easy it really is.
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).
4. Click on **EMR** from the list of all services. This will bring you to the Amazon EMR dashboard page
5. Click **Create cluster**
6. Click **Go to advanced options** at the top of the page
7. For Release, select **emr-5.11.0**. Tick **Hadoop, Pig, and Hive and Hue** checkboxes. Tick **Use for Hive table metadata** checkbox, this is important because it enables us to access the Glue Data Catalog and query the data on S3 with Hive. 
8. Click **Next**
9. For this lab, we only need 1 node cluster, so go ahead and change Core instances from **2** to **0**. For Network and EC2 subnet, make sure to pick a VPC and subnet that is publicly accessible as we will SSH into the master node later. Note, if you run into capacity limit error to run m3.xlarge instance, change instance type to m4.xlarge by clicking on the pencil icon next to m3.xlarge in Master node type.



1. Click **Next**
2. For Cluster name, enter **YourInitials\_bigdata**. Leaving everything else default and click **Next**
3. Select an EC2 key pair that you have private PEM file and make sure Master security group allows inbound SSH port from public (anywhere).If you do not have a EC2 key pair, refer to Appendix below for instructions to create one.
4. Click **Create cluster**
5. Take a break and wait until the cluster status shows **Waiting**
6. Click on the EMR cluster with name **YourInitials\_bigdata**. In Summary page, click on SSH link.



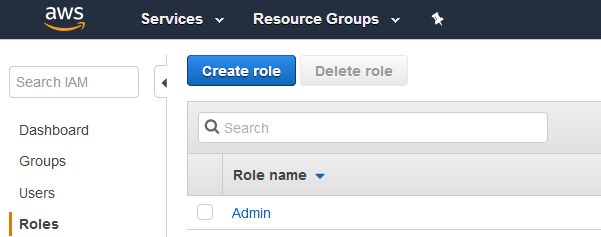
1. Follow the instruction and then SSH into the master node**.** Note, if you have issues connecting to the instance, make sure security group is opened to allow SSH inbound connection.
2. Enter **hive** in the terminal
3. Run the following HiveQL query, make sure to change database name with your initials

select sensorname, sensorvalue, anomalyscore from **YourInitial\_bigdata**.analytic\_csv2parquet where anomalyscore > 2 limit 10;

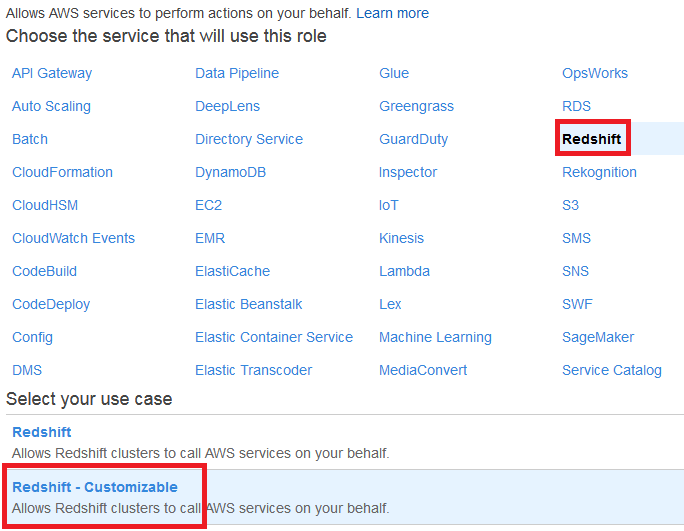
1. Feel free to run queries against other tables available to explore the data.

# Query Data with Amazon Redshift Spectrum and AWS Glue Data Catalog (Optional)

1. Note this section by no means a proper Redshift workshop, we are only going to focus on one of the Redshift features, Spectrum to learn how to query data in external datasource through AWS Glue Data Catalog.
2. **Prerequisites –** In order to connect and work with Amazon Redshift, you need a SQL query client such as SQL Workbench/J with JDBC or ODBC for RedShift installed. You can download and install Workbench/J from <http://www.sql-workbench.net>. Follow links provided for instructions on how to install and configure [JDBC](http://docs.aws.amazon.com/redshift/latest/mgmt/configure-jdbc-connection.html%22%20%5Co) and [ODBC](http://docs.aws.amazon.com/redshift/latest/mgmt/configure-odbc-connection.html%22%20%5Co%20%22Configure%20an%20ODBC%20Connection) for Redshift.
3. First we need create a role for Redshift Spectrum to have access to Glue Data Catalog and S3 resources.
4. Click on **IAM** from the list of all services. This will bring you to the IAM dashboard page.
5. Click **Roles** on the left hand panel and then click **Create role**



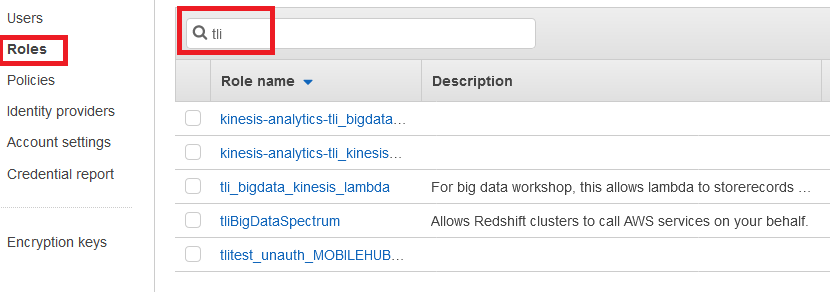
1. Select **Redshift** and select **Redshift - Customizable** for your use case



1. Click **Next: Permissions**
2. In attach permission policies page, add **AWSGlueServiceNotebookRole** and **AmazonS3FullAccess**.



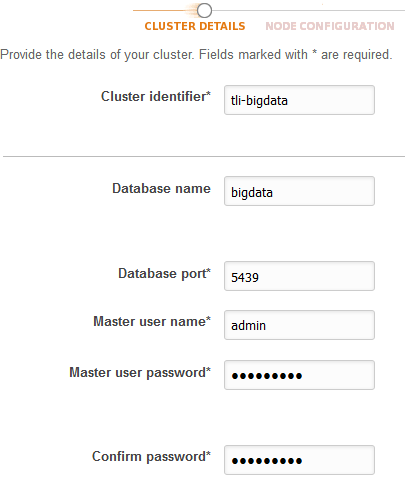
1. Click **Next: Review**
2. Name it **YourInitialsBigDataSpectrumRole** and Click **Create Role**
3. In the IAM roles, page search for the new role you created



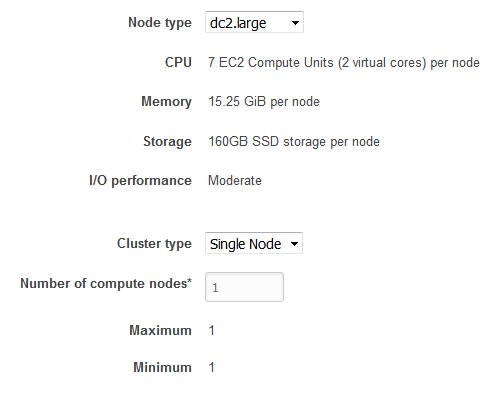
1. Click **YourInitialsBigDataSpectrumRole**



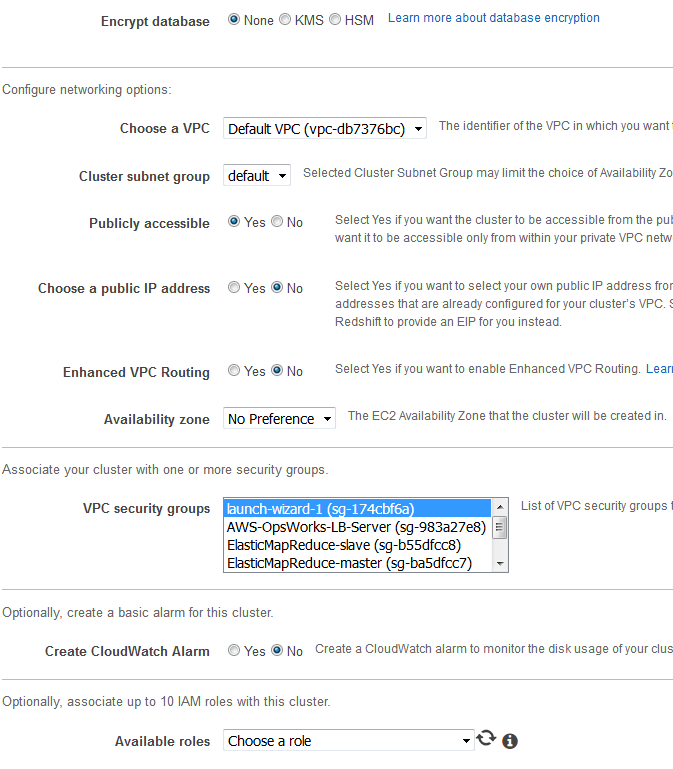
1. Copy the **Role ARN** for later use
2. Back to the AWS Management Console <https://console.aws.amazon.com/>.
3. In the upper-right corner of the AWS Management Console, confirm you are in the desired AWS region (e.g., N. Virginia).
4. Click on **Redshift** from the list of all services. This will bring you to the Amazon Redshift dashboard page
5. Click **Launch cluster**
6. In Cluster Details page, enter the following
   * Cluster identifier: **YourInitials-bigdata**
   * Database name: **bigdata**
   * Database port: **5439**
   * Master user name: **admin**
   * Master user password: **YourPassword**
   * Confirm password: **YourPassword**



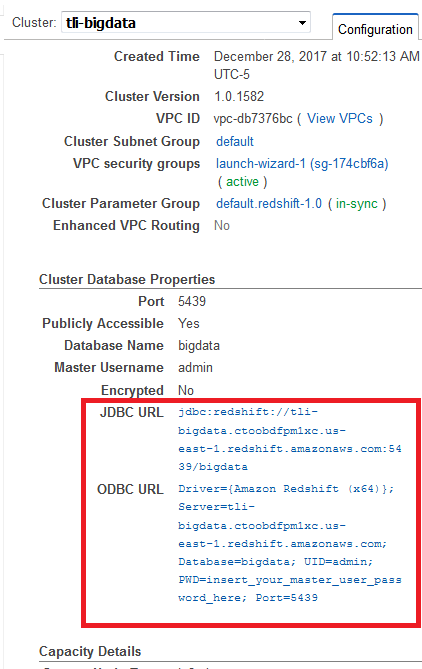
1. Click **Continue**
2. Leave everything default, single node dc2.large cluster is enough for this exercise.



1. Click **Continue**
2. In Additional Configuration page, enter the following
   * Cluster parameter group: **default**
   * Encrypt database: **None**
   * Choose a VPC: **Pick a VPC with Public Access**
   * Cluster subnet group: **default**
   * Publicly accessible: **Yes**
   * Choose a public IP address: **No**
   * Enhanced VPC Routing: **No**
   * Availablity zone: **No Preference**
   * VPC security groups: **Pick one,** make sure the selected security group allows inbound database port 5439 from public (anywhere).
   * Create CloudWatch Alarm: **No**
   * Available roles: **YourInitialsBigDataSpectrumRole**



1. Click **Continue**
2. Review configuration and click **Launch**
3. Take a break and wait until the cluster status shows **available**
4. Click YourInitals-bigdata cluster
5. In the detail page, copy the JDBC or ODBC URL and use it to connect to the Redshift cluster with your SQL query client such as Workbench/J.



1. Once connected with your SQL query client. Execute the following SQL, note to change region if you are not using N. Virginia region.

create external schema **YourInitial**\_bigdata from data catalog

database '**YourInitial\_bigdata**'

iam\_role '**Role ARN Copied in Step 12'**

region 'us-east-1';

1. You should receive a message indicating YourInitials\_bigdata schema created successfully. Execute the following SQL

select sensorname, sensorvalue, anomalyscore from **YourInitial\_bigdata**.analytic\_csv2parquet where anomalyscore > 2 limit 10;

# Conclusion

In this lab, you have learned how to leverage AWS Glue, Amazon Athena, Amazon EMR, Amazon Redshift, and Amazon QuickSight to continue to process, transform, query, analyze and visualize the data stored in your S3 data lake. In the next lab, you will learn how to use Amazon SageMaker to train the stored data, then host the model, and finally calling the SageMaker Endpoint for prediction/inference.

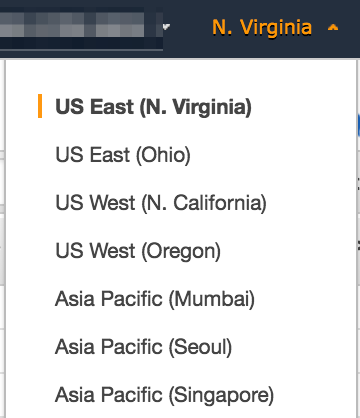
# Appendix

**Create New EC2 Key Pair**

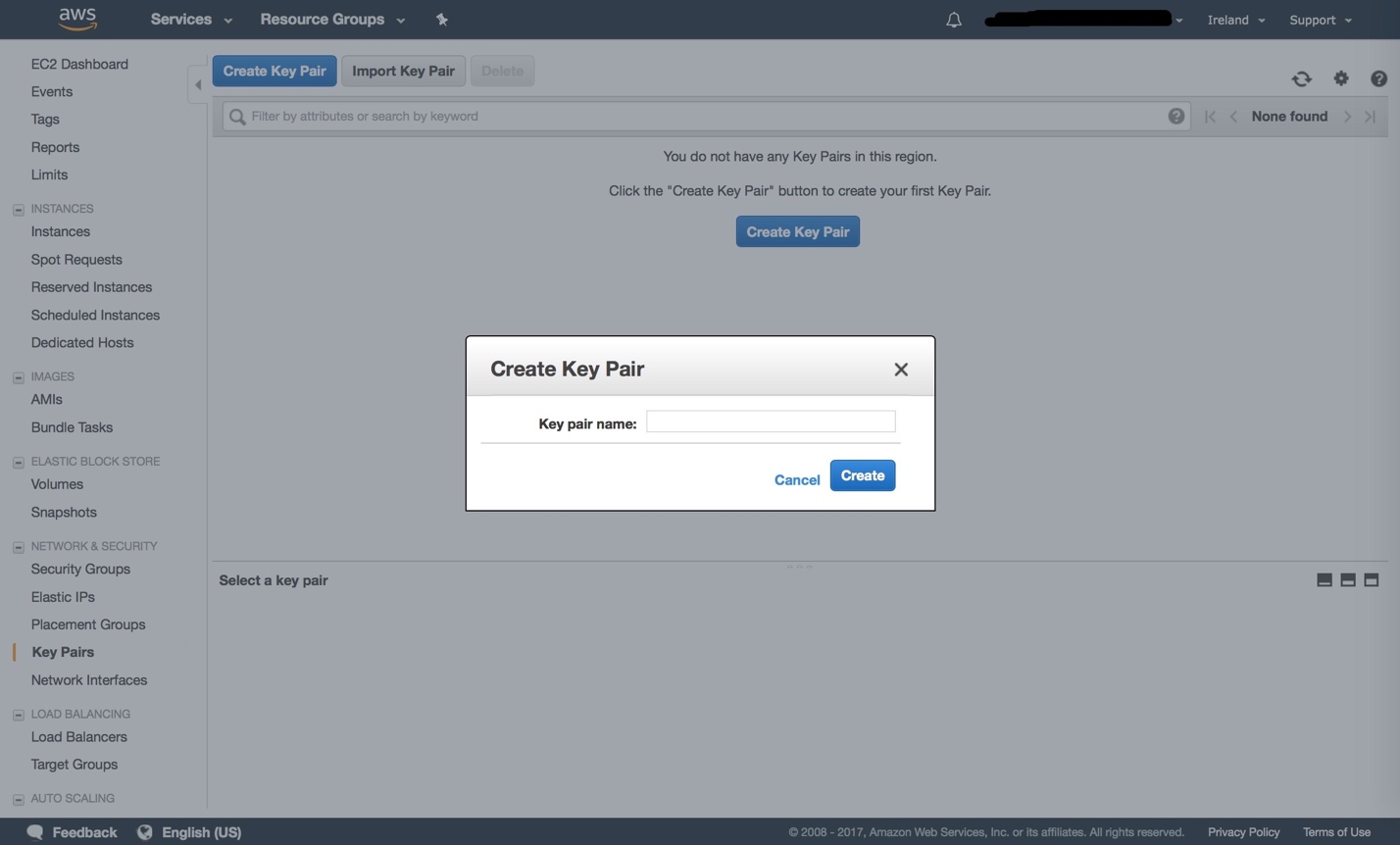
Before launching CloudFormation, you will need to create an EC2 Key Pair which will be used to allow administrative access into the EC2 Instances created by CloudFormation.

Sign in to the AWS management console. Open the EC2 page by clicking the searching for EC2 or going to <https://console.aws.amazon.com/ec2/home>.

Once on the EC2 page, look in the upper right corner for the AWS Region and change this to **US East (N. Virginia)**



* Under **Network & Security**, select **Key Pairs** on the left and select **Create Key Pair**. Recommended name “DMSlab”. Please append your initials if multiple people are running in the same account, such as “DMSlab\_ws”.



* + Mac users: You will receive a download of a pem.txt file. Rename the file and remove the .txt to make it a .pem file. For example: DMSlab\_ws.pem. Note where this file was downloaded.
  + Windows users: You will be prompted to download a .pem file. Save this, and note where this was downloaded.

You have completed creating a new EC2 Key Pair.