Prerequisites

I.   Lab Facilitator's Checklist

If you're not a lab facilitator, this section isn't pertinent. You can skip ahead.

Facilitators, here is a summary of things required to prepare for the lab. It is recommended to plan at least a few weeks ahead. More detail steps are provided in the next section.

- Ensure that all participants have access to a shared AWS account.
- All participants should have their own their own [EC2 key pair](#) from the region the lab is being hosted. If you don't plan to provide these to the participants ahead of the lab, you should provide guidance to create one as these steps aren't documented in the exercise.
- Create an IAM role, and attach [this policy](#) along with this [trust policy](#). You will need to provide the participants the name of this role, so that they can run a CloudFormation template to create their personal resources.
- Create a VPC for this lab using this [Cloud Formation template.](#) This will serve as the base environment for the workshop. You can execute the template with the IAM role you created if you don't have administrator level access.
- Ensure an S3 bucket is prepared in your account with the data required by the workshop. More details on setting that up is provided below.
- Note that the Redshift team recommends running the lab in either us-west-2 or us-east-1 as these regions have the highest capacity. If you plan to run this lab with a large group in a different region, inform the Redshift service team so that capacity can be prepared accordingly.
- Ensure the account limits can support the number of participants.

Create Lab Environment

1. **Prepare account limits**
   i. **Glue Data Catalog API:** If (6 X #participants) exceeds 100, you will need to submit a support ticket and request for the api request rate for the Glue Data catalog to be increased to that number.
   ii. **M4.Large Instances**: each participant will use a m4.large instance as their Windows bastion host. Increase the default of 20 as needed.
   iii. **Redshift Nodes:** Each participants will launch 4 nodes. The default account limit per region is 200. Adjust this limit for very large groups.
   iv. Finally, if you are running a large event outside of us-east-1 or us-west-2, inform the Redshift service team so that they can prepare Redshift Spectrum capacity accordingly.

2. **Create the workshop VPC**

   i. Create a custom IAM policy with the following specifications: [redshift-spectrum-lab-build-access.](#)

ii.   Create an IAM role called "redshift-spectrum-lab-builder." Attach the redshift-spectrum-lab-build-access policy to this role, and configure the trust policy with the following specifications.

iii.   Download the immersion-day-redshift-spectrum-vpc.template. This template will be used to create the VPC for hosting the participants' workshop resources.

iv.   Create a stack using the CloudFormation template you just downloaded. Name the stack "workshop-vpc." You will be provided a number of configuration options. Select your availability zone of choice, and leave the rest with default values.

v.   The participants will be launching Windows bastion hosts during the lab. A custom AMI has been created with SQL workbench and Redshift drivers pre-installed to accelerate the lab. Public AMIs are available in us-west-2 and us-east-1. Copy one of those AMI to the region that is hosting your lab if it is different from these two. The AMI IDs are:

- us-west-2: ami-48996b30
- us-east-1: ami-4ee71134

Provide the appropriate AMI ID to the participants. Note that at the time this lab was created the AMI was installed with the following:

- Redshift Driver: JDBC 4.2–compatible driver.
- SQL Workbench Build 123 (2017-09-25)

It's possible that this AMI may have to be updated in the future, so please take note. In the event there is an unexpected compatibility issue in the future, the bastion host will have to be created with the base windows AMI, and SQL Workbench will have to be installed manually by following our documentation.

3. Prepare your S3 buckets

The data sets for this workshop add up to approximately 730GB of data in S3. You should expect the storage costs to be under $20 for the duration of the event.

i.   Create two buckets: <unique_id>-redshift-spectrum-datastore-csv10 and <unique_id>-redshift-spectrum-datastore-parquet1. Replace <unique_id> to ensure your bucket name is globally unique.

ii.   Copy the lab datasets over to the buckets that you just created. Note that you must be an authenticated user to download these data sets. You can run these two AWS CLI commands to copy the data:

- aws s3 sync --source-region us-east-1 s3://redshift-spectrum-bigdata-blog-datasets/clickstream-csv10 s3://<my-unique-id>-redshift-spectrum-datastore-csv10

- aws s3 sync --source-region us-east-1 s3://redshift-spectrum-bigdata-blog-datasets/clickstream-parquet1 s3://<my-unique-id>-redshift-spectrum-datastore-parquet1

Note that the copy process may take hours.

4. **[Highly Recommended]** This next steps involve setting up the environments for each of the lab participants. The diagram below illustrates the resources that will be created. The primary resources will be a Windows bastion host with a client pre-installed, and a 4-node (dc1.large) Redshift cluster.
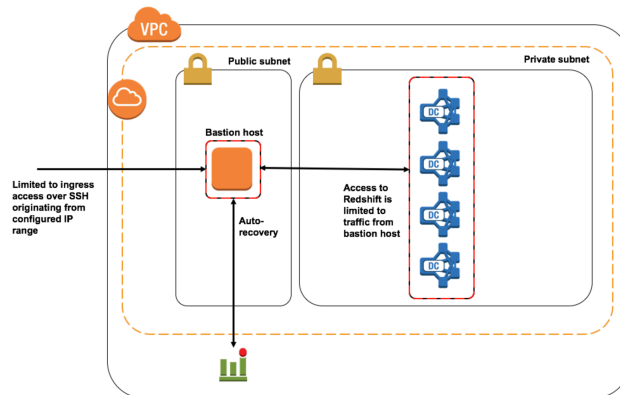
Note that it is possible to have the participants complete this step during the lab; however, based on SA feedback, it is highly recommended that you do this for the customer. The steps are largely around configuring a CloudFormation template, which doesn't add value to a Redshift Spectrum workshop. Doing this will also give you the assurance that the account limits are properly set.

Repeat the following steps for each participant. A good strategy to minimize effort and cost could be to have participants work in groups of 2-3 for this lab.

1. Navigate to the CloudFormation service, and locate the stack "workshop-vpc." This stack should have been launched by the workshop facilitators.

2. Click on the "workshop-vpc" link. You will be taken to a page with a number of subsections. Expand the subsection "**Outputs**." Take note of the four values as shown in the screenshot. You will need these to create your personal lab environment.

workshop-vpc

| | Other Actions ▾ | Update Stack |

Stack name:	workshop-vpc
Stack ID:	arn:aws:cloudformation:us-west-2:803235869972:stack/workshop-vpc/386ae6a0-9a84-11e7-91ca-503f20f2adae
Status:	CREATE_COMPLETE
Status reason:	
IAM Role:	redshift-spectrum-poc-env-builder (arn:aws:iam::803235869972:role/redshift-spectrum-poc-env-builder)
Description:	This template creates an isolated PoC environment for Redshift Spectrum

▾ Outputs

Key	Value	Description	Export Name
AvailabilityZone	us-west-2a	The AZ where the subnets are created	workshop-vpc-AvailabilityZone
PublicSubnet	subnet-1e74bf56	public subnet	workshop-vpc-PublicSubnet
StackVPC	vpc-48b9dc2e	The ID of the VPC	workshop-vpc-VPCID
AnalyticalPrivateSubnet	subnet-58894310	private subnet where Redshift cluster …	workshop-vpc-PrivateAnalyticalSubnet

3. Download the CloudFormation template linked here. This template defines resources required by the lab, and will instruct CloudFormation to create these resources for you.

4. Return to the main page of the CloudFormation service, and click on "Create Stack."

5. Select the **Upload to Amazon S3** option, and upload the template you just downloaded. Click next.

6. You will be presented with a form that requires your input. The first parameter that requires your input is "**Stack name**." Provide a name that is unique among your peers, and remember this name. You will be responsible for deleting this stack at the end of the lab.

Specify Details

Specify a stack name and parameter values. You can use or change the default parameter values, which are defined in the AWS CloudFormation template. Learn more.

Stack name student1-env

7. The next set of parameters informs the template about the VPC where your lab resources will be deployed.

VPC Configuration

Workshop VPC ID	replace with vpc id	VPC that will house the workshop resources
Public Subnet ID	replace with public subnet id	public subnet where your bastion host will be launched
Private Subnet ID	replace with private subnet id	private subnet where your Redshift cluster will be launched

Refer back to the output values from the "workshop-vpc" stack that you were asked to take note of.

- For **Workshop VPC ID**, enter the value associated with the key "StackVPC."
- For **Public Subnet ID**, enter the value associated with the key "PublicSubnet."
- For **Private Subnet ID**, enter the value associated with the key "AnalyticalPrivateSubnet."

8. The parameters in the next section are used to configure your bastion host.

Bastion Host Configuration

Bastion Host AMI ID		Provide the custom bastion host ami id
Instance Type	m4.large	Instance type for your bastion host
EC2 Key Pair		Name of an existing EC2 key pair, which you will use to log into the Bastion host
Whitelist CIDR Block	0.0.0.0/0	Allow inbound traffic to the bastion host from this CIDR range.

For **Bastion Host AMI ID**, enter the custom ami id that was provided by your lab facilitator. It should resemble the pattern like ami-xxxxxxx.

Leave the **Instance Type** as the default value. For **EC2 Key Pair**, enter the name of your private key (don't include the .pem extension in your input text). Leave **Whitelist CIDR Block** as the default.

9. The next section consists of configurations required to provision your Redshift cluster. Leave all the values as their defaults. You will need to provide a **DB User Password**. Your password needs to be between 8 to 64 characters, and has to contain at least one digit and capital letter. Click next once you have provided a valid password.

Redshift Cluster Configuration

Cluster Type	multi-node	The type of cluster
Number of Compute Nodes	4	The number of compute nodes in the cluster. For multi-node clusters, the NumberOfNodes parameter must be greater than 1
Node Instance Type	dc1.large	The type of node to be provisioned
Listener Port	5439	The port number on which the cluster accepts incoming connections.
Database Name	ssb	The name of the first database to be created when the cluster is created
DB User Name	admin	The user name that is associated with the master user account for the cluster that is being created
DB User Password		The password that is associated with the master user account for the cluster that is being created.

10. Scroll down to the **Permissions** sub section, and select the "redshift-spectrum-lab-builder" role. If you don't see the role, ask your lab facilitator for the name of the role that was created for this workshop. This role provides CloudFormation with the necessary permissions to launch the resources for your lab environment. Click on the Next button at the bottom of the page.

Permissions

You can choose an IAM role that CloudFormation uses to create, modify, or delete resources in the stack. If you don't choose a role, CloudFormation uses the permissions defined in your account. Learn more.

IAM Role	redshift-spectrum-poc-env-buil ⬍
	Enter role arn
	arn:aws:iam::803235869972:role/

11. Select the checkbox "**I acknowledge that AWS CloudFormation might create IAM resources**." Click on the Create button at the bottom of the page.

Capabilities

ℹ **The following resource(s) require capabilities: [AWS::IAM::Role]**

This template contains Identity and Access Management (IAM) resources that might provide entities access to make changes to your AWS account. Check that you want to create each of these resources and that they have the minimum required permissions. Learn more.

☐ **I acknowledge that AWS CloudFormation might create IAM resources.**

12. It will take 5-10 minutes for your resources to be provisioned. Your environment has been created once the status of your stack shows up as "CREATE_COMPLETE" on the main CloudFormation service page. You can continue with the lab once your stack reports this status.

	Stack Name	Created Time	Status	Description
☐	student1-env	2017-09-15 22:35:01 UTC-0700	CREATE_COMPLETE	This template creates an isolated PoC environment for Redshift .
☐	workshop-vpc	2017-09-15 19:10:33 UTC-0700	CREATE_COMPLETE	This template creates an isolated PoC environment for Redshift .
☐	production-vpc	2017-04-12 12:23:08 UTC-0700	CREATE_COMPLETE	Provides networking configuration for a standard, public facing .