



Amazon Web Services Data Engineering Immersion Day

Lab 1. Hydrating the Data Lake with DMS
August 2020

Table of Contents

<i>Introduction.....</i>	<i>2</i>
<i>Get Started Using the Lab Environment</i>	<i>3</i>
<i>Create the Subnet Group.....</i>	<i>6</i>
<i>Create the Replication Instance.....</i>	<i>7</i>
<i>Create the DMS Source Endpoint</i>	<i>9</i>
<i>Create the Target Endpoint.....</i>	<i>11</i>
<i>Create a task to perform the initial full copy</i>	<i>14</i>
<i>(Optional) Create a DMS endpoint to perform ongoing replication.....</i>	<i>19</i>
<i>(Optional) Create a task to perform ongoing replication</i>	<i>20</i>

Introduction

This lab will give you an understanding of the AWS Database Migration Service (AWS DMS). You will migrate data from an existing Amazon Relational Database Service (Amazon RDS) Postgres database to an Amazon Simple Storage Service (Amazon S3) bucket that you create.



In this lab you will complete the following tasks:

1. Create a subnet group within the DMS Lab VPC
2. Create a DMS replication instance
3. Create a source endpoint
4. Create a target endpoint
5. Create a task to perform the initial migration of the data.

Optionally, you can add ongoing replication of data changes on the source (***Only one DMS replication instance will enable this feature.***)

6. Create target endpoint for CDC files to place these files in a separate location than the initial load files
7. Create a task to perform the ongoing replication of data changes

Your instructor has created and populated the RDS Postgres database that you will use as your source endpoint in this lab.

If you'd like to run the workshop on your own after the AWS hosted event, please follow the lab instruction here: <https://github.com/aws-samples/data-engineering-for-aws-immersion-day>

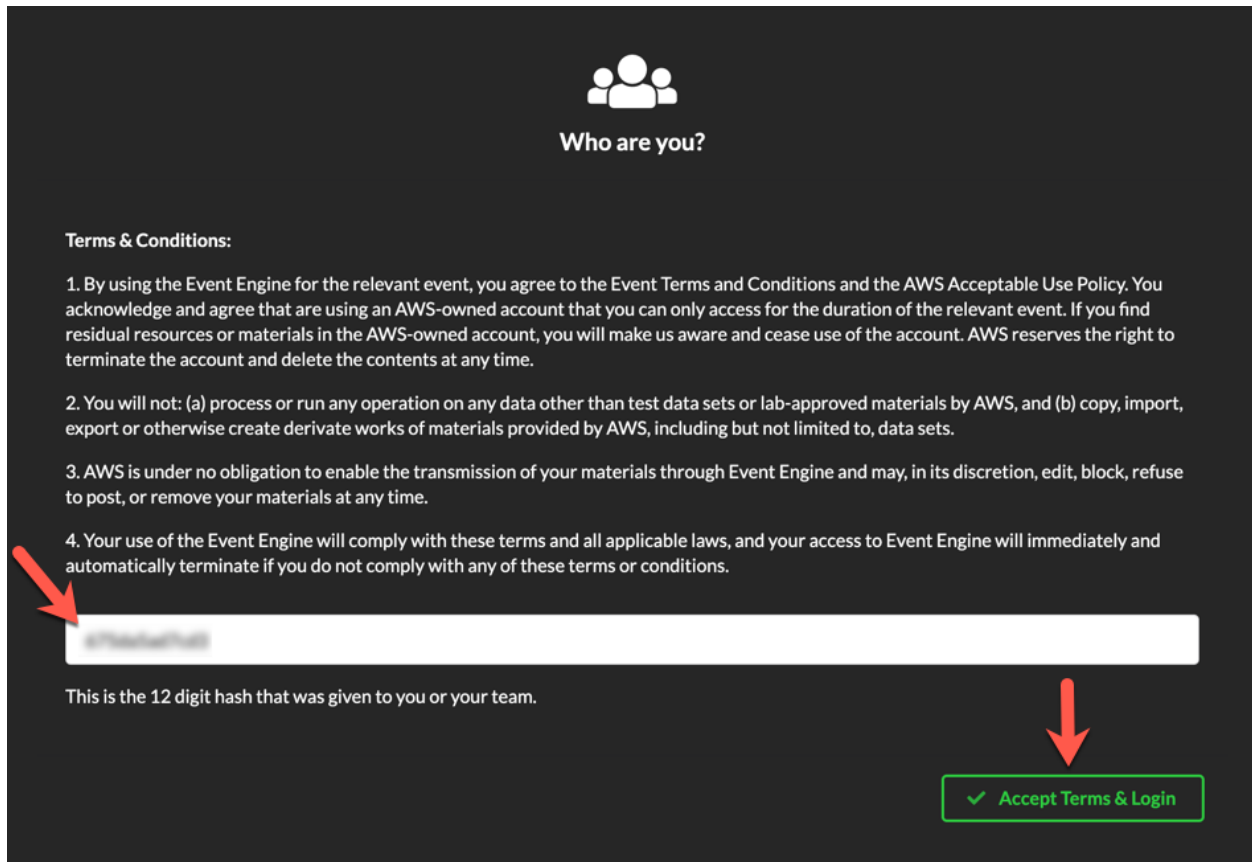
Get Started Using the Lab Environment

Please skip this section if you are running the lab on your own AWS account.

Today, you are attending a formal event and you will have been sent your access details beforehand. If in the future you might want to perform these labs in your own AWS environment by yourself, you can follow instructions on GitHub - <https://github.com/aws-samples/data-engineering-for-aws-immersion-day>.

A 12-character access code (or 'hash') is the access code that grants you permission to use a dedicated AWS account for the purposes of this workshop.

1. Go to <https://dashboard.eventengine.run/>, enter the access code and click Proceed:



Who are you?

Terms & Conditions:

1. By using the Event Engine for the relevant event, you agree to the Event Terms and Conditions and the AWS Acceptable Use Policy. You acknowledge and agree that are using an AWS-owned account that you can only access for the duration of the relevant event. If you find residual resources or materials in the AWS-owned account, you will make us aware and cease use of the account. AWS reserves the right to terminate the account and delete the contents at any time.
2. You will not: (a) process or run any operation on any data other than test data sets or lab-approved materials by AWS, and (b) copy, import, export or otherwise create derivate works of materials provided by AWS, including but not limited to, data sets.
3. AWS is under no obligation to enable the transmission of your materials through Event Engine and may, in its discretion, edit, block, refuse to post, or remove your materials at any time.
4. Your use of the Event Engine will comply with these terms and all applicable laws, and your access to Event Engine will immediately and automatically terminate if you do not comply with any of these terms or conditions.

This is the 12 digit hash that was given to you or your team.

✓ Accept Terms & Login

2. On the Team Dashboard web page you will see a set of parameters that you will need during the labs. Best to save them to a text file locally, alternatively you can always go to this page to review them. Replace the parameters with the corresponding values from here where indicated in subsequent labs:

Lab 1. Hydrating the Data Lake with DMS


Because you're at a formal event, some AWS resources have been pre-deployed for your convenience, for example:

- The source database connection in RDS DB Info module








RDS DB Info [Readme](#)

Outputs:
No outputs defined

- S3 Bucket, IAM role for the DMS lab etc


 **Modules**

Environment Setup [Readme](#)

Outputs:
S3 Bucket name
mod-3fccddd609114925-dmslabs3bucket-1ngcgzzcnd15u 
BusinessAnalystUser
mod-3fccddd609114925-BusinessAnalystUser-MBOXFZLQLOXX 
DMSLabRoleS3 ARN
arn:aws:iam::377243295828:role/mod-3fccddd609114925-DMSLabRoleS3-O2VT1RSN43SG 
Glue Lab Role
mod-3fccddd609114925-GlueLabRole-YLTJA13WW6WT 
S3BucketWorkgroupA
mod-3fccddd609114925-s3bucketworkgroupa-tbon3m1mkunh 
S3BucketWorkgroupB
mod-3fccddd609114925-s3bucketworkgroupb-18ygl8nfp8ead 
WorkgroupManagerUser
mod-3fccddd609114925-WorkgroupManagerUser-5IVE0UQNIBG4 

3. On the Team Dashboard, please click AWS Console to log into the AWS Management Console:

Team Dashboard

 **Event**

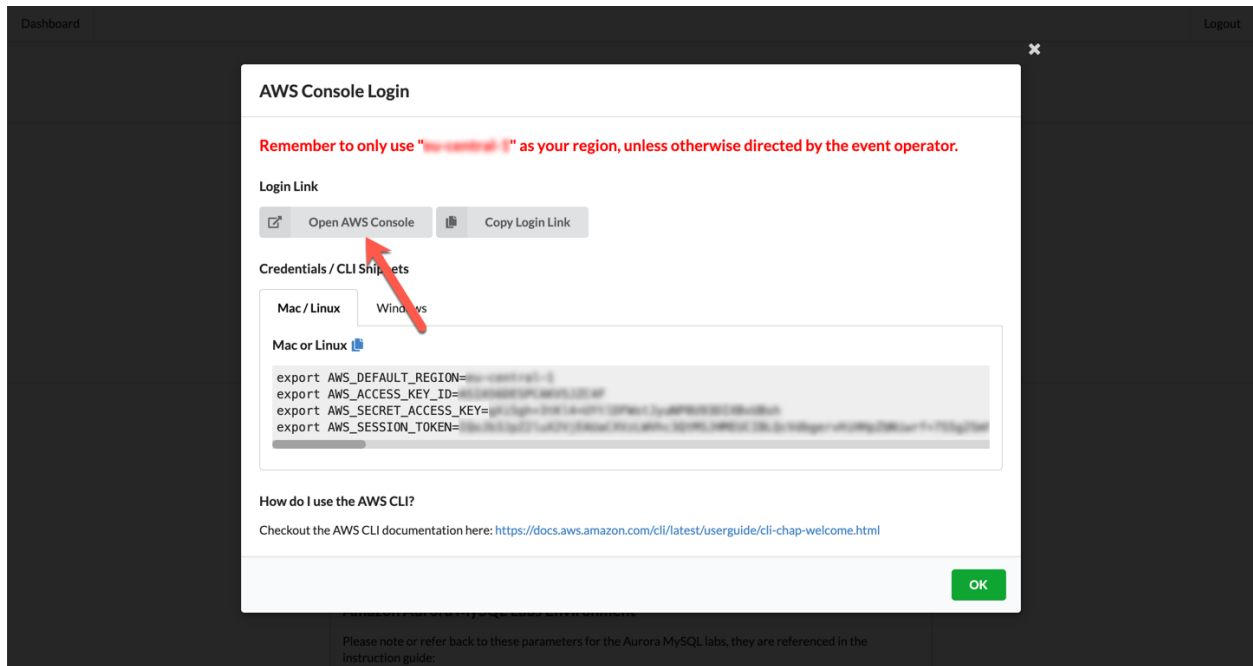
[AWS Console](#) [SSH Key](#)

Event **Data Engineering Immersion Day - Test**
Team Name:

Event ID: d2302d4ae9ff4ea2857846b74f7de7e2
Team ID: 1c2f7ad7ec044b0b8276f917c5983133

Lab 1. Hydrating the Data Lake with DMS

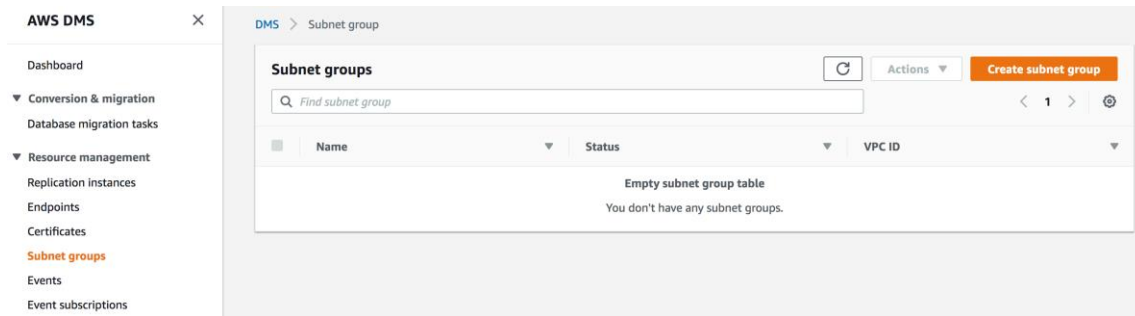
4. Click Open AWS Console. For the purposes of this workshop, you will not need to use command line and API access credentials:



Once you have completed these steps, you can continue with the rest of this lab.

Create the Subnet Group

1. Navigate to the DMS Console:
<https://console.aws.amazon.com/dms/v2/home?region=us-east-1#subnetGroup>
2. On the DMS console, select **Subnet Groups**.



3. Click **Create subnet group**.
 - a. In the **Name** box, type a descriptive name that you will easily recognize (e.g., "dms-lab-subnet-grp").
 - b. In the **Description** box, type an easily recognizable description (e.g., "Replication instance for production data system").
 - c. For **VPC**, select the pre-created VPC ending with **dmslstudv1**.
The subnet list populates in the Available Subnets pane.
 - d. Select as many subnets as you want and click Add. The selected subnets move to the Subnet Group pane. Note: DMS requires at least two separate availability zones to be selected.

Lab 1. Hydrating the Data Lake with DMS

AWS DMS

Dashboard

▼ Migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

Subnet groups

Events

Event subscriptions

What's new 2

Notifications

DMS > Subnet groups > Create subnet group

Create replication subnet group

Subnet group configuration

Name
A regionally scoped unique identifier you will use to identify your Replication Subnet Group
dms-lab-subnet-grp

Description
Free form text to describe your Replication Subnet Group
Replication instance for production data system

VPC
vpc-0f62b5136d50966b8 ... ▼

Add subnets

Add Subnet(s) to this Subnet Group. You may add subnets one at a time or add all the subnets related to this VPC. You may make additions/edits after this group is created.

subnet-0b5196c1746607c18 - public_subnet
us-east-1c 10.0.0.64/26 Public

subnet-0c80a2e5158507319 - private_subnet
us-east-1a 10.0.0.0/26 Private

subnet-039e4137660a20465 - private_subnet
us-east-1b 10.0.0.128/26 Private

► Tags

Cancel Create subnet group

4. Click **Create subnet group**
5. On the DMS console, the subnet group status displays **Complete**.

DMS > Subnet group

Subnet groups (1)

Find subnet group

	Name	Status	VPC ID
<input type="checkbox"/>	dms-lab-subnet-grp	Complete	vpc-0314e829ba12d9481

Create subnet group

Create the Replication Instance

1. On the DMS console, select **Replication instances**.

AWS DMS

Dashboard

▼ Conversion & migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

DMS > Replication Instance

Replication instances

Find replication instance

	Name	Class	Status	Engine version	Availability zone	VPC	Public	Public IP address	Private IP address
Empty replication instance table You don't have any replication instances.									

Create replication instance

2. Click **Create replication instance**.

Lab 1. Hydrating the Data Lake with DMS

- For **Name**, type a name for the replication instance that you will easily recognize. (e.g., "DMS-Replication-Instance").
- For **Description**, type a description you will easily recognize. (e.g., "DMS Replication Instance").
- For **Instance class**, choose **dms.t2.medium**
- Select **Engine version** as **3.3.1**
- For **VPC**, select the name of the VPC that you created earlier with AWS CloudFormation template. VPC name ending with **dmslstudv1**

AWS DMS ×

Dashboard

▼ Conversion & migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

Subnet groups

Events

Event subscriptions

What's new

Notifications

Replication instance configuration

Name
The name must be unique among all of your replication instances in the current AWS region.

Replication instance name must not start with a numeric value

Description

The description must only have unicode letters, digits, whitespace, or one of these symbols: _:/=+@. 1000 maximum character.

Instance class
Choose an appropriate instance class for your replication needs. Each instance class provides differing levels of compute, network and memory capacity.

Billing is based on [DMS pricing](#).

Engine version
Choose an AWS DMS version to run on your replication instance.

Allocated storage (GiB)
Choose the amount of storage space you want for your replication instance. AWS DMS uses this storage for log files and cached transactions while replication tasks are in progress.

VPC
Choose an Amazon Virtual Private Cloud (VPC) where your replication instance should run.

☐ **Multi AZ**
If you choose this option, AWS DMS will perform a multi-AZ deployment, with a primary instance in one availability zone (AZ) and a standby instance in another AZ. This configuration provides a highly available, fault-tolerant replication environment.

- Click **Advanced security and network configuration** section.
- Select the security group with **sgdefault** in the name.

Lab 1. Hydrating the Data Lake with DMS

AWS DMS

Dashboard

▼ Migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

Subnet groups

Events

Event subscriptions

What's new **2**

Notifications

☒ **Publicly accessible**
If you choose this option, AWS DMS will assign a public IP address to your replication instance, and you'll be able to connect to databases outside of your Amazon VPC.

▼ **Advanced security and network configuration**

Replication subnet group
Choose a subnet group for your replication instance. The subnet group defines the IP ranges and subnets that your replication instance can use within the Amazon VPC you've chosen.

dms-lab-subnet-grp

Availability zone
Choose an availability zone (AZ) where you want your replication instance to run. The default is "No preference", meaning that AWS DMS will determine which AZ to use.

No Preference

VPC security group(s)
Choose one or more security groups for your replication instances. The security group(s) specify inbound and outbound rules to control network access to your replication instance.

Use default

mod-3fcdd609114925-sgdefault-19HPA00P1U78F

KMS master key [Info](#)

(Default) aws/dms

► **Maintenance**

3. Click **Create**.
4. The DMS console displays **creating** for the instance status. When the replication instance is ready, the status changes to **available**. While replication instance is spinning up, you can proceed to next step for DMS endpoint creation.

DMS > Replication instances

Replication instances (2)

Find replication instance

Actions Create replication instance

	Name	Class	Status	Engine version	Availability zone	VPC	Public	Public IP address
<input type="checkbox"/>	dms-replication-instance	dms.t2.large	Available	3.3.1	us-east-1b	vpc-0537f7268d522ba73	Yes	35.175.68.214

Create the DMS Source Endpoint

Please proceed to create your endpoints, without waiting for the step above.

1. On the DMS console, select **Endpoints**

AWS DMS

Dashboard

▼ Conversion & migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

DMS > Endpoint

Endpoints

Find endpoint


Actions Create endpoint

	Name	Type	Status	Engine	Server name	Port	Migration Hub Mapping	ARN	Certificate ARN
Empty endpoint table You don't have any endpoints.									

2. Click **Create endpoint**.

Lab 1. Hydrating the Data Lake with DMS

- select **Source endpoint** type.
- For **Endpoint identifier**, select an easily recognizable name (e.g. **rds-source-endpoint**)
- For Source engine, select **postgres**
- For **Server name**, get the information from **RDS DB Info** module on your event engine dashboard.

RDS DB Info  [Readme](#)

Outputs:
No outputs defined

If you are running the lab outside of AWS hosted event, enter the **DMSInstanceEndpoint** parameter value from [dmslab-instructor CloudFormation Outputs](#) tab

- For Port, enter **5432**
- For SSL mode, choose **none**
- For User name, type **master**
- For Password, type **master123**
- For Database name, type **sportstickets**

AWS DMS ×

Create endpoint

Endpoint type [Info](#)

☒ **Source endpoint**
A source endpoint allows AWS DMS to read data from a database (on-premises or in the cloud), or from other data source such as Amazon S3.

☐ **Target endpoint**
A target endpoint allows AWS DMS to write data to a database, or to other data source.

☐ Select RDS DB Instance

Endpoint configuration

Endpoint identifier [Info](#)
A label for the endpoint to help you identify it.
prodendpoint-postgre

Source engine
The type of database engine this endpoint is connected to.
postgres

Server name
dmslabinstance.c1ny3gywsvdz.us-east-1.rds.amazonaws.com

Port
The port the database runs on for this endpoint.
5432

Secure Socket Layer (SSL) mode
The type of Secure Socket Layer enforcement
none

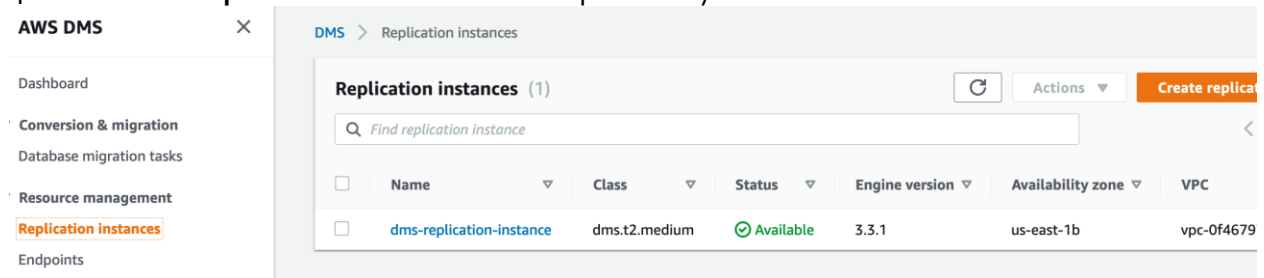
User name [Info](#)
master

Password [Info](#)

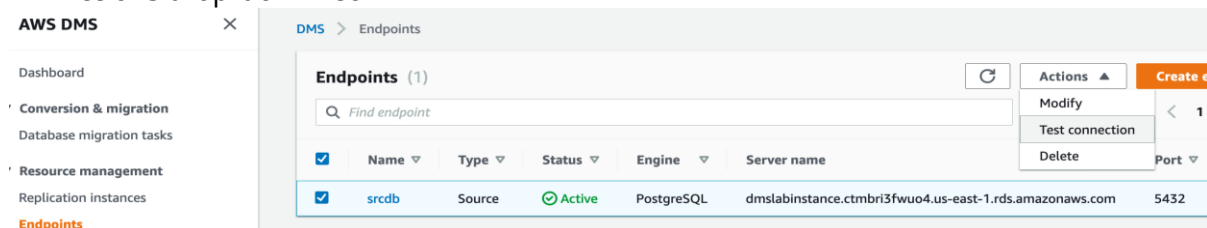
Database name
sportstickets

Lab 1. Hydrating the Data Lake with DMS

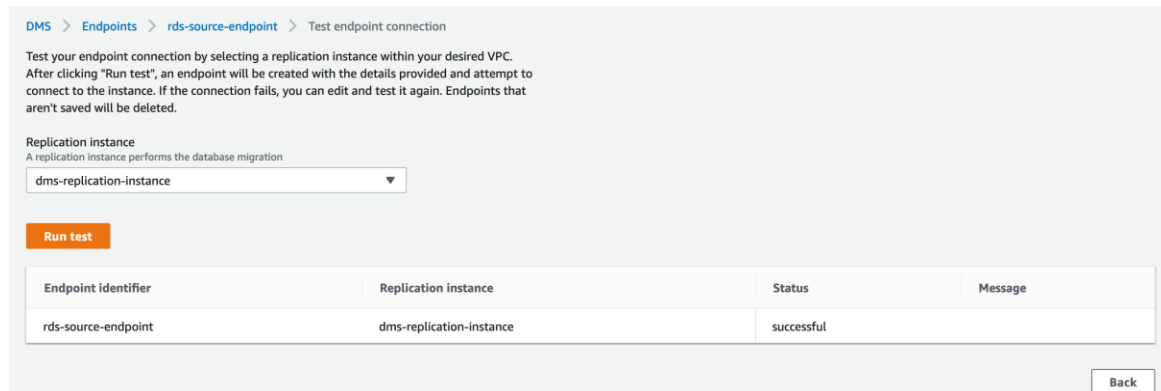
- Click **Create endpoint** to create the endpoint. When available, the endpoint status changes to **active**.
- Check the **replication instance** created previously. Make sure the status is **available**.



- Select your newly created source **endpoint**, and choose **Test connection** on the **Actions** drop-down list.



- Click **Run test**. This step tests connectivity to the source database system. If successful, the message "Connection tested successfully" appears. **You may need to wait for the DMS replication instance to become available first.**

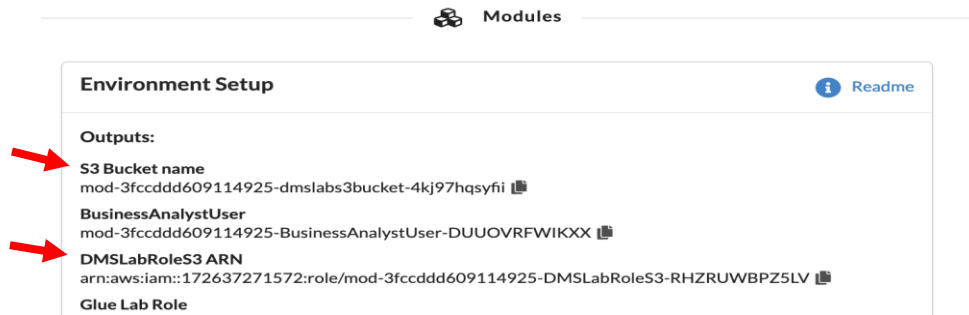


Create the Target Endpoint

Before start, make sure you have the following values handy (on your event dashboard).

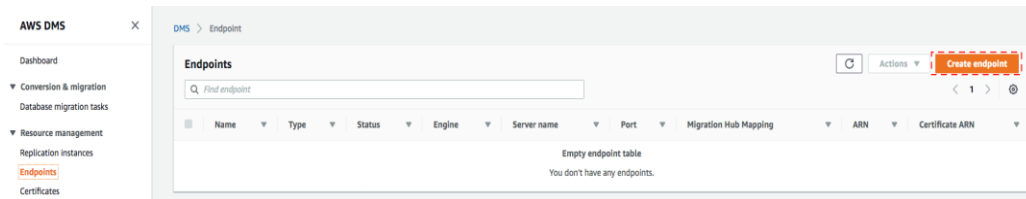
- DMSLabRoleS3** ARN – It looks like "arn:aws:iam::<account number>:role/xxx-DMSLabRoleS3-xxxx"
- BucketName** - It looks like "xxx-dmslabs3bucket-xxxx"

Lab 1. Hydrating the Data Lake with DMS



If you are running the lab outside of AWS hosted event, can find them in [dmslab-student Cloudformation](#) Outputs tab.

1. On the DMS console, select **Endpoints**.



2. Click **Create endpoint**.
 - a. For Endpoint type, select **Target endpoint**.
 - b. For Endpoint identifier, type an easily recognized name such as **s3-target-endpoint**.
 - c. For Target engine, choose **s3**.
 - d. For Service access role ARN, paste the **DMSLabRoleS3** value noted earlier
 - e. For Bucket name, paste the value of **BucketName** noted earlier
 - f. For Bucket folder, type **tickets**

Lab 1. Hydrating the Data Lake with DMS

AWS DMS ×

Dashboard

▼ Conversion & migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

Subnet groups

Events

Event subscriptions

DMS > Create endpoint

Create endpoint

Endpoint type [Info](#)

☐ Source endpoint
A source endpoint allows AWS DMS to read data from a database (on-premises or in the cloud), or from other data source such as Amazon S3.

☒ Target endpoint
A target endpoint allows AWS DMS to write data to a database, or to other data source.

☐ Select RDS DB instance

Endpoint configuration

Endpoint identifier [Info](#)
A label for the endpoint to help you identify it.
targetendpoint

Target engine
The type of database engine this endpoint is connected to.
s3 ▼

Service access role ARN
Role that can access target
arn:aws:iam::341259728059:role/dmslab-student-DMSLabRoleS3-8DVW2RR7J7QZ

Bucket name
The name of an Amazon S3 bucket where DMS will read the files from
dmslab-student-dmslabs3bucket-woti4bf73cw3

Bucket folder
The Amazon S3 bucket path where the CSV files can be found
tickets

- g. Click **Endpoint-specific settings** to expand the section.
- h. In the **Extra connection attributes** box, type **addColumnNames=true**
- i. This attribute includes the column names in the files in the S3 bucket.
- j. Expand the **Test endpoint connection (optional)** section, and choose your **VPC** name with **dmslstudv1** on the drop-down list.
- k. Click **Run test**. This step tests connectivity to the source database system. If successful, the message "Connection tested successfully" appears.

Lab 1. Hydrating the Data Lake with DMS

AWS DMS

Dashboard

▼ Conversion & migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

Subnet groups

Events

Event subscriptions

▼ **Endpoint-specific settings**

Extra connection attributes
Type any additional connection parameters here. See the documentation for more information.

addColumnName=true

▼ **Test endpoint connection (optional)**

Test your endpoint connection by selecting a replication instance within your desired VPC. After clicking "Run test", an endpoint will be created with the details provided and attempt to connect to the instance. If the connection fails, you can edit and test it again. Endpoints that aren't saved will be deleted.

VPC
vpc-0314e829ba12d9481 - dmslstudv1

Replication instance
A replication instance performs the database migration
dms-replication-instance

Run test

After clicking "Run test", an endpoint will be created with the details provided and attempt to connect to the instance. If the connection fails, you can edit and test it again. Endpoints that aren't saved will be deleted.

Endpoint identifier	Replication instance	Status	Message
targetendpoint	dms-replication-instance	successful	

Cancel **Create endpoint**

3. Click **Create Endpoint**. When available, the endpoint status changes to **active**.

AWS DMS

Dashboard

▼ Conversion & migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

DMS > Endpoint

Endpoints (2)

Find endpoint

	Name	Type	Status	Engine	Server name	Port	Migration Hub Mapping	ARN
<input type="checkbox"/>	prodendpoint-postgre	Source	Active	PostgreSQL	dmslabinstance.c1ny3gywsvdz.us-east-1.rds.amazonaws.com	5432		arn:aws:dms:sus-east-1:341259728059:endp
<input type="checkbox"/>	targetendpoint	Target	Active	Amazon S3	-	-		arn:aws:dms:sus-east-1:341259728059:endp

Create a task to perform the initial full copy

1. On the DMS console, select **Database Migration Tasks**.

AWS DMS

Dashboard

▼ Conversion & migration

Database migration tasks

Replication instances

Endpoints

Certificates

DMS > Database migration tasks

Database migration tasks

Find task

	Name	Status	Source	Target	Type	Progress	Elapsed time	Tables loaded	Tables loading	Tables queued	Tables errored
Empty replication task table You don't have any replication tasks.											

Lab 1. Hydrating the Data Lake with DMS

2. Click **Create Task**.

- Type an easily recognized **Task name** e.g. **dms-full-dump-task**
- Select your **Replication instance** from drop down.
- Select your **Source endpoint** from drop down.
- Select your **Target endpoint** from drop down.
- For, **Migration type** choose **Migrate existing data**.

The screenshot shows the AWS DMS console interface. On the left is a navigation sidebar with 'AWS DMS' at the top, followed by 'Dashboard', 'Migration' (expanded), 'Database migration tasks' (selected), 'Resource management', and 'What's new'. The main content area is titled 'Create database migration task' and contains a 'Task configuration' section with the following fields:

- Task identifier:
- Replication instance:
- Source database endpoint:
- Target database endpoint:
- Migration type:

- Under **Task Settings**, select the **Enable CloudWatch logs** check box.

The screenshot shows the 'Task settings' section of the AWS DMS console. It includes the following configuration options:

- Target table preparation mode: ☒ Drop tables on target
- Include LOB columns in replication: ☒ Limited LOB mode
- Maximum LOB size (KB):
- Enable validation: ☐
- Enable CloudWatch logs: ☒

A blue information box at the bottom states: 'CloudWatch logs usage will be charged at standard rates. See here for more details.'

Lab 1. Hydrating the Data Lake with DMS

- g. Go to **Table Mappings**.
- h. Click on **Add new selection rule** and select **Enter a Schema** in **Schema** field.
- i. For Schema name, type **dms_sample** and keep the settings for the remaining fields

▼ Table mappings

Editing mode

☒ Guided UI
Set up your table mapping rules using a step-by-step guided interface.

☐ JSON editor [Learn more](#)
Enter your table mapping rules directly, in JSON format.

Specify at least one selection rule with an include action. After you do this, you can add one or more transformation rules.

▼ Selection rules

Choose the schema and/or tables you want to include with, or exclude from, your migration task. [Info](#) Add new selection rule

▼ where schema name is like '%' and table name is like %, include

Schema
Enter a schema

Schema name
Use the % character as a wildcard
dms_sample

Table name
Use the % character as a wildcard
%

Action
Choose "Include" to migrate your selected objects, or "Exclude" to ignore them during the migration.
Include

3. Click **Create task**. Your task is created and starts automatically. (Note: The complete creation and data extraction process takes around 5 minutes.)
4. Once complete, the console displays 100% complete.

DMS > Database migration tasks

Database migration tasks (1)

[Refresh](#) [Actions](#) [Create task](#)

<input type="checkbox"/>	Name	Status	Source	Target	Type	Progress	Elapsed time	Tables loaded	Tables loading	Tables queue
<input type="checkbox"/>	dmstask	Load complete	src-rds	targets3	Full load	100%	5 m	16	0	

5. Select your task and explore the summary. Under **Table** statistics tab you can review all table information loaded in S3 from RDS by DMS

DMS > Database migration tasks > dms-full-dump-task

dms-full-dump-task

[Refresh](#) [Actions](#)

Summary

Status: Load complete Type: Full load Source: rds-source-dms

[Overview details](#) [Table statistics](#) [CloudWatch metrics](#) [Mapping rules](#) [Premigration assessments](#) [Assessment](#)

Overview details

Basic configuration

Task ARN
arn:aws:dms:us-east-1:172637271572:task:3RMQ2OYSPGZKP5R5AFNYM5T45MCBDVHRX7QMTWY

Progress
100%

Last failure message
-

Replication instance
[dms-replication-instance](#)

Created
8/21/2020, 5:51:01 PM GMT+1000

Migration task logs [Info](#)

[View CloudWatch logs](#)

Lab 1. Hydrating the Data Lake with DMS

Summary

Status

✔ Load complete

Type

Full load

Source

rds-source-dms

Target

s3-target

Overview details

Table statistics

CloudWatch metrics

Mapping rules

Premigration assessments

New

Assessment results

Tags

Table statistics (16)

↻

Validate again

Reload table d

Total rows include loaded source table rows from Inserts, Deletes, Updates, DDLs, and Full load rows.

🔍 Find schema

<input type="checkbox"/>	Schema name ▾	Table ▾	Load state ▾	Inserts ▾	Deletes ▾	Updates ▾	DDLs ▾	Full load rows ▾	Total rows ▾
<input type="checkbox"/>	dms_sample	seat_type	Table completed	0	0	0	0	6	6
<input type="checkbox"/>	dms_sample	seat	Table completed	0	0	0	0	594,152	594,152
<input type="checkbox"/>	dms_sample	mlb_data	Table completed	0	0	0	0	2,230	2,230
<input type="checkbox"/>	dms_sample	player	Table completed	0	0	0	0	5,157	5,157
<input type="checkbox"/>	dms_sample	ticket_purchase_hist	Table completed	0	0	0	0	5,219,270	5,219,270
<input type="checkbox"/>	dms_sample	person	Table completed	0	0	0	0	7,025,584	7,025,584

- Open the S3 console to view the data that was copied by DMS. The s3 bucket looks like `<random_string>-dmslabs3bucket-<random_string>`
- Click on the bucket used as the DMS target and navigate to `/tickets/dms_sample/` to view the loaded tables, one folder per table

Amazon S3 > mod-3fccddd609114925-dmslabs3bucket-4kj97hqsyfii > tickets > dms_sample

mod-3fccddd609114925-dmslabs3bucket-4kj97hqsyfii

Overview

Q












Type a prefix and press Enter to search. Press ESC to clear.

Upload

Create folder

Download

Actions


<input type="checkbox"/>	Name	Last modified
<input type="checkbox"/>	 mlb_data	--
<input type="checkbox"/>	 name_data	--
<input type="checkbox"/>	 nfl_data	--
<input type="checkbox"/>	 nfl_stadium_data	--
<input type="checkbox"/>	 person	--
<input type="checkbox"/>	 player	--
<input type="checkbox"/>	 seat	--
<input type="checkbox"/>	 seat_type	--
<input type="checkbox"/>	 sport_division	--
<input type="checkbox"/>	 sport_league	--
<input type="checkbox"/>	 sport_location	--

8. Download one of the files:
 - a. Navigate further to **mlb_data/LOAD00000001.csv**, select the check box next to the file name and click **Download** in the pop-up window.
 - b. Click **Save File**.
 - c. Open the file.

You will notice that the file contains the column headers in the first row as requested by the "addColumnNames=true" connection attribute we included when we created the s3 target endpoint. Note that column names are included in the file in the first row.

	A	B	C	D	E
1	id	sport_team_id	last_name	first_name	full_name
2	1	131	Adam Loewen	Adam	Loewen
3	11	131	A.J. Pollock	A.J.	Pollock
4	21	131	Alex Sanabia	Alex	Sanabia
5	31	131	Andrew Chafin	Andrew	Chafin
6	41	131	Andy Marte	Andy	Marte
7	51	131	Archie Bradley	Archie	Bradley
8	61	131	Ben Francisco	Ben	Francisco
9	71	131	Braden Shipley	Braden	Shipley
10	81	131	Bradin Hagens	Bradin	Hagens
11	91	131	Brandon Drury	Brandon	Drury
12	101	131	Brett Jackson	Brett	Jackson

You may notice that the primary key column was loaded in scientific notation:

 LOAD00000001.csv - Notepad

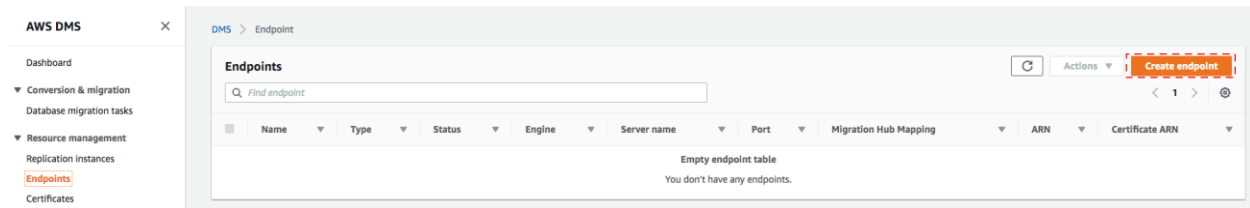
File Edit Format View Help

```
mlb_id,mlb_name,mlb_pos,mlb_team,mlb_team_long,bats,throws,birth_year,bp
+5.065600000000000e+05,Alexi Amarista,3B,SD,San Diego Padres,L,R,1989,+
+4.582100000000000e+05,Alexi Casilla,2B,TB,Tampa Bay Rays,S,R,1984,+4.5
+4.683960000000000e+05,Alexi Ogando,P,ATL,Atlanta Braves,R,R,1983,+4.9
+4.696860000000000e+05,Alfredo Aceves,P,NYY,New York Yankees,R,R,1982,+
+4.516280000000000e+05,Alfredo Figaro,P,TEX,Texas Rangers,R,R,1984,+5.2
+5.540540000000000e+05,Alfredo Gonzalez,C,CWS,Chicago White Sox,R,R,19
+5.012450000000000e+05,Alfredo Marte,LF,BAL,Baltimore Orioles,R,R,1989,
+4.305800000000000e+05,Alfredo Simon,P,CIN,Cincinnati Reds,R,R,1981,+4.
+4.553780000000000e+05,Ali Solis,C,LAD,Los Angeles Dodgers,R,R,1987,+5.
+4.888520000000000e+05,Allan Dykstra,1B,TB,Tampa Bay Rays,L,R,1987,+5.7
+5.018000000000000e+05,Allen Craig,1B,BOS,Boston Red Sox,R,R,1984,+5.16
```

This is due to the tables at the source having primary key as **double precision**. Keep in mind that DMS allows you to perform additional transformations, for example type casting at load time. Here we will proceed without making any further modifications.

(Optional) Create a DMS endpoint to perform ongoing replication

1. Navigate to the DMS console: <https://console.aws.amazon.com/dms/v2/home?region=us-east-1#dashboard> and select **Endpoints**:



2. Click **Create endpoint**.
 - a. For **Endpoint type**, select **Target**
 - b. For **Endpoint identifier**, type **rds-cdc-endpoint**
 - c. For **Target engine**, choose **s3**.
 - d. For Service access role ARN, paste the **DMSLabRoleS3** number noted earlier
 - e. For Bucket name, paste the **S3 Bucket Name** noted earlier
 - f. For **Bucket folder**, type **cdc**.

Create endpoint

Endpoint type [Info](#)

☐ Source endpoint
 A source endpoint allows AWS DMS to read data from a database (on-premises or in the cloud), or from other data source such as Amazon S3.

☒ Target endpoint
 A target endpoint allows AWS DMS to write data to a database, or to other data source.

☐ Select RDS DB instance

Endpoint configuration

Endpoint identifier [Info](#)
A label for the endpoint to help you identify it.

s3cdc

Target engine
The type of database engine this endpoint is connected to.

s3

Service access role ARN
Role that can access target

arn:aws:iam::720560070661:role/dmslab-student-DMSLabRoleS3-6WRA37YEG4SI

Bucket name
The name of an Amazon S3 bucket where DMS will read the files from

dmslab-student-dmslabs3bucket-tjtm2ypm2jvr

Bucket folder
The Amazon S3 bucket path where the CSV files can be found

cdc

- g. Click **Endpoint-specific settings** to expand the section.

Lab 1. Hydrating the Data Lake with DMS

- h. In the **Extra connection attributes** box, type **addColumnNames=true** to include column names in the files in the S3 bucket.
- i. Expand the **Test endpoint connection (optional)** section, and choose your **dmslstudv1** name on the VPC drop-down list.
- j. Click **Run test**. This step tests connectivity to the source database system. If successful, the message "Connection tested successfully" appears.

▼ Endpoint-specific settings

Extra connection attributes
Type any additional connection parameters here. See the documentation for more information.

▼ Test endpoint connection (optional)

Test your endpoint connection by selecting a replication instance within your desired VPC. After clicking "Run test", an endpoint will be created with the details provided and attempt to connect to the instance. If the connection fails, you can edit and test it again. Endpoints that aren't saved will be deleted.

VPC

Replication instance
A replication instance performs the database migration

Run test

After clicking "Run test", an endpoint will be created with the details provided and attempt to connect to the instance. If the connection fails, you can edit and test it again. Endpoints that aren't saved will be deleted.

Endpoint identifier	Replication instance	Status	Message
cdcendpoint	dms-replication-instance	successful	

Cancel **Create endpoint**

3. Click **Create endpoint**.
4. When available, the endpoint status changes to active.

DMS > Endpoints

Endpoints (3) Find endpoint Actions Create endpoint

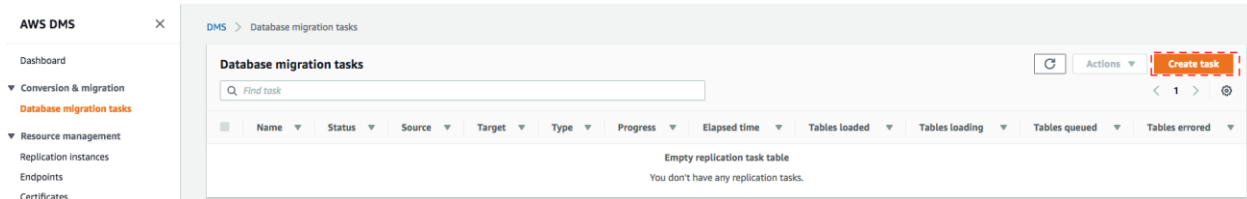
	Name	Type	Status	Engine	Server name	Port	Migration Hub Mapping	ARN
<input checked="" type="checkbox"/>	rds-cdc-endpoint	Target	Active	Amazon S3	-	-		arn:aws:dms:us-east-1:132701118127:endpoint:QCCRWAZQTXI
<input type="checkbox"/>	rds-source-endpoint	Source	Active	PostgreSQL	dmslabinstance.c8msbe8b7bwx.us-east-1.rds.amazonaws.com	5432		arn:aws:dms:us-east-1:132701118127:endpoint:SWCXG2HRI7A
<input type="checkbox"/>	s3-target-endpoint	Target	Active	Amazon S3	-	-		arn:aws:dms:us-east-1:132701118127:endpoint:T53323213IIX5F

(Optional) Create a task to perform ongoing replication

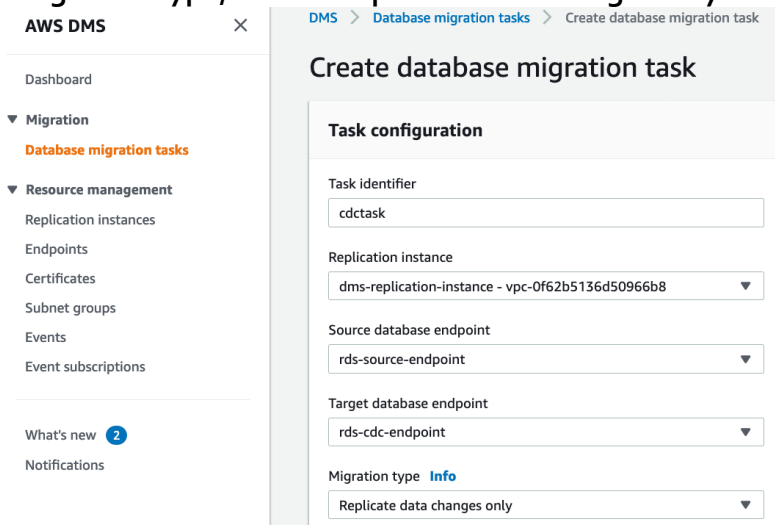
Lab 1. Hydrating the Data Lake with DMS

Before start the lab, ask your instructor generate some new data in the source database.

1. Navigate to the DMS console: <https://console.aws.amazon.com/dms/v2/home?region=us-east-1#dashboard> and select **Database Migration Tasks**.

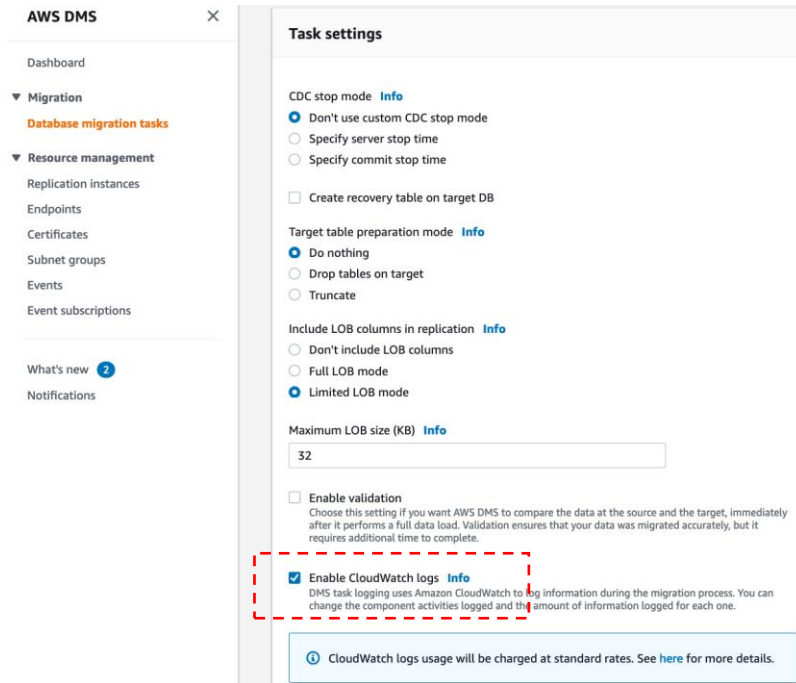


2. Click **Create Task**.
 - a. Type **cdctask** as **Task Identifier**
 - b. Select your **Replication instance**.
 - c. Select your **Source endpoint**.
 - d. Select **Target endpoint** as **rds-cdc-endpoint** created in the previous section.
 - e. For **Migration type**, choose **Replicate data changes only**.



- f. In **Task Settings**, Select the **Enable CloudWatch logs** check box. Do not enable the validation.

Lab 1. Hydrating the Data Lake with DMS



AWS DMS

Dashboard

▼ Migration

Database migration tasks

▼ Resource management

Replication instances

Endpoints

Certificates

Subnet groups

Events

Event subscriptions

What's new 2

Notifications

Task settings

CDC stop mode [Info](#)

☒ Don't use custom CDC stop mode

☐ Specify server stop time

☐ Specify commit stop time

☐ Create recovery table on target DB

Target table preparation mode [Info](#)

☒ Do nothing

☐ Drop tables on target

☐ Truncate

Include LOB columns in replication [Info](#)

☐ Don't include LOB columns

☐ Full LOB mode

☒ Limited LOB mode

Maximum LOB size (KB) [Info](#)

32

☐ Enable validation

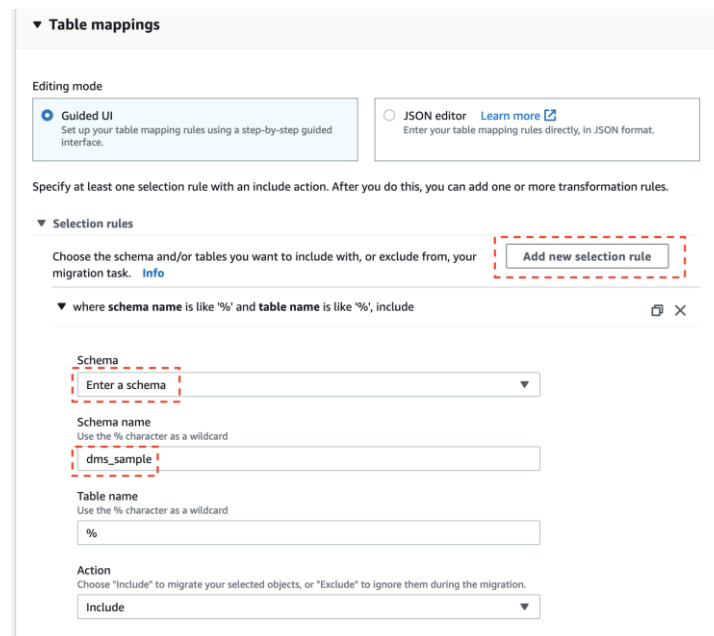
Choose this setting if you want AWS DMS to compare the data at the source and the target, immediately after it performs a full data load. Validation ensures that your data was migrated accurately, but it requires additional time to complete.

☒ Enable CloudWatch logs [Info](#)

DMS task logging uses Amazon CloudWatch to log information during the migration process. You can change the component activities logged and the amount of information logged for each one.

CloudWatch logs usage will be charged at standard rates. See [here](#) for more details.

- g. Go to **Table Mappings**.
- h. Click on **Add new selection rule** and select **Enter a Schema** in Schema field.
- i. For **Schema name**, type **dms_sample** and keep the values in the remaining fields



▼ Table mappings

Editing mode

☒ Guided UI

Set up your table mapping rules using a step-by-step guided interface.

☐ JSON editor [Learn more](#)

Enter your table mapping rules directly, in JSON format.

Specify at least one selection rule with an include action. After you do this, you can add one or more transformation rules.

▼ Selection rules

Choose the schema and/or tables you want to include with, or exclude from, your migration task. [Info](#)

[Add new selection rule](#)

▼ where schema name is like '%' and table name is like '%', include

Schema

Enter a schema

Schema name

Use the % character as a wildcard

dms_sample

Table name

Use the % character as a wildcard

%

Action

Choose "Include" to migrate your selected objects, or "Exclude" to ignore them during the migration.

Include

3. Click **Create task**. Your task is created and starts automatically. You can see status as **ongoing replication**, after couple of minutes. Once complete, the console displays 100% complete.

Lab 1. Hydrating the Data Lake with DMS

DMS > Database migration tasks

Database migration tasks (2)

Find task

< 1 > ⚙

	Name	Status	Source	Target	Type	Progress	Elapsed time	Tables loaded	Tables loading	Tables queued
<input type="checkbox"/>	dms-task	Load complete	prodendpoint-postgre	targetendpoint	Full load	100 %	9 m	16	0	0
<input checked="" type="checkbox"/>	newcdc	Replication ongoing	prodendpoint-postgre	cdcendpoint	Ongoing replication	100 %	0 m	16	0	0

- By now, your instructor has generated some CDC activity, which above migration task will capture. You may need to wait for 5 to 10 minutes for the new data to be picked up.
- Once the CDC data gets replicated, you can navigate to CDC task details, and under **Table statistics** tab review the details, as shown below:

Note: In case you see DMS CDC task in fail/error status. Make sure your replication instance version is 3.3.1 and it is large enough (dms.t2.medium or above) to run CDC replication task

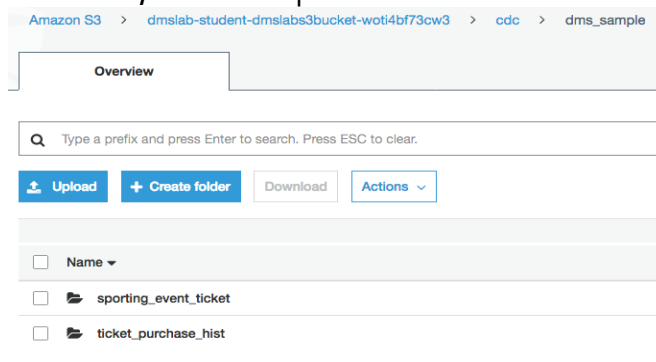
Table statistics (16)

Find schema

< 1 > ⚙

	Schema name	Table	Load state	Inserts	Deletes	Updates	DDLs	Full load rows	Total	Validation state	Validation pending
<input type="checkbox"/>	dms_sample	seat_type	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	seat	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	mlb_data	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	player	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	ticket_purchase_hist	Table completed	11,002	0	0	0	0	11,002	Not enabled	0
<input type="checkbox"/>	dms_sample	person	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	name_data	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	sport_team	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	sport_league	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	sporting_event	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	sporting_event_ticket	Table completed	0	0	11,002	0	0	11,002	Not enabled	0
<input type="checkbox"/>	dms_sample	sport_division	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	sport_location	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	sport_type	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	nfl_stadium_data	Table completed	0	0	0	0	0	0	Not enabled	0
<input type="checkbox"/>	dms_sample	nfl_data	Table completed	0	0	0	0	0	0	Not enabled	0

- Open the [S3 console](#) to view the data that was copied by DMS:
- Click on the bucket used as the DMS target and navigate to `/cdc/dms_sample/` to view the loaded tables, one folder per table



Lab 1. Hydrating the Data Lake with DMS

8. Download one of the files:
 - a. Select the check box next to the object name and click Download in the pop-up window.
 - b. Click **Save File**.
 - c. Open the file.

You will notice that the file contains the column headers in the first row as requested by the **addColumnNames=true** connection attribute we included when we created the s3 target endpoint.

The screenshot shows the Amazon S3 console interface. The breadcrumb trail is: Amazon S3 > dmslab-student-dmslab3bucket-wot4b73cw3 > cdc > dms_sample > sporting_event_ticket. The 'Overview' tab is active. A search bar is present. Below the search bar are buttons for 'Upload', '+ Create folder', 'Download', and 'Actions'. A table lists two files:

Name	Last modified	Size
<input checked="" type="checkbox"/> 20190529-230604667.csv	May 29, 2019 4:06:05 PM GMT-0700	19.3 MB
<input type="checkbox"/> 20190529-230729693.csv	May 29, 2019 4:07:30 PM GMT-0700	53.9 MB

A pop-up window for the selected file '20190529-230604667.csv' is shown on the right. It includes buttons for 'Download', 'Copy path', and 'Select from'. Below these buttons, it says 'Latest version'. Further down, there is an 'Overview' section with details:

- Key: 20190529-230604667.csv
- Size: 19.3 MB
- Expiration date: N/A
- Expiration rule: N/A
- ETag: ad6b57f453cd4e9745035e8d50dfb3bc-4
- Last modified: May 29, 2019 4:06:05 PM GMT-0700
- Object URL: https://s3.amazonaws.com/dmslab-student-dmslab3bucket-wot4b73cw3/cdc/dms_sample/sporting_event_ticket/20190529-230604667.csv

Note that file name has a timestamp. You can see the header is included and the operation column is added at the beginning of each row. The file below shows updates (U) to the table along with the values after the update. Inserts (I) show data after the insert and Deletes (D) show data before the delete.

	A	B	C	D	E	F	G	H	I	J
	Op	id	sporting_event_id	sport_location_id	seat_level	seat_section	seat_row	seat	ticketholder_id	ticket_price
1	U	145192591	3931	4	2	10 A		2	2898028	98
2	U	145192601	3931	4	2	10 A		1	2898028	98
3	U	145192581	3931	4	2	10 A		3	2898028	98
4	U	145192501	3931	4	2	10 B		1	2898028	98
5	U	145187751	3931	4	2	13 B		2	2898028	49
6	U	145187741	3931	4	2	13 B		3	2898028	49
7	U	145187721	3931	4	2	13 C		2	2898028	49
8	U	145187711	3931	4	2	13 C		3	2898028	49
9	U	145187731	3931	4	2	13 C		1	2898028	49
10	U	145187701	3931	4	2	14 A		1	2898028	49
11	U	145187681	3931	4	2	14 A		3	2898028	49
12	U	145187691	3931	4	2	14 A		2	2898028	49
13	U	145187471	3931	4	2	14 B		3	2898028	49
14	U	145187671	3931	4	2	14 B		1	2898028	49
15	U	145187481	3931	4	2	14 B		2	2898028	49
16	U	145187451	3931	4	2	14 C		2	2898028	49
17	U	145187461	3931	4	2	14 C		1	2898028	49
18	U	145190341	3931	4	2	14 C		3	2898028	49
19	U	145183201	3931	4	2	15 A		4	2898028	49
20	U	145179691	3931	4	2	15 A		1	2898028	49
21	U	145179661	3931	4	2	15 A		4	2898028	49
22	U	145179671	3931	4	2	15 A		3	2898028	49
23	U	145179681	3931	4	2	15 A		2	2898028	49
24	U	145190321	3931	4	2	15 A		2	2898028	49