



# **Amazon Web Services**

## **Data Engineering Immersion Day**

---

Lab 1. Hydrating the Data Lake with DMS

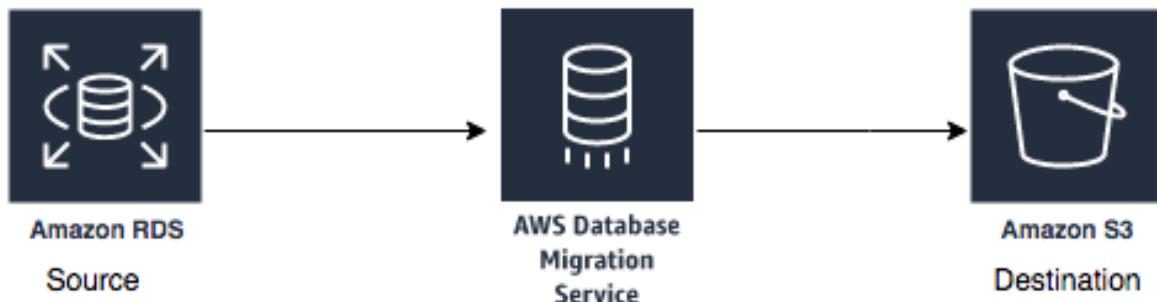
***August 2020***

## Table of Contents

<i>Introduction</i> .....	2
<i>Get Started Using the Lab Environment</i> .....	3
<i>Create the Subnet Group</i> .....	6
<i>Create the Replication Instance</i> .....	7
<i>Create the DMS Source Endpoint</i> .....	9
<i>Create the Target Endpoint</i> .....	11
<i>Create a task to perform the initial full copy</i> .....	13
<i>(Optional) Create a DMS endpoint to perform ongoing replication</i> .....	18
<i>(Optional) Create a task to perform ongoing replication</i> .....	20

## 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>

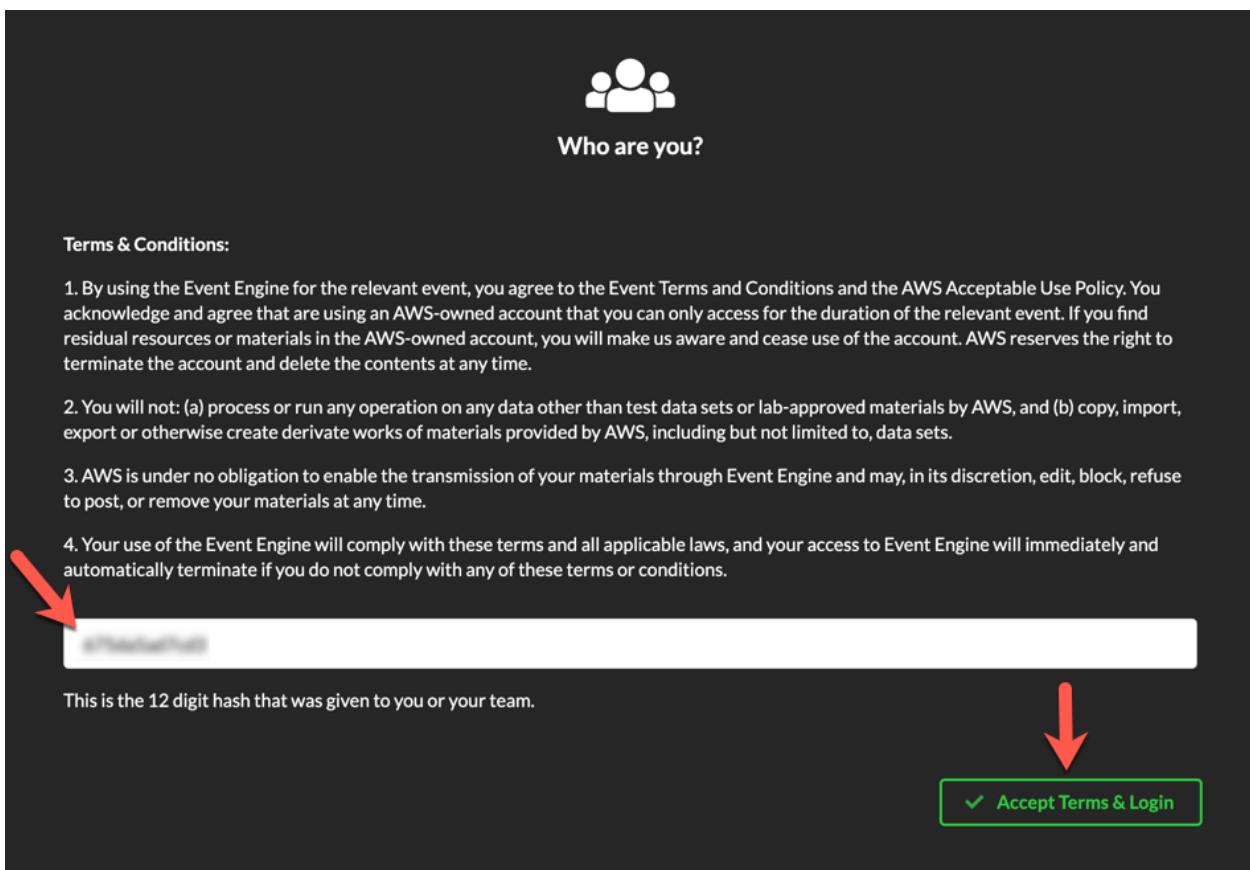
## Get Started Using the Lab Environment

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:



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

Outputs:

No outputs defined

Readme

- S3 Bucket, IAM role for the DMS lab etc

Environment Setup

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

Readme

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

Team Dashboard

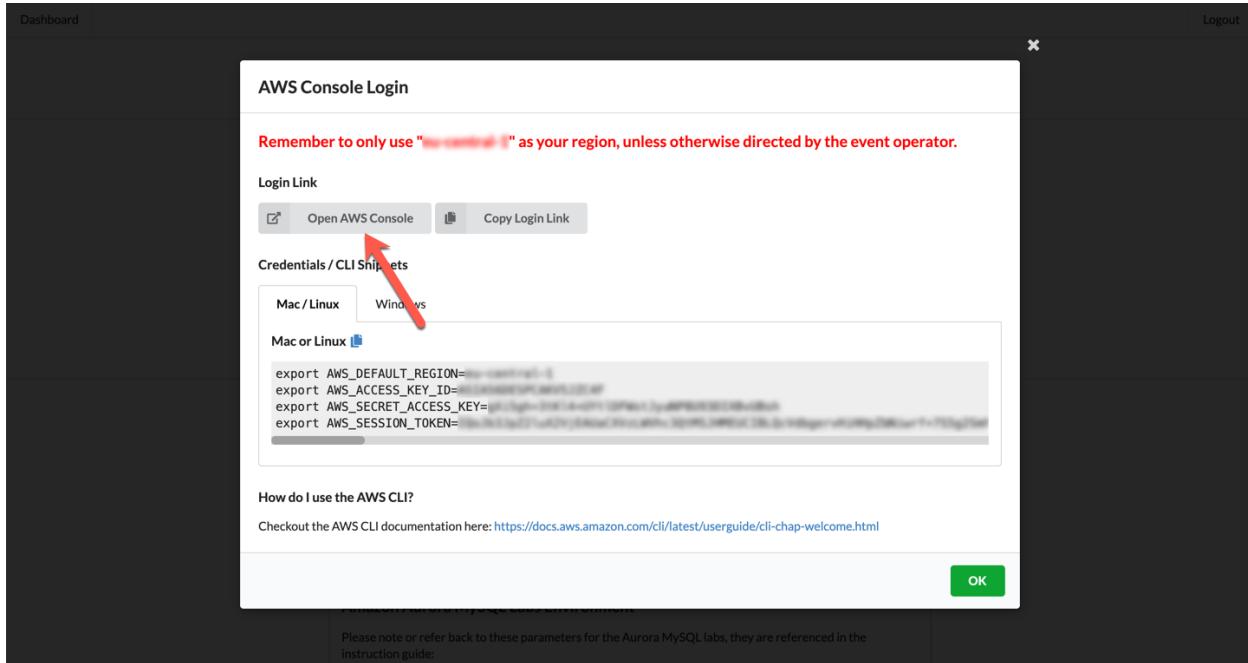
Event

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.

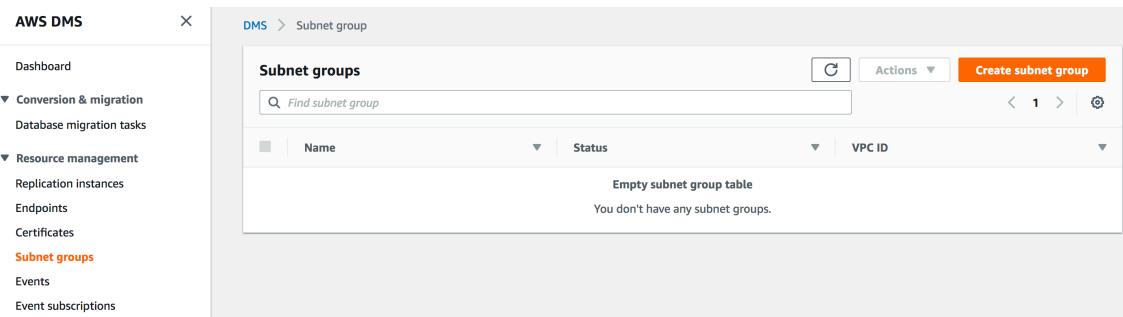
## Lab 1. Hydrating the Data Lake with DMS

### Create the Subnet Group

1. Navigate to the DMS Console:

<https://console.aws.amazon.com/dms/v2/home?region=us-east-1#firstRun>

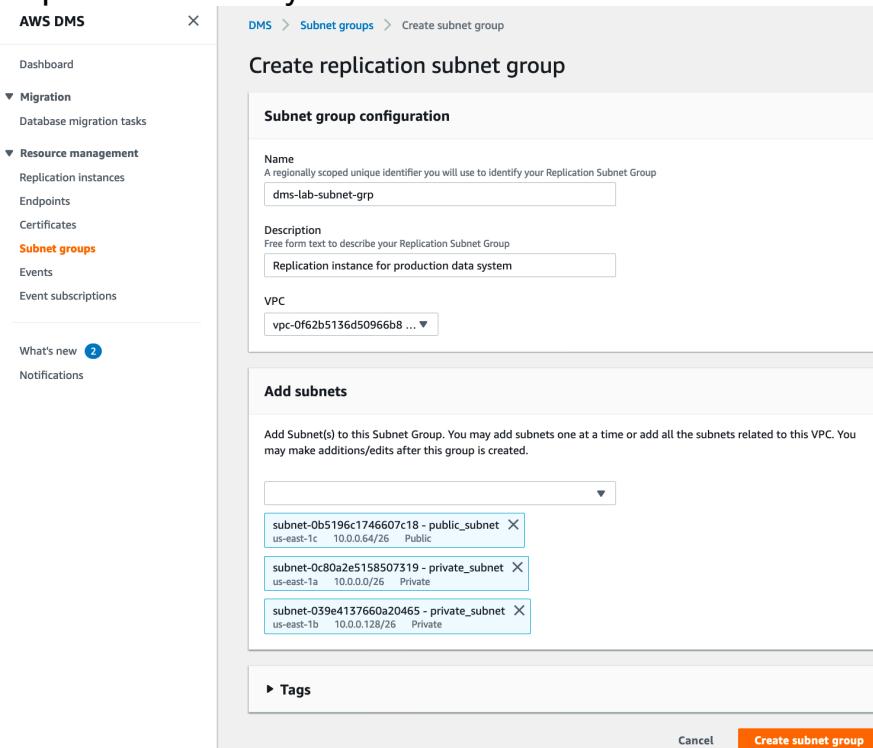
2. On the DMS console, select **Subnet Groups**.



The screenshot shows the AWS DMS Subnet Groups page. The left sidebar has a 'Subnet groups' section highlighted in orange. The main area is titled 'Subnet groups' and contains a search bar and a table header with columns for Name, Status, and VPC ID. Below the table, a message says 'Empty subnet group table' and 'You don't have any subnet groups.'

3. Click **Create subnet group**.

- In the Identifier box, type a descriptive name that you will easily recognize (e.g., "dms-lab-subnet-grp").
- In the Description box, type an easily recognizable description (e.g., "**Replication instance for production data system**").
- For VPC, select the pre-created VPC ending with **dmslstudv1**. The subnet list populates in the Available Subnets pane.
- 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.



The screenshot shows the 'Create replication subnet group' dialog box. It has two main sections: 'Subnet group configuration' and 'Add subnets'. In 'Subnet group configuration', the 'Name' field is set to 'dms-lab-subnet-grp' and the 'Description' field is set to 'Replication instance for production data system'. In the 'VPC' dropdown, 'vpc-0f62b5136d50966b8 ...' is selected. In the 'Add subnets' section, three subnets are listed: 'subnet-0b5196c1746607c18 - public\_subnet' (us-east-1c, 10.0.64/26, Public), 'subnet-0c80a2e5158507319 - private\_subnet' (us-east-1a, 10.0.0/26, Private), and 'subnet-039e4137660a20465 - private\_subnet' (us-east-1b, 10.0.128/26, Private). At the bottom right are 'Cancel' and 'Create subnet group' buttons.

## Lab 1. Hydrating the Data Lake with DMS

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

The screenshot shows the AWS DMS Subnet group list. At the top, there's a search bar labeled 'Find subnet group'. Below it is a table with columns: Name, Status, and VPC ID. There is one row in the table:

Name	Status	VPC ID
dms-lab-subnet-grp	Complete	vpc-0314e829ba12d9481

## Create the Replication Instance

1. On the DMS console, select Replication instances.

The screenshot shows the AWS DMS Replication instance list. On the left is a sidebar with 'AWS DMS' at the top, followed by 'Dashboard', 'Conversion & migration' (with 'Database migration tasks' under it), 'Resource management' (with 'Replication instances' under it, which is highlighted), 'Endpoints', and 'Certificates'. The main area has a header 'Replication instances' with a search bar 'Find replication instance'. Below is a table with columns: Name, Class, Status, Engine version, Availability zone, VPC, Public, Public IP address, and Private IP address. A message at the bottom says 'Empty replication instance table' and 'You don't have any replication instances.'

2. Click Create replication instance.

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

## 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
- 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.

- f. Click **Advanced** to expand the section.
- g. Select the security group with **sgdefault** in the name.

**AWS DMS**

- Dashboard
- ▼ Conversion & migration
  - Database migration tasks
- ▼ Resource management
  - Replication instances**
  - Endpoints
  - Certificates
  - Subnet groups
  - Events
  - Event subscriptions

**Advanced security and network configuration**

**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.

**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.

**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.

**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.

**KMS master key** [Info](#)  
(Default) aws/dms

Account  
Description  
Key ARN

**Maintenance**

[Cancel](#) **Create**

3. Click **Create**.

- 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.

Replication instances (2)									
	Name	Class	Status	Engine version	Availability zone	VPC	Public	Public IP address	Actions
	dms-replication-instance	dms.t2.large	Available	3.3.1	us-east-1b	vpc-0537f7268d522baf3	Yes	35.175.68.214	<a href="#">Create replication instance</a>

## Create the DMS Source Endpoint

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

- On the DMS console, select **Endpoints**

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

- Click **Create endpoint**.

- 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**

**Outputs:**

No outputs defined

If you are running the lab yourself, 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**



## Lab 1. Hydrating the Data Lake with DMS

### i. For Database name, type **sportstickets**

The screenshot shows the 'Create endpoint' wizard in the AWS DMS console. The left sidebar shows the navigation menu with 'Endpoints' selected. The main form is titled 'Create endpoint' and has two tabs: 'Endpoint type' and 'Endpoint configuration'. Under 'Endpoint type', the 'Source endpoint' radio button is selected. In the 'Endpoint configuration' tab, the 'Endpoint identifier' is 'prodendpoint-postgre', 'Source engine' is 'postgres', 'Server name' is 'dmslabinstance.c1ny3gywsvdz.us-east-1.rds.amazonaws.com', 'Port' is '5432', 'User name' is 'master', 'Database name' is 'sportstickets', and 'Secure Socket Layer (SSL) mode' is set to 'none'.

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

The screenshot shows the 'Replication instances' page in the AWS DMS console. The left sidebar shows the navigation menu with 'Replication instances' selected. The main table lists one replication instance: 'dms-replication-instance' (Status: Available, Engine version: 3.3.1, Availability zone: us-east-1b, VPC: vpc-0f4679).

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

The screenshot shows the 'Endpoints' page in the AWS DMS console. The left sidebar shows the navigation menu with 'Endpoints' selected. The main table lists one endpoint: 'srcdb' (Name: srcdb, Type: Source, Status: Active, Engine: PostgreSQL, Server name: dmslabinstance.ctmbri3fwuo4.us-east-1.rds.amazonaws.com, Port: 5432).

## Lab 1. Hydrating the Data Lake with DMS

6. 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.**

The screenshot shows the AWS DMS console with the path: DMS > Endpoints > rds-source-endpoint > Test endpoint connection. The page title is "Test endpoint connection". Below it, a note says: "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." A dropdown menu labeled "Replication instance" is set to "dms-replication-instance". A large orange "Run test" button is visible. Below the button is a table with one row:

Endpoint identifier	Replication instance	Status	Message
rds-source-endpoint	dms-replication-instance	successful	

A "Back" button is located at the bottom right of the table area.

## Create the Target Endpoint

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

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

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

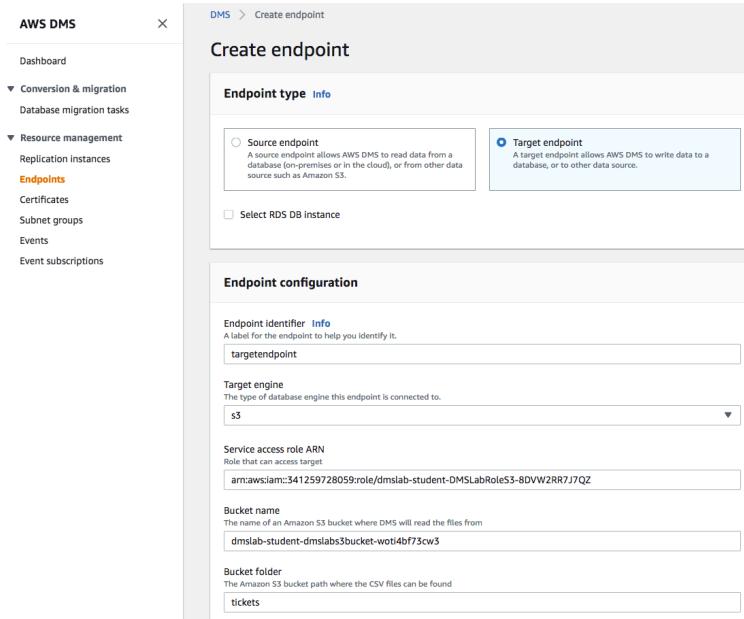
1. On the DMS console, select **Endpoints**.

The screenshot shows the AWS DMS console under the "Endpoints" section. The left sidebar has "Dashboard", "Conversion & migration", "Database migration tasks", "Resource management", "Replication instances", "Endpoints" (which is highlighted in orange), and "Certificates". The main area is titled "Endpoints" with a search bar and a "Create endpoint" button. Below the search bar is a table header with columns: Name, Type, Status, Engine, Server name, Port, Migration Hub Mapping, ARN, and Certificate ARN. The table body displays the message: "Empty endpoint table" and "You don't have any 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**.

## Lab 1. Hydrating the Data Lake with DMS

- 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**



- g. Click **Endpoint-specific settings** to expand the section.
- h. In the **Extra connection attributes** box, type **addColumnName=true**. This attribute includes the column names in the files in the S3 bucket.
- i. Expand the **Test endpoint connection (optional) section**, and choose your VPC name with **dmslstudv1** 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.

## Lab 1. Hydrating the Data Lake with DMS

The screenshot shows the 'AWS DMS' console with the 'Endpoints' section selected. A new endpoint is being created. The 'Test endpoint connection' step is active, showing a dropdown for 'VPC' containing 'vpc-0314e829ba12d9481 - dmslstudv1'. Below it, a dropdown for 'Replication instance' contains 'dms-replication-instance'. An orange 'Run test' button is highlighted with a red dashed box. Below the button, a note states: '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.' A table below shows the endpoint details: 'targetendpoint' (Identifier), 'dms-replication-instance' (Replication instance), 'successful' (Status). The 'Create endpoint' button at the bottom right is also highlighted with a red dashed box.

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

The screenshot shows the 'AWS DMS' console with the 'Endpoints' section selected. Two endpoints are listed: 'prodendpoint-postgre' (Source, Active, PostgreSQL) and 'targetendpoint' (Target, Active, Amazon S3). The 'Create endpoint' button is visible at the top right.

Name	Type	Status	Engine	Server name	Port	Migration Hub Mapping	ARN
prodendpoint-postgre	Source	Active	PostgreSQL	dmslabinstance.c1ny3gywsvdz.us-east-1.rds.amazonaws.com	5432		arn:aws:dms:us-east-1:541259728059:endpoint:prodendpoint-postgre
targetendpoint	Target	Active	Amazon S3	-	-		arn:aws:dms:us-east-1:541259728059:endpoint:targetendpoint

### Create a task to perform the initial full copy

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

The screenshot shows the 'AWS DMS' console with the 'Database migration tasks' section selected. The table is empty, displaying the message: 'Empty replication task table' and 'You don't have any replication tasks.' The 'Create task' button is visible at the top right.

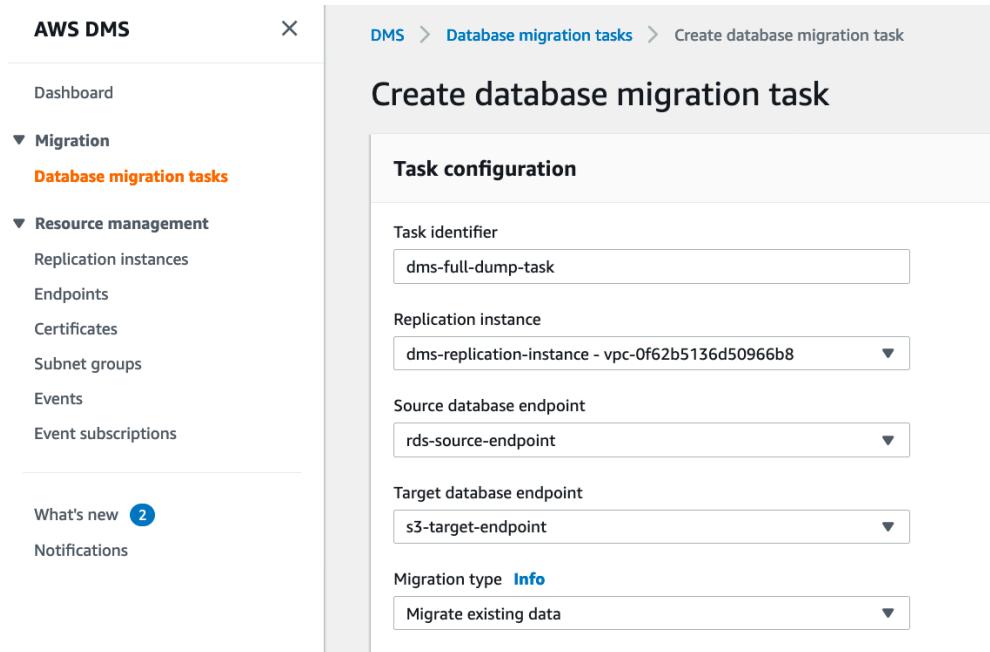
Name	Status	Source	Target	Type	Progress	Elapsed time	Tables loaded	Tables loading	Tables queued	Tables errored
Empty replication task table You don't have any replication tasks.										

2. Click **Create Task**.

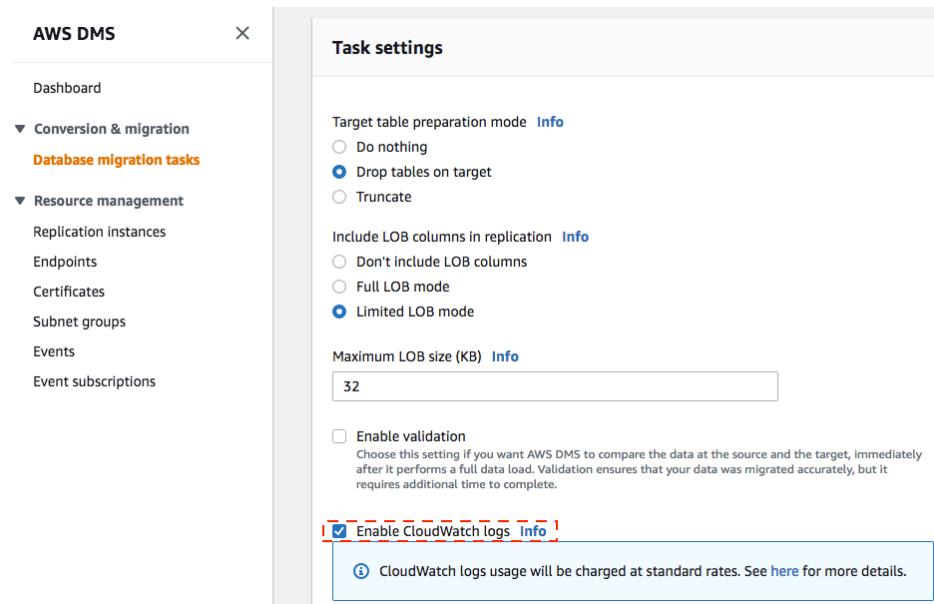
a. Type an easily recognized **Task name** e.g. **dms-full-dump-task**

## Lab 1. Hydrating the Data Lake with DMS

- b. Select your **Replication instance** from drop down.
- c. Select your **Source endpoint** from drop down.
- d. Select your **Target endpoint** from drop down.
- e. For, **Migration type** choose **Migrate existing data**.



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



- g. Go to **Table Mappings**.
- h. Click on **Add new selection rule** and select **Enter a Schema in Schema** field.

## Lab 1. Hydrating the Data Lake with DMS

- 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)

Find task

C Actions Create task

	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

## Lab 1. Hydrating the Data Lake with DMS

The screenshot shows the AWS DMS console with the following details:

- Task Summary:**
  - Status: Load complete
  - Type: Full load
  - Source: prodendpoint-postre
  - Target: targetendpoint
- Overview details:**
  - Task ARN: arn:aws:dms:us-east-1:1341259728059:task:MUYIRRLBYT4SEZESVNFNGAUL4
  - Type: Full load
  - Source: prodendpoint-postre
  - Last failure message: -
  - Started: 5/29/2019, 10:55:51 AM GMT-0700
  - Status: Load complete
  - Replication instance: dms-replication-instance
  - Target: targetendpoint
  - Created: 5/29/2019, 10:55:15 AM GMT-0700
  - Migration task logs: Info
  - View logs
- Table statistics (16):**

Schema name	Table	Load state	Inserts	Deletes	Updates	DDLs	Full load rows	Total	Validation state	Validation pending
dms_sample	seat_type	Table completed	0	0	0	0	6	6	Not enabled	0
dms_sample	seat	Table completed	0	0	0	0	603,631	603,631	Not enabled	0
dms_sample	mlb_data	Table completed	0	0	0	0	2,230	2,230	Not enabled	0
dms_sample	player	Table completed	0	0	0	0	5,157	5,157	Not enabled	0
dms_sample	ticket_purchase_hist	Table completed	0	0	0	0	6,038,756	6,038,756	Not enabled	0
dms_sample	person	Table completed	0	0	0	0	7,025,584	7,025,584	Not enabled	0
dms_sample	name_data	Table completed	0	0	0	0	5,360	5,360	Not enabled	0
dms_sample	sport_team	Table completed	0	0	0	0	62	62	Not enabled	0
dms_sample	sport_league	Table completed	0	0	0	0	2	2	Not enabled	0
dms_sample	sporting_event	Table completed	0	0	0	0	1,158	1,158	Not enabled	0
dms_sample	sporting_event_ticket	Table completed	0	0	0	0	15,212,460	15,212,460	Not enabled	0
dms_sample	sport_division	Table completed	0	0	0	0	14	14	Not enabled	0
dms_sample	sport_location	Table completed	0	0	0	0	62	62	Not enabled	0
dms_sample	sport_type	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	nfl_stadium_data	Table completed	0	0	0	0	32	32	Not enabled	0
dms_sample	nfl_data	Table completed	0	0	0	0	2,928	2,928	Not enabled	0

6. Open the S3 console to view the data that was copied by DMS:  
<https://s3.console.aws.amazon.com/s3/home?region=us-east-1#>
7. Click on the bucket used as the DMS target and navigate to **/tickets/dms\_sample/** to view the loaded tables, one folder per table

## Lab 1. Hydrating the Data Lake with DMS

The screenshot shows the Amazon S3 console interface. The path is: Amazon S3 > dmslab-student-dmslabs3bucket-wotl4bf73cw3 > tickets > dms\_sample. The 'Overview' tab is selected. A search bar at the top says 'Type a prefix and press Enter to search. Press ESC to clear.' Below it are buttons for 'Upload', '+ Create folder', 'Download', and 'Actions'. The main area lists files and folders: mlb\_data, name\_data, nfl\_data, nfl\_stadium\_data, person, player, seat, seat\_type, sport\_division, sport\_league, sport\_location, sport\_team, sporting\_event, sporting\_event\_ticket, and ticket\_purchase\_hist.

### 8. Download one of the files:

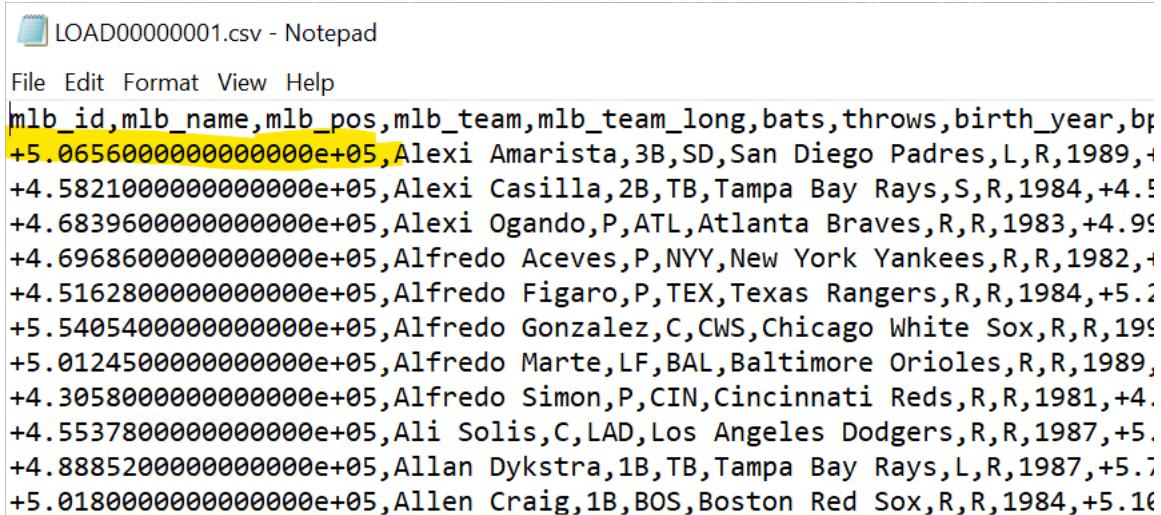
- Navigate further to **mlb\_data/LOAD00000001.csv**, select the check box next to the file name and click **Download** in the pop-up window.
- Click Save File.**
- Open the file.

You will notice that the file contains the column headers in the first row as requested by the “addColumnName=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	<b>id</b>	<b>sport_team_id</b>	<b>last_name</b>	<b>first_name</b>	<b>full_name</b>
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:

## Lab 1. Hydrating the Data Lake with DMS



LOAD00000001.csv - Notepad

File Edit Format View Help

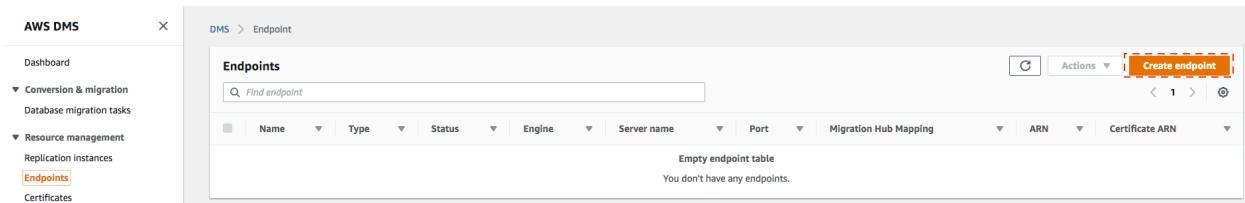
```
mlb_id,mlb_name,mlb_pos,mlb_team,mlb_team_long,bats,throws,birth_year,bp
+5.06560000000000e+05,Alexi Amarista,3B,SD,San Diego Padres,L,R,1989,+4.5
+4.58210000000000e+05,Alexi Casilla,2B,TB,Tampa Bay Rays,S,R,1984,+4.5
+4.68396000000000e+05,Alexi Ogando,P,ATL,Atlanta Braves,R,R,1983,+4.9
+4.69686000000000e+05,Alfredo Aceves,P,NYY,New York Yankees,R,R,1982,+4.5
+4.51628000000000e+05,Alfredo Figaro,P,TEX,Texas Rangers,R,R,1984,+5.2
+5.54054000000000e+05,Alfredo Gonzalez,C,CWS,Chicago White Sox,R,R,1985,+4.5
+5.01245000000000e+05,Alfredo Marte,LF,BAL,Baltimore Orioles,R,R,1989,+4.5
+4.30580000000000e+05,Alfredo Simon,P,CIN,Cincinnati Reds,R,R,1981,+4.5
+4.55378000000000e+05,Ali Solis,C,LAD,Los Angeles Dodgers,R,R,1987,+5.7
+4.88852000000000e+05,Allan Dykstra,1B,TB,Tampa Bay Rays,L,R,1987,+5.7
+5.01800000000000e+05,Allen Craig,1B,BOS,Boston Red Sox,R,R,1984,+5.1
```

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**:



AWS DMS

DMS > Endpoint

Endpoints

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**.

- 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**.

## Lab 1. Hydrating the Data Lake with 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.

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

**Service access role ARN**  
Role that can access target

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

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

- g. Click **Endpoint-specific settings** to expand the section.
- h. In the **Extra connection attributes** box, type **addColumnName=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
cdccendpoint	dms-replication-instance	successful	

**Create endpoint**

## Lab 1. Hydrating the Data Lake with DMS

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

Name	Type	Status	Engine	Server name	Port	Migration Hub Mapping	ARN
rds-cdc-endpoint	Target	Active	Amazon S3	-	-	-	arn:aws:dms:us-east-1:132701118127:endpoint:QQCRWAZQTXI
rds-source-endpoint	Source	Active	PostgreSQL	dmslabinstance.c8msbe8b7bxw.us-east-1.rds.amazonaws.com	5432	-	arn:aws:dms:us-east-1:132701118127:endpoint:SWCXG2HRIT7A
s3-target-endpoint	Target	Active	Amazon S3	-	-	-	arn:aws:dms:us-east-1:132701118127:endpoint:T533232IIX5I

## (Optional) Create a task to perform ongoing replication

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**.

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.										

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**.

Task configuration

Task identifier: cdctask

Replication instance: dms-replication-instance - vpc-0f62b5136d50966b8

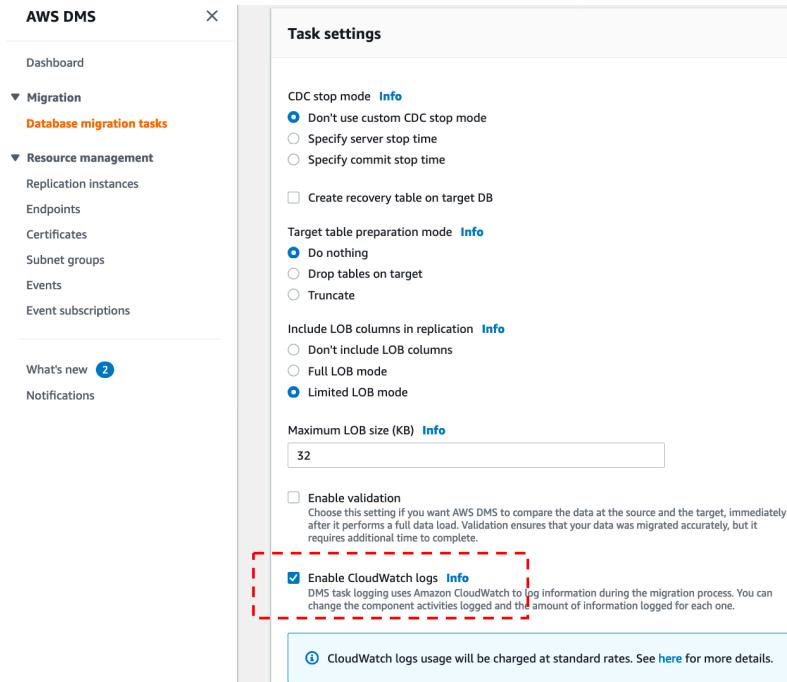
Source database endpoint: rds-source-endpoint

Target database endpoint: rds-cdc-endpoint

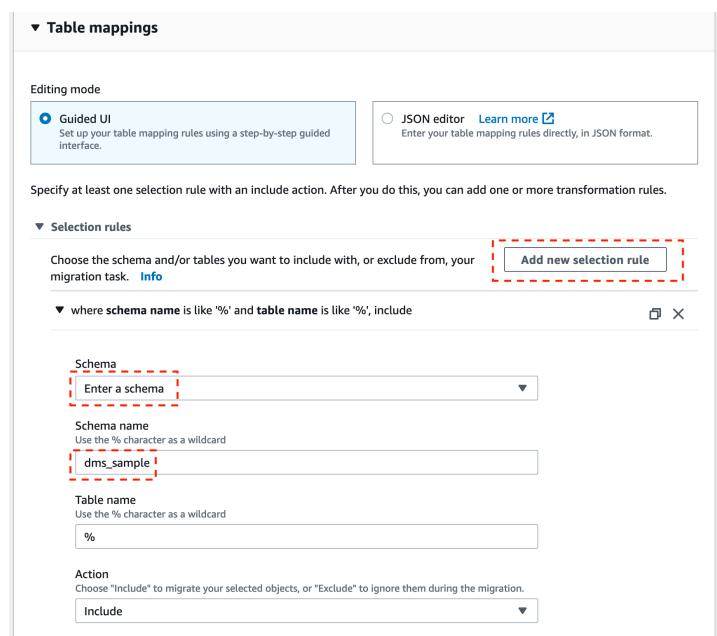
Migration type: Replicate data changes only

## Lab 1. Hydrating the Data Lake with DMS

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



- 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



## Lab 1. Hydrating the Data Lake with DMS

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.

Name	Status	Source	Target	Type	Progress	Elapsed time	Tables loaded	Tables loading	Tables queued
dms-task	Load complete	prodendpoint-postgre	targetendpoint	Full load	100 %	9 m	16	0	0
newcdc	Replication ongoing	prodendpoint-postgre	cdcdendpoint	Ongoing replication	100 %	0 m	16	0	0

4. 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.
5. 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

Schema name	Table	Load state	Inserts	Deletes	Updates	DDLs	Full load rows	Total	Validation state	Validation pending
dms_sample	seat_type	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	seat	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	mib_data	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	player	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	ticket_purchase_hist	Table completed	11,002	0	0	0	11,002	11,002	Not enabled	0
dms_sample	person	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	name_data	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_team	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_league	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sporting_event	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sporting_event_ticket	Table completed	0	0	11,002	0	11,002	11,002	Not enabled	0
dms_sample	sport_division	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_location	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_type	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	nfl_stadium_data	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	nfl_data	Table completed	0	0	0	0	0	0	Not enabled	0

6. Open the S3 console to view the data that was copied by DMS:  
<https://s3.console.aws.amazon.com/s3/home?region=us-east-1#>

## Lab 1. Hydrating the Data Lake with DMS

7. Click on the bucket used as the DMS target and navigate to **/cdc/dms\_sample/** to view the loaded tables, one folder per table

The screenshot shows the Amazon S3 console interface. The path is: Amazon S3 > dmslab-student-dmslabs3bucket-wotl4bf73cw3 > cdc > dms\_sample. The 'Overview' tab is selected. A search bar at the top says 'Type a prefix and press Enter to search. Press ESC to clear.' Below it are buttons for 'Upload', '+ Create folder', 'Download', and 'Actions'. A dropdown menu under 'Actions' is open. The main list shows two items: 'sporting\_event\_ticket' and 'ticket\_purchase\_hist', each with a checkbox next to its name.

8. Download one of the files:

- Select the check box next to the object name and click Download in the pop-up window.
- Click **Save File**.
- Open the file.

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

The screenshot shows the Amazon S3 console interface. The path is: Amazon S3 > dmslab-student-dmslabs3bucket-wotl4bf73cw3 > cdc > dms\_sample > sporting\_event\_ticket. The 'Overview' tab is selected. A search bar at the top says 'Type a prefix and press Enter to search. Press ESC to clear.' Below it are buttons for 'Upload', '+ Create folder', 'Download', and 'Actions'. A dropdown menu under 'Actions' is open. The main list shows two files: '20190529-230604667.csv' and '20190529-230729693.csv'. The first file has a checked checkbox. A modal window is open over the list, titled '20190529-230604667.csv'. It contains three buttons: 'Download', 'Copy path', and 'Select from'. The modal is divided into sections: 'Latest version' and 'Overview'. The 'Overview' section displays the following details:  
Key: 20190529-230604667.csv  
Size: 19.3 MB  
Expiration date: N/A  
Expiration rule: N/A  
ETag: acfb57f453c0449745035e6d50dfb3bc-4  
Last modified: May 29, 2019 4:06:05 PM GMT-0700  
Object URL: https://s3.amazonaws.com/dmslab-student-dmslabs3bucket-wotl4bf73cw3/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.

### Lab 1. Hydrating the Data Lake with DMS

A	B	C	D	E	F	G	H	I	J
Op	id	sporting_eve	sport_location_id	seat_level	seat_section	seat_row	seat	ticketholder_id	ticket_price
U	145192591	3931	4	2	10 A		2	2898028	98
U	145192601	3931	4	2	10 A		1	2898028	98
U	145192581	3931	4	2	10 A		3	2898028	98
U	145192501	3931	4	2	10 B		1	2898028	98
U	145187751	3931	4	2	13 B		2	2898028	49
U	145187741	3931	4	2	13 B		3	2898028	49
U	145187721	3931	4	2	13 C		2	2898028	49
U	145187711	3931	4	2	13 C		3	2898028	49
U	145187731	3931	4	2	13 C		1	2898028	49
U	145187701	3931	4	2	14 A		1	2898028	49
U	145187681	3931	4	2	14 A		3	2898028	49
U	145187691	3931	4	2	14 A		2	2898028	49
U	145187471	3931	4	2	14 B		3	2898028	49
U	145187671	3931	4	2	14 B		1	2898028	49
U	145187481	3931	4	2	14 B		2	2898028	49
U	145187451	3931	4	2	14 C		2	2898028	49
U	145187461	3931	4	2	14 C		1	2898028	49
U	145190341	3931	4	2	14 C		3	2898028	49
U	145183201	3931	4	2	15 A		4	2898028	49
U	145179691	3931	4	2	15 A		1	2898028	49
U	145179661	3931	4	2	15 A		4	2898028	49
U	145179671	3931	4	2	15 A		3	2898028	49
U	145179681	3931	4	2	15 A		2	2898028	49
U	145190321	3931	4	2	15 A		2	2898028	49