Amazon Web Services
Data Engineering Immersion Day

Database Migration Services Lab

September 2021

Table of Contents

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.

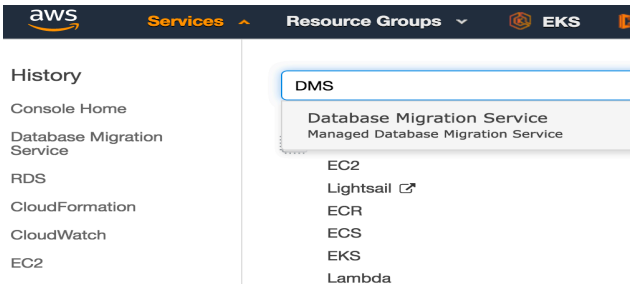
Optionally, you can add ongoing replication of data changes on the source: ***(Only one of the DMS replication instances will enable this feature.)***

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

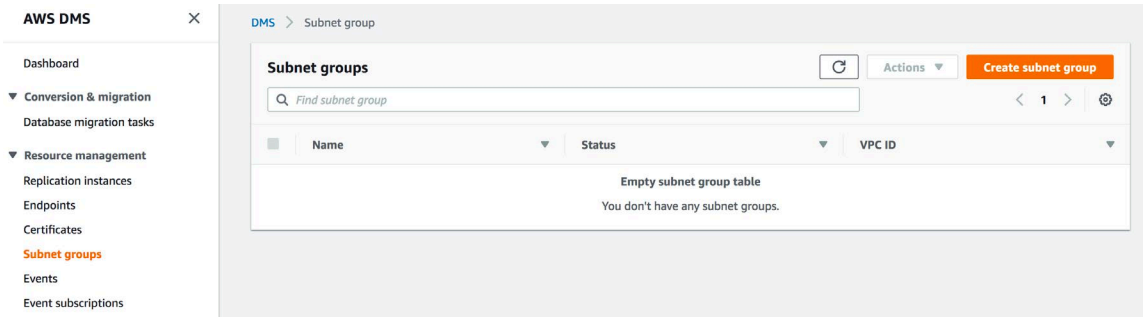
Your instructor has already created and populated the RDS Postgres database that you will use as your source endpoint in this lab.

Create the Subnet Group

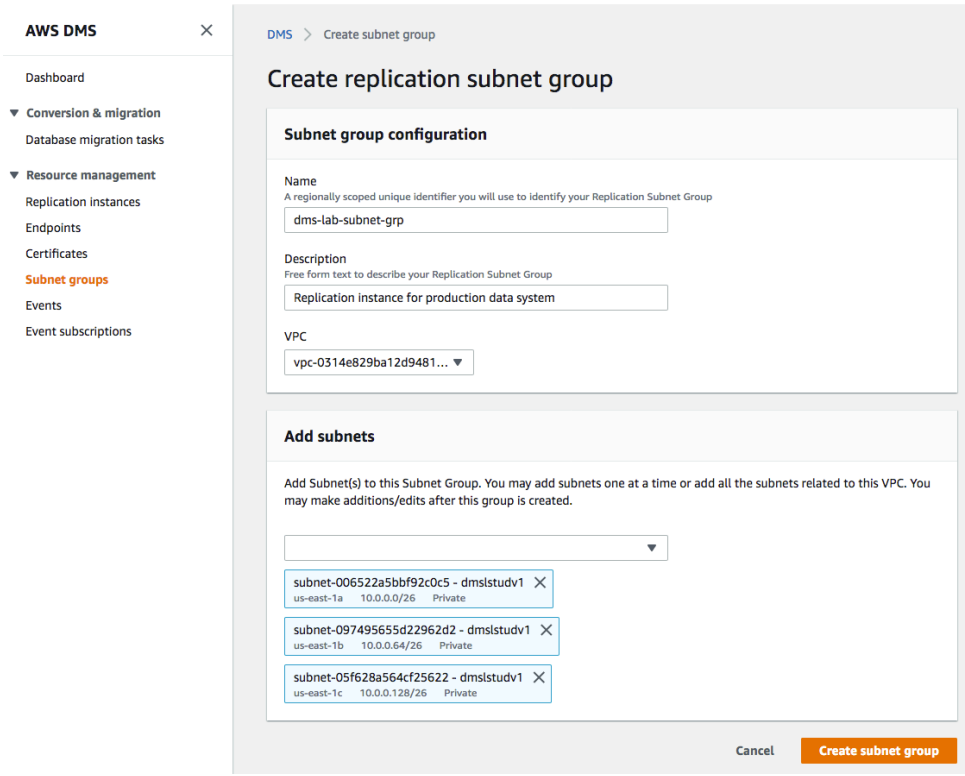
1. Search **DMS service** in the AWS console



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

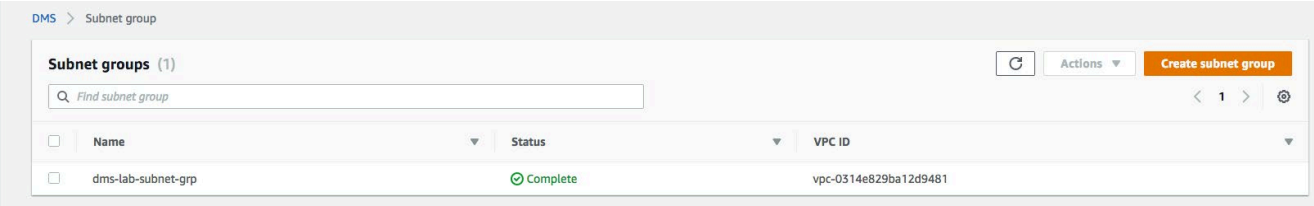


3. Click **Create subnet group**.
4. In the Identifier box, type a descriptive name that you will easily recognize (e.g., dms-lab-subnet-grp).
5. In the Description box, type an easily recognizable description (e.g., Replication instance for production data system).
6. For VPC, select the name of the VPC that you created earlier (e.g., dmslstudv1). The subnet list populates in the Available Subnets pane.
7. 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.



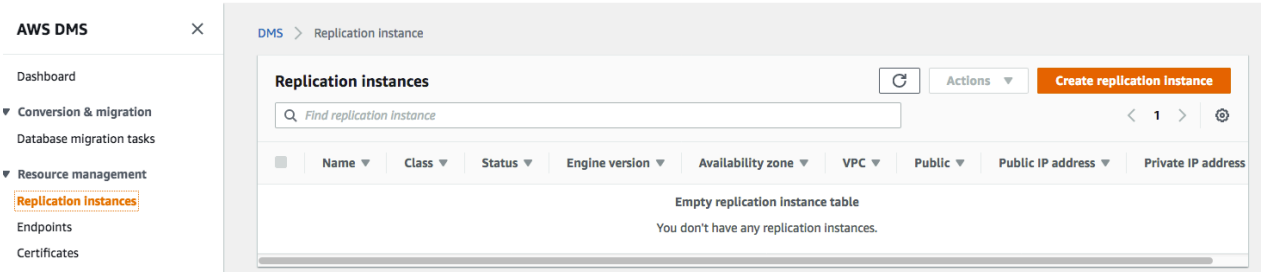
8. Click Create subnet group

9. On the DMS console, the subnet group status displays Complete.



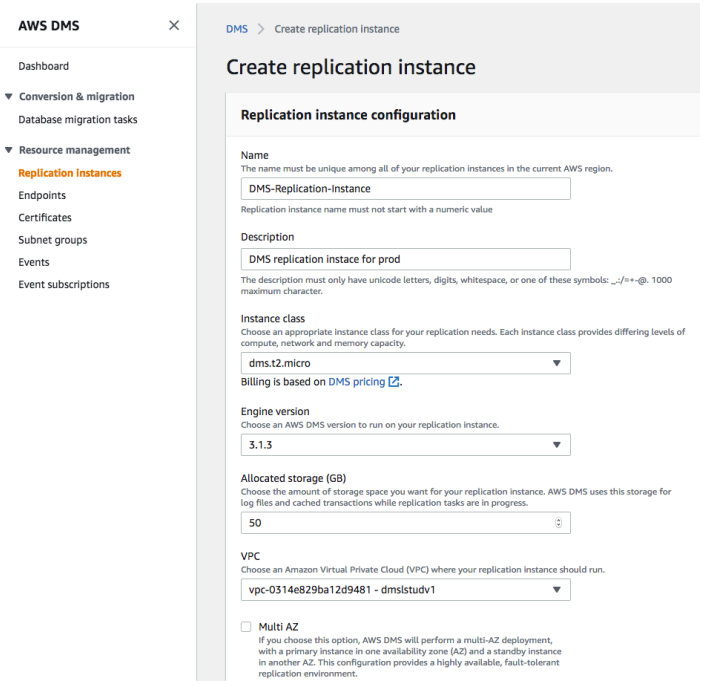
Create the Replication Instance

1. On the DMS console, select **Replication instances**.
2. Click **Create replication instance**.

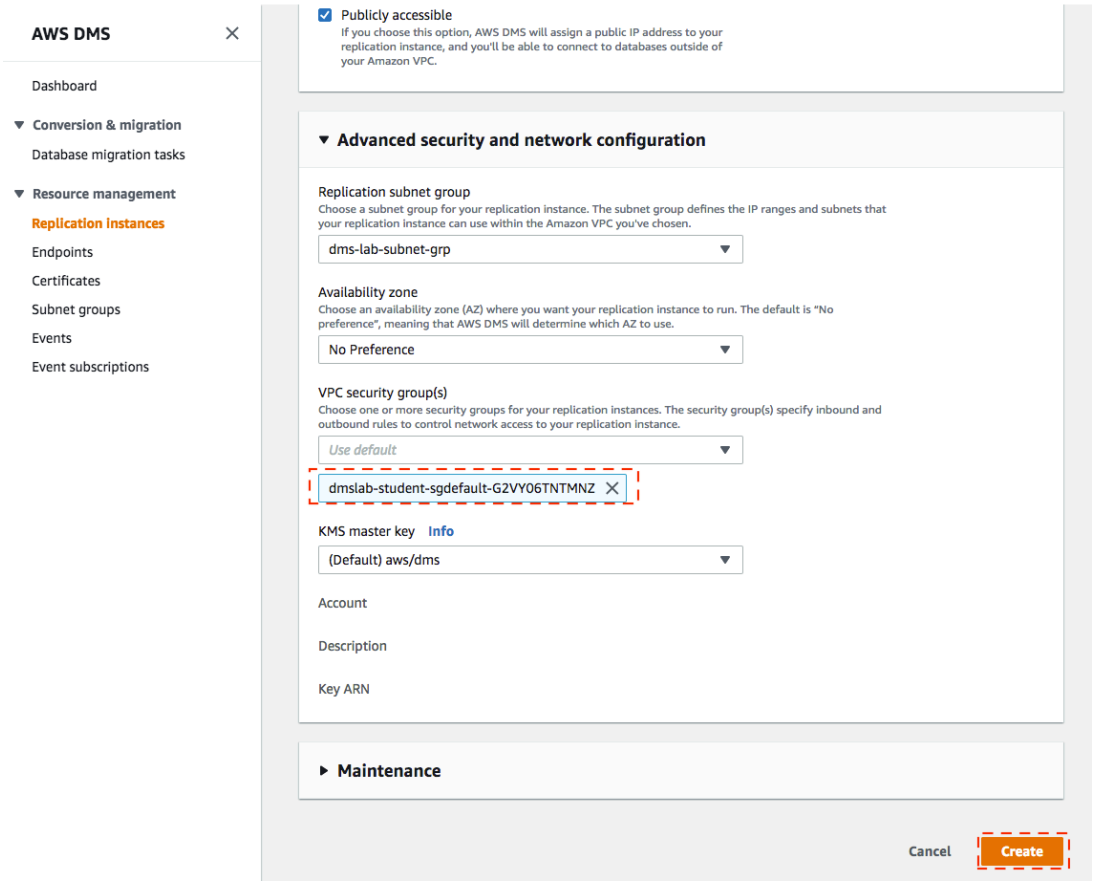


3. For Name, type a name for the replication instance that you will easily recognize.
4. For Description, type a description you will easily recognize. (e.g., DMS-Replication-Instance).
5. For Instance class, choose **dms.t2.micro**.
6. For VPC, choose the **dmslsstudv1** that you created earlier in pre-lab.

   NOTE: Keep the existing default settings. (You may see a newer engine version than what is shown in the example image.)

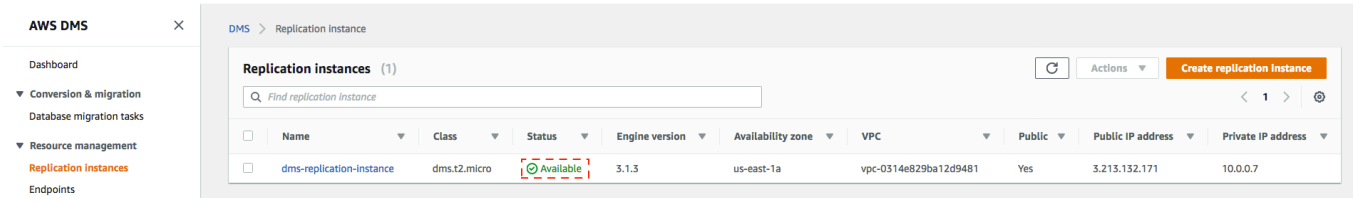


7. Click **Advanced** to expand the section.
8. Select the security group with **sgdefault** in the name.



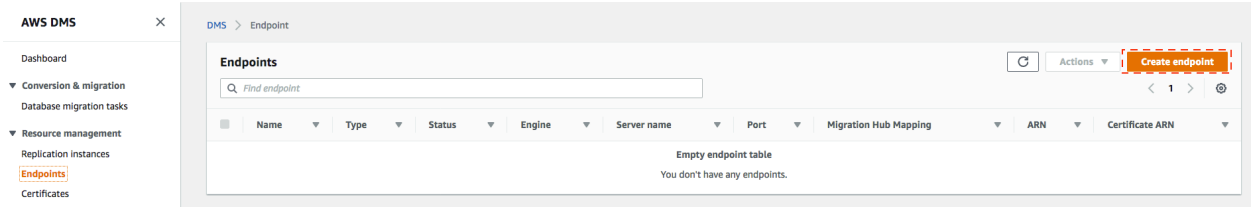
9. Click **Create**.

10. The DMS console displays **creating** for the instance status. When the replication instance is ready, the status changes to **available**.

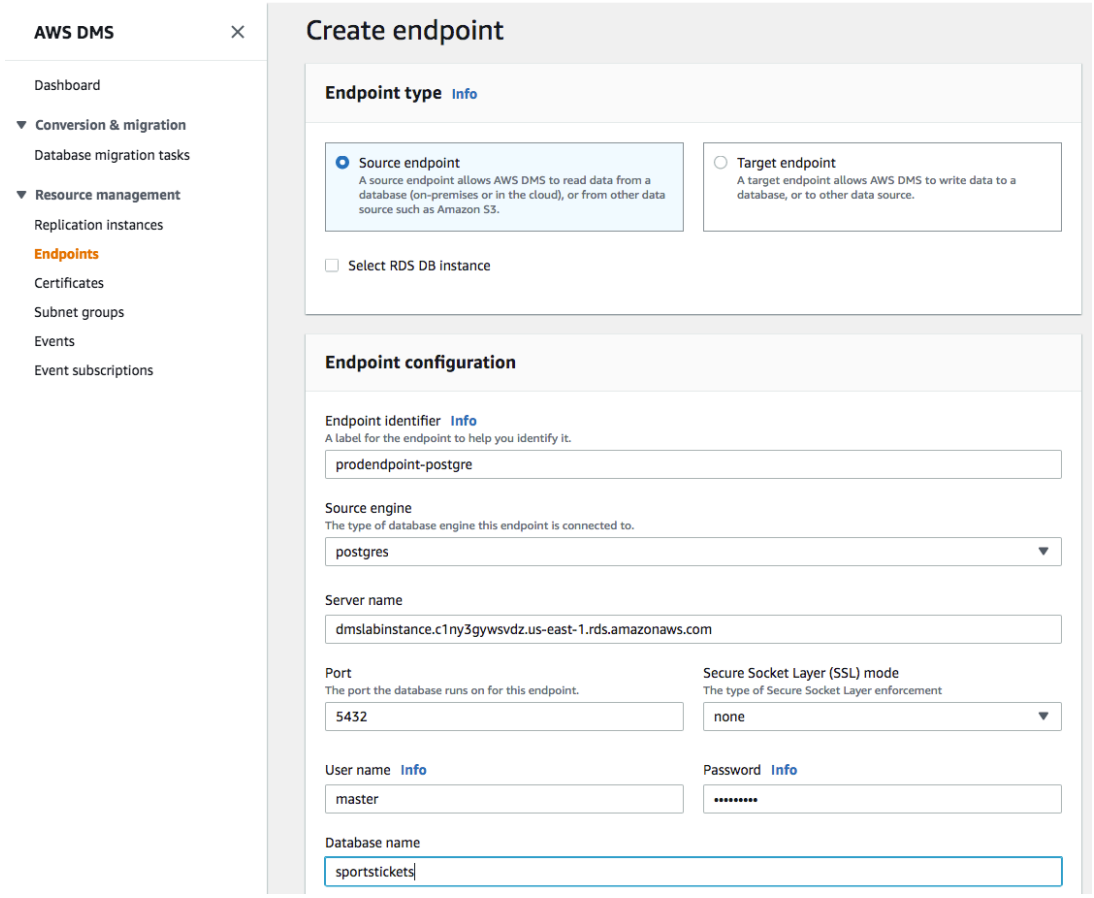


Create the Source Endpoint

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

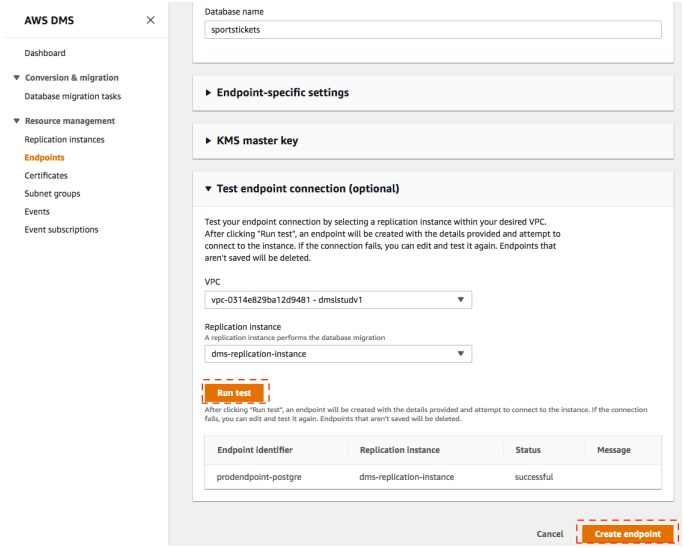


2. Click **Create endpoint**.
3. On the Create endpoint page, for Endpoint type, select **Source**.
4. For Endpoint identifier, select your easily recognized name.
5. For Source engine, select **postgres**.
6. Enter the **Server name** provided by your instructor, or if you ran instructor lab then take recorded endpoint from the same pre-lab.
7. For Port, enter **5432**.
8. For SSL mode, choose **none**.
9. For User name, type **adminuser**.
10. For Password, type **admin123**.
11. For Database name, type **sportstickets**.



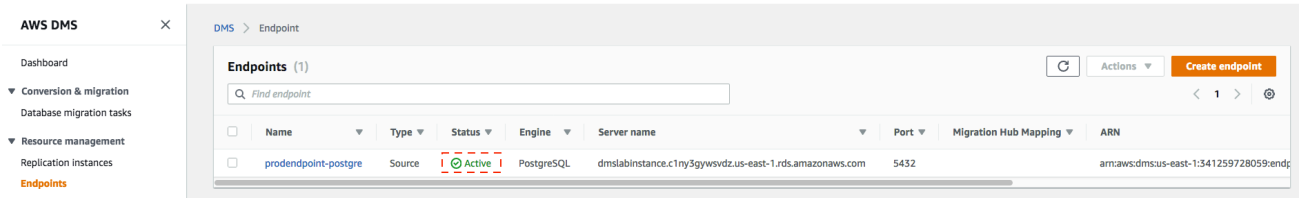
12. Expand the **Test endpoint connection (optional) section**, and choose your DMS Lab VPC name on the VPC drop-down list.

13. Click **Run test**. This step tests connectivity to the source database system. If successful, the message “Connection tested successfully” appears.



14. Click **Create endpoint** to create the endpoint.

15. When available, the endpoint status changes to **active**.



IAM Policy for DMS->S3 Access

Now that we have created the source endpoint from which we want to replicate and/or export data from, we now need a security policy and role that DMS can run under to store the results against our target.

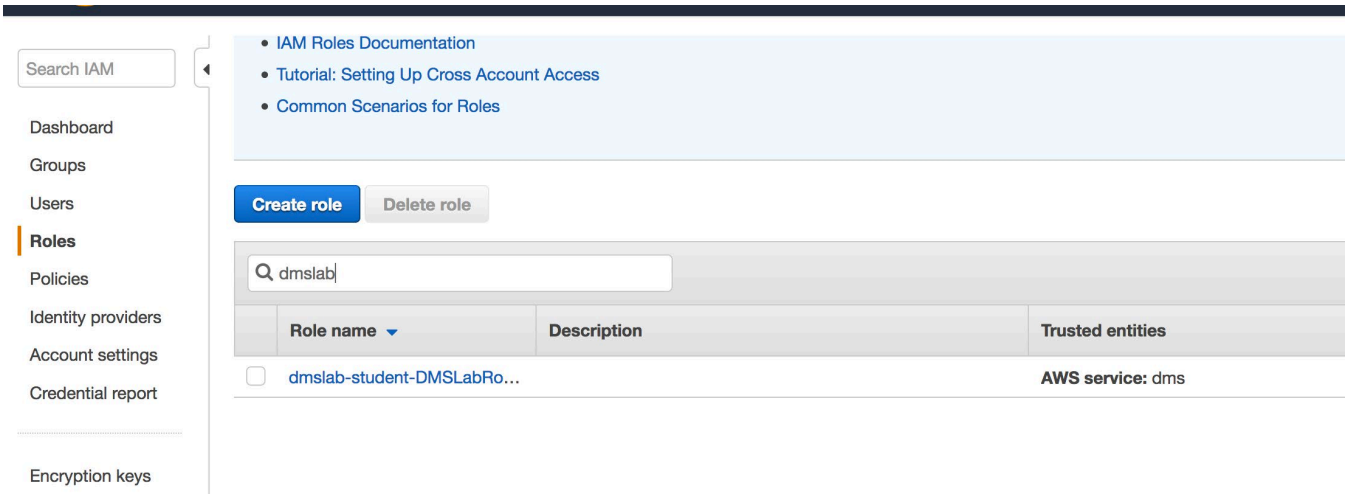
The policy and role have been created for you in the student prelab by an AWS CloudFormation template, with a permission set that allows the DMS service to access the S3 bucket.

Below IAM policy for the IAM role granted to your S3 bucket endpoint, enabling DMS to write to the S3 bucket. This policy grants *GetObject, PutObject, DeleteObject and ListBucket* to a bucket with a name that starts with dmslab. See the following code for an example:

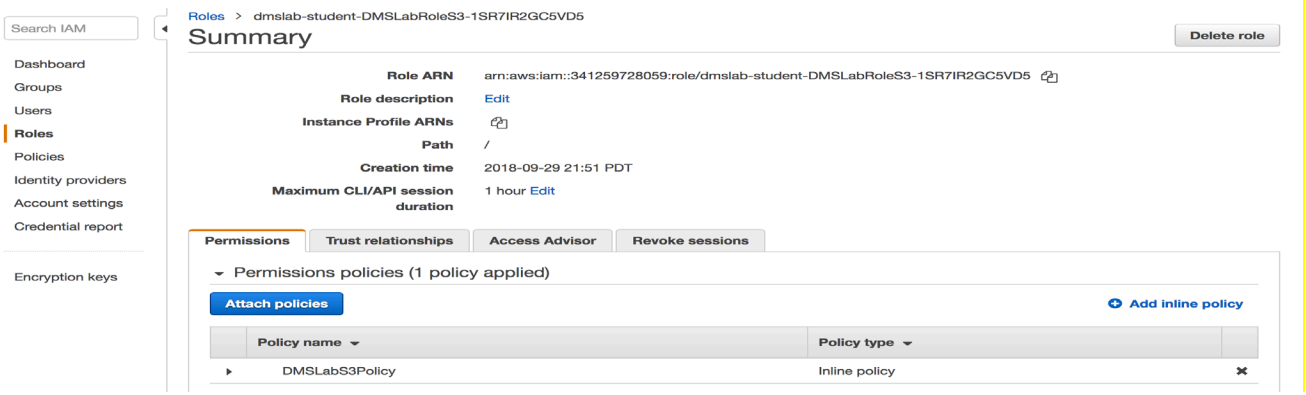
```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Action": [
            "s3:GetObject",
            "s3:PutObject",
            "s3:DeleteObject"
         ],
         "Resource": [
            "arn:aws:s3:::dmslabstudent-dmslabs3bucket-1eegnc2tj056l/*”
         ],
         "Effect": "Allow"
      },
      {
         "Action": [
            "s3:ListBucket"
         ],
         "Resource": [
            " arn:aws:s3:::dmslabstudent-dmslabs3bucket-1eegnc2tj056l”
         ],
         "Effect": "Allow"
      }
   ]
}
```

Explore the IAM Role by following below steps:

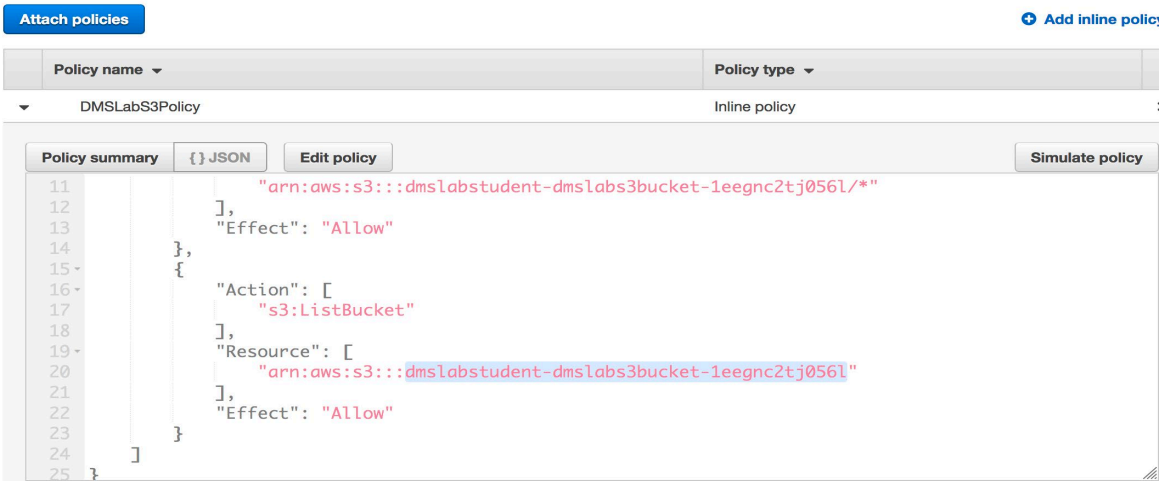
1. On the IAM console, select Roles.
2. On the Roles page, in the search box, type **dmslab** to filter the results.
3. Click the **DMSLabRoleS3** role name.



4. Look at the **ROLE ARN** value for this role.



5. Expand the *DMSLabS3Policy*, record the **s3 bucket name** from the Resource section. NOTE: don't copy the ARN prefix '*arn:aws:s3:::*'

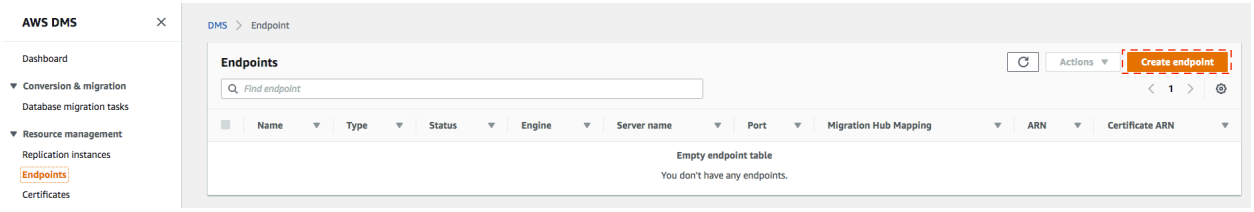


Create the Target Endpoint

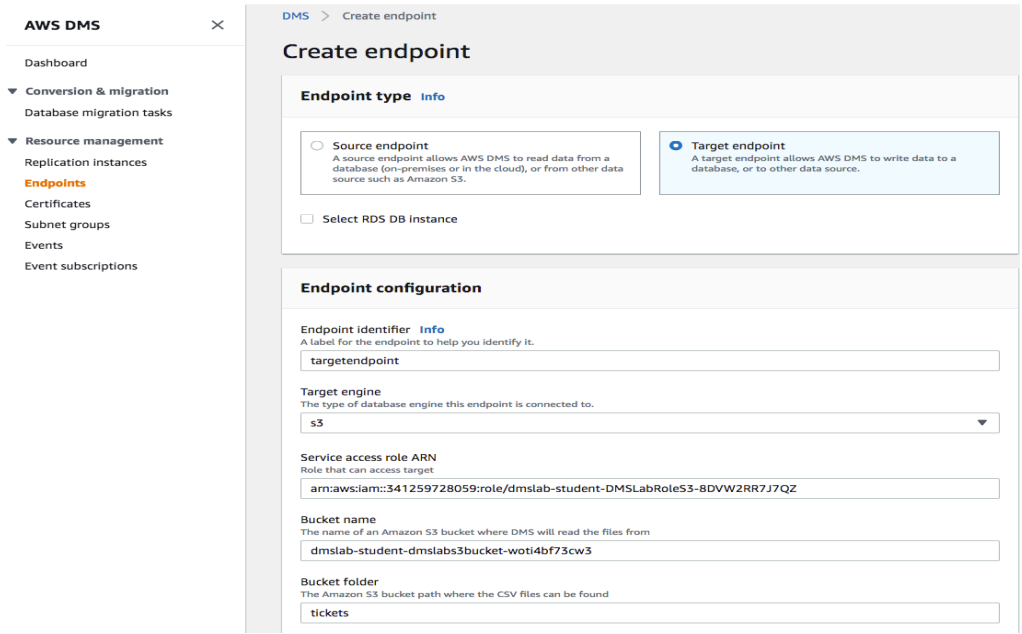
Before start, make sure you have the following information on hand:

- **DMSLabRoleS3 ARN** - arn:aws:iam::xxxx:role/xxxxx
- **s3 Bucket Name** - xxxx-dmslabs3bucket-xxxxx

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

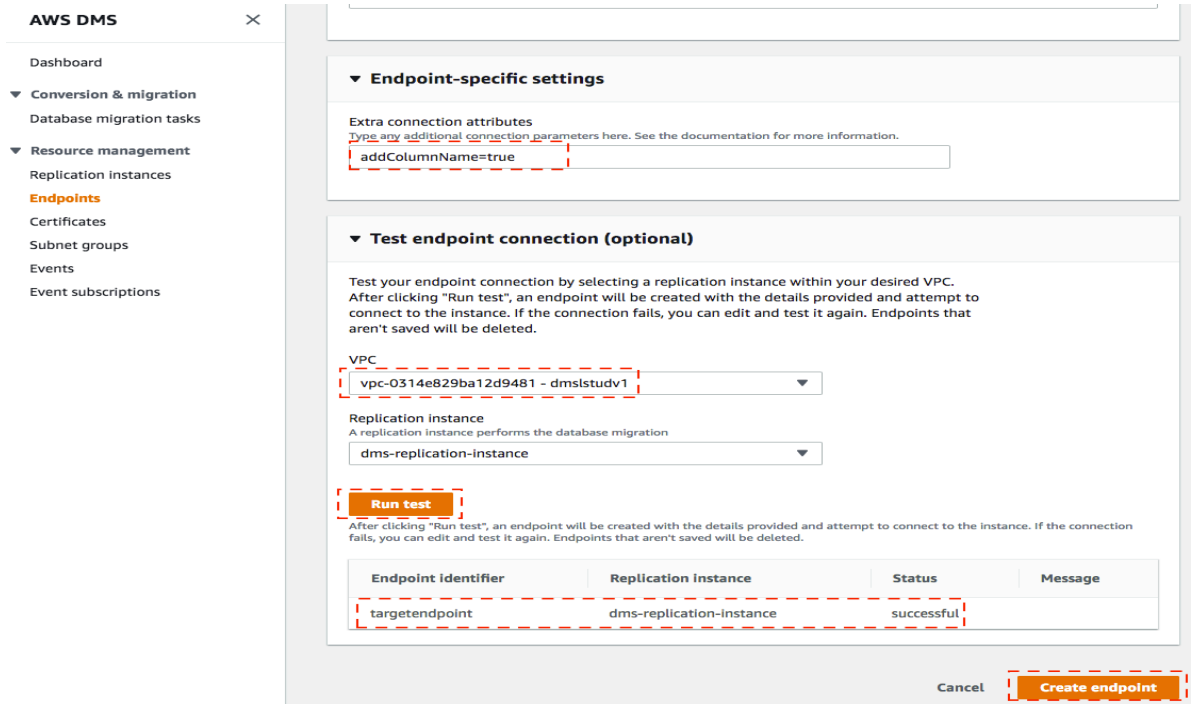


2. Click **Create endpoint**.
3. For Endpoint type, select **Target**.
4. For Endpoint identifier, type an easily recognized name.
5. For Target engine, choose **s3**.
6. For Service access role ARN, paste the **DMSLabRoleS3 ARN** number
7. For Bucket name, paste the **s3 bucket name**
8. For Bucket folder, type **tickets**.

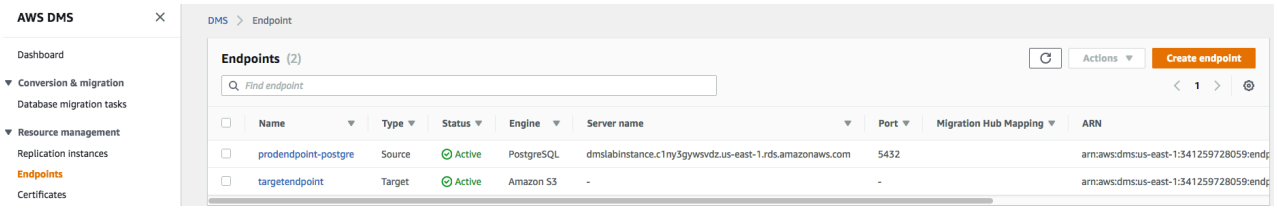


9. Click **Endpoint-specific settings** to expand the section.
10. In the **Extra connection attributes** box, type **addColumnName=true**. This attribute includes the column names in the files in the S3 bucket.

11. Expand the **Test endpoint connection (optional) section**, and choose your **dmslstudv1** name on the **VPC** drop-down list.
12. Click **Run test**. This step tests connectivity to the source database system. If successful, the message “Connection tested successfully” appears.
13. Click **Create Endpoint**. If the button is grey out, just click **Cancel** button.

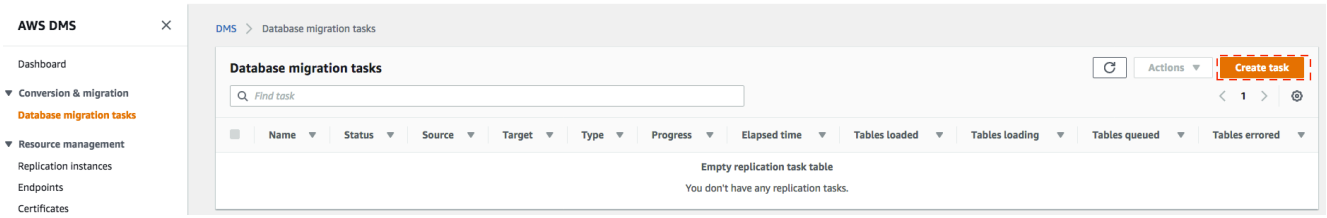


When available, the endpoint status changes to **active**.



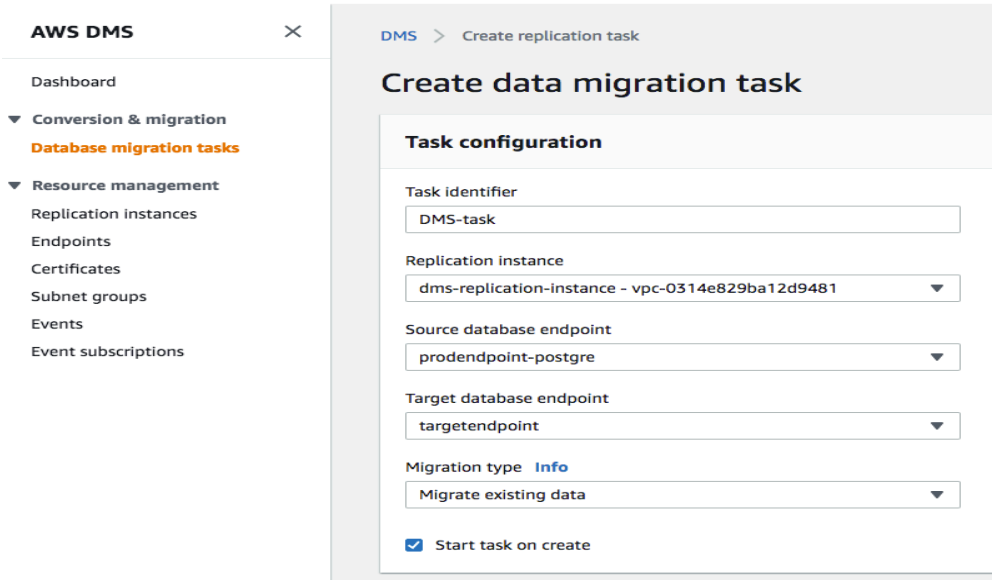
Create a task to perform the initial full copy

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

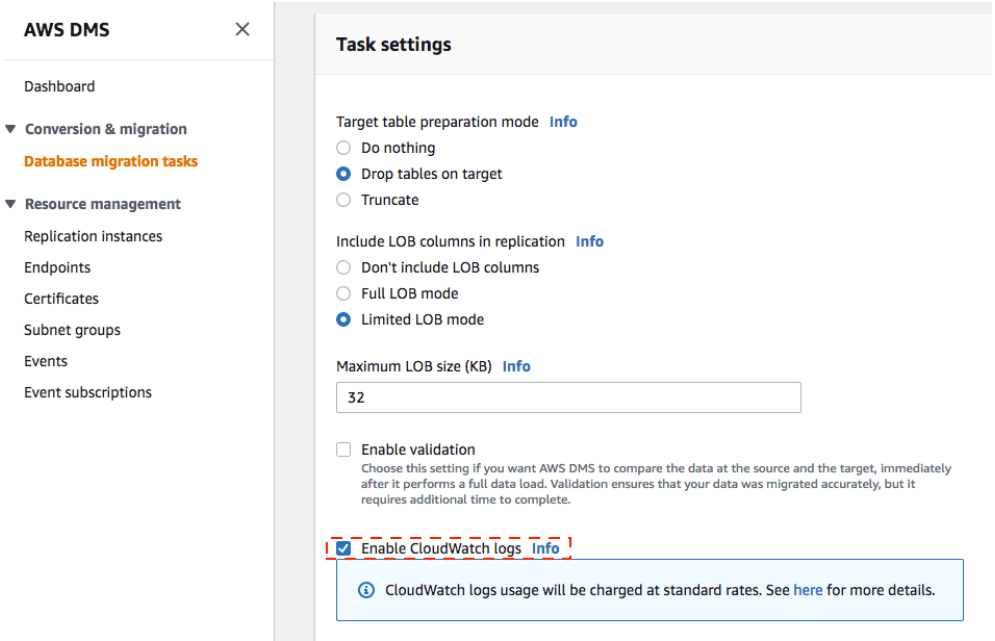


2. Click **Create Task**.

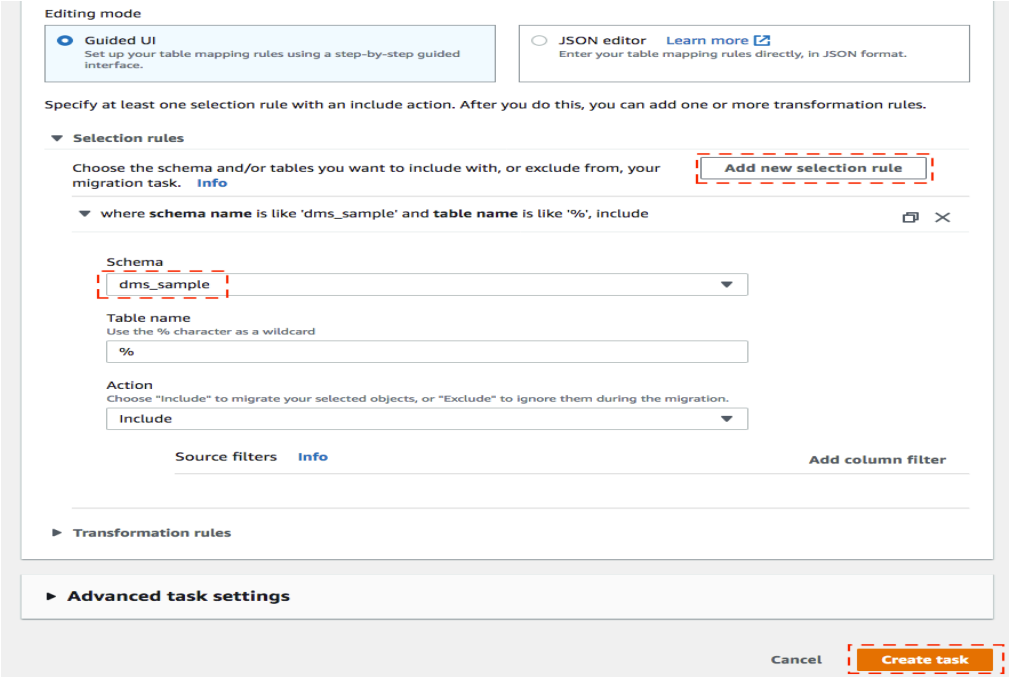
3. Type an easily recognized **Task name**.
4. Select your **Replication instance**.
5. Select your **Source endpoint**.
6. Select your **Target endpoint**.
7. For Migration type, choose **Migrate existing data**.
8. Select the **Start task on create** check box.



9. Expand **Task Settings.**
10. Select the **Enable CloudWatch logs** check box.

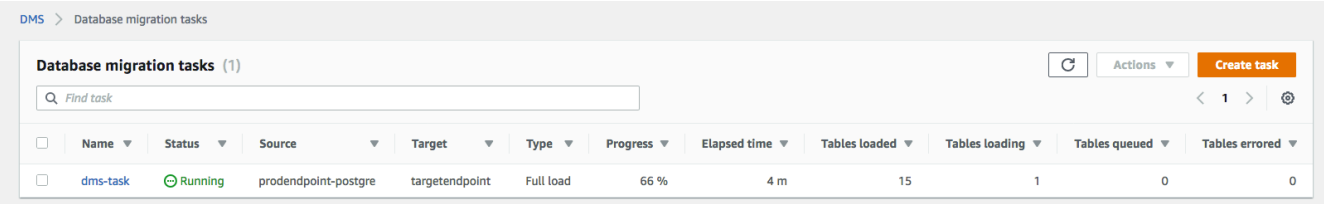


11. Go to **Table Mappings**.
12. Click on **Add new selection rule**
13. For Schema name, select **dms_sample** from drop down. Keep the settings for the remaining fields

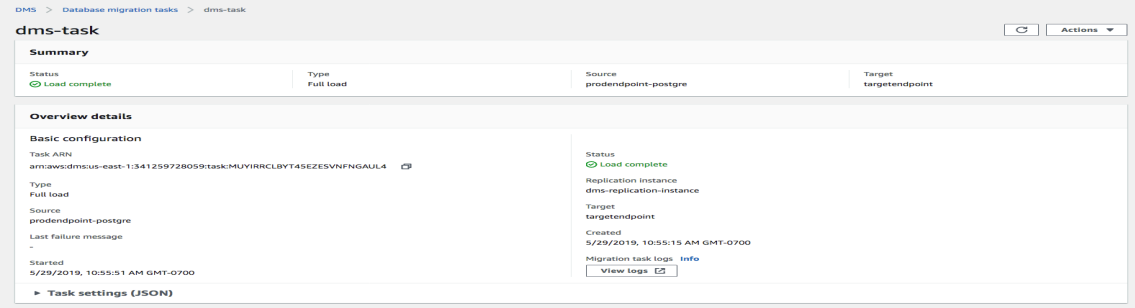


14. Click **Create task**.

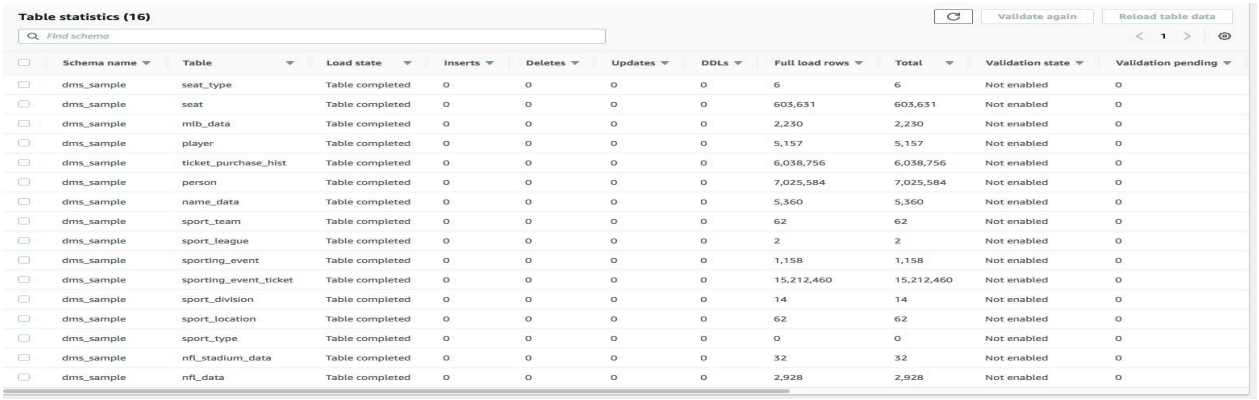
Your task is created and starts automatically. (Note: The complete creation and data extraction process takes 5 to 15 minutes.)



Once complete, the console displays 100% complete. Select your task and explore the summary:

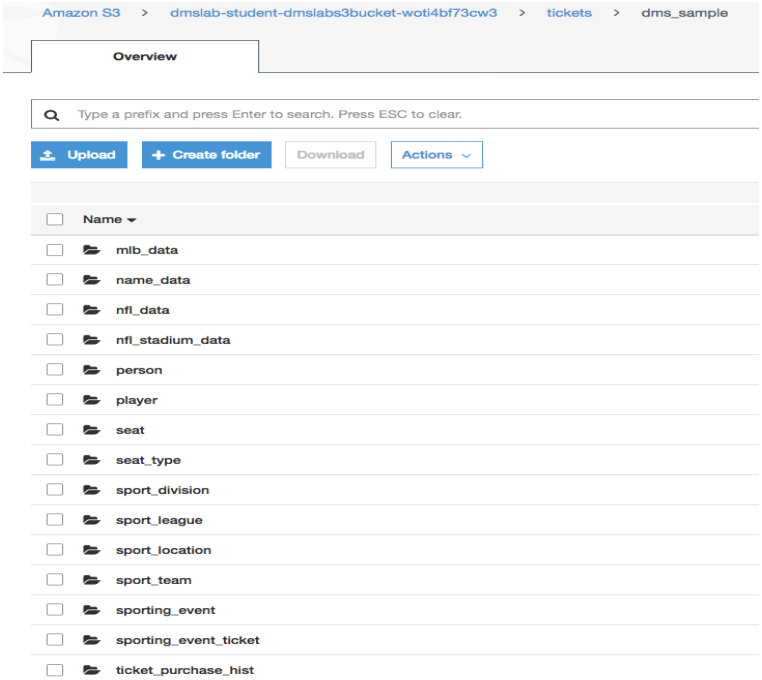


Scroll down and you can observe all table information loaded in S3 from RDS by DMS

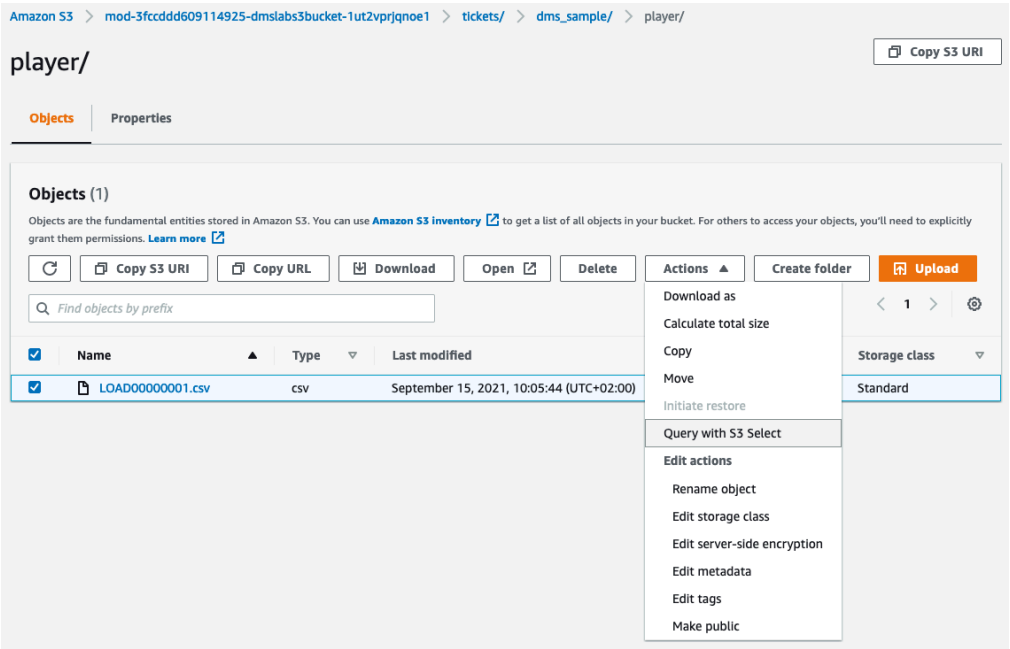


15. Open the S3 console and view the data that was copied by DMS.

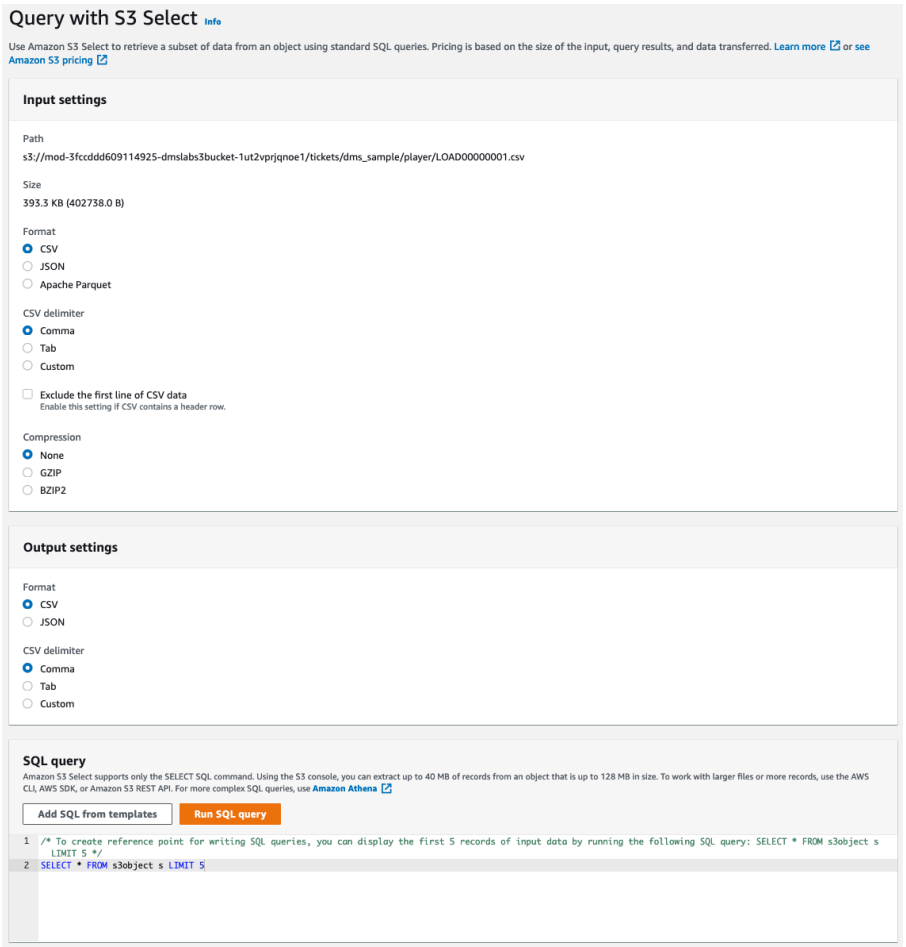
    Your S3 bucket name will look like below :
    BucketName/bucket_folder_name/schema_name/table_name/objects/



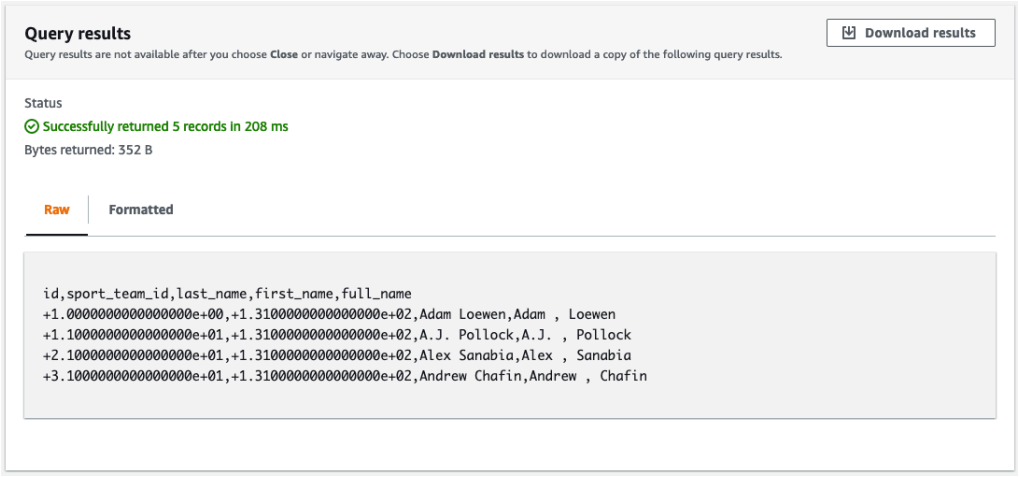
16. Navigate to one of the files and review it using [S3 Select]:
    a. Navigate in to the directory named **player** and select the check box next to the file name.
    b. Click the **Actions** dropdown button and choose **Query with S3 Select**.



    c. In the Query with S3 Select page, leave the default value for *Input Settings* and SQL Query and click **Run SQL query**.



d. It will execute the specified SQL query and return the first 5 lines from the CSV file.



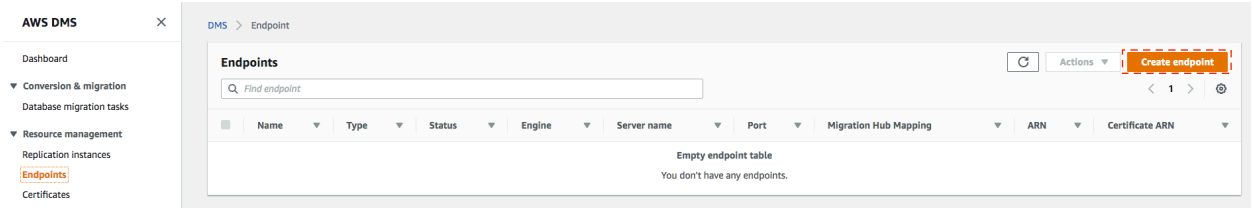
You will notice that the file contains the column headers in the first row as requested by the **“addColumnName=true”** connection attribute we included when we created the s3 target endpoint. Note that column names are included in the file in the first row.

Explore the objects in the S3 directory further.

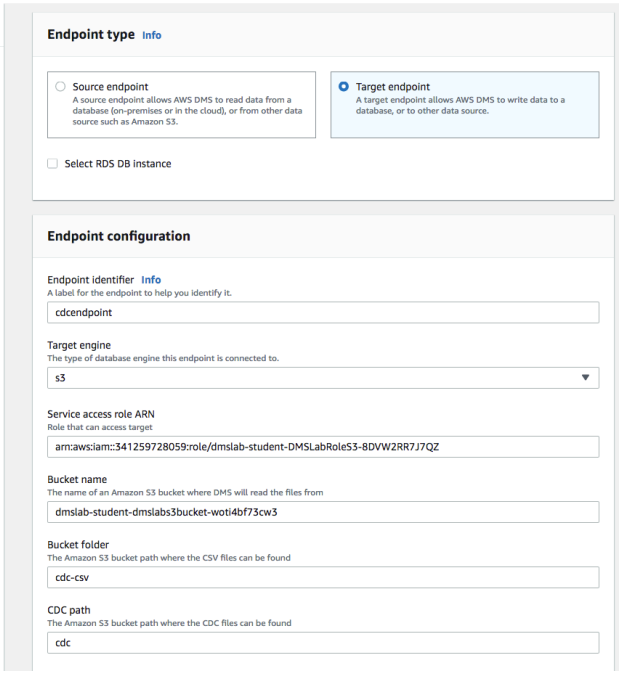
(Optional) Create the CDC endpoint to replicate ongoing changes

As of now we are enabling only one schema replication for CDC

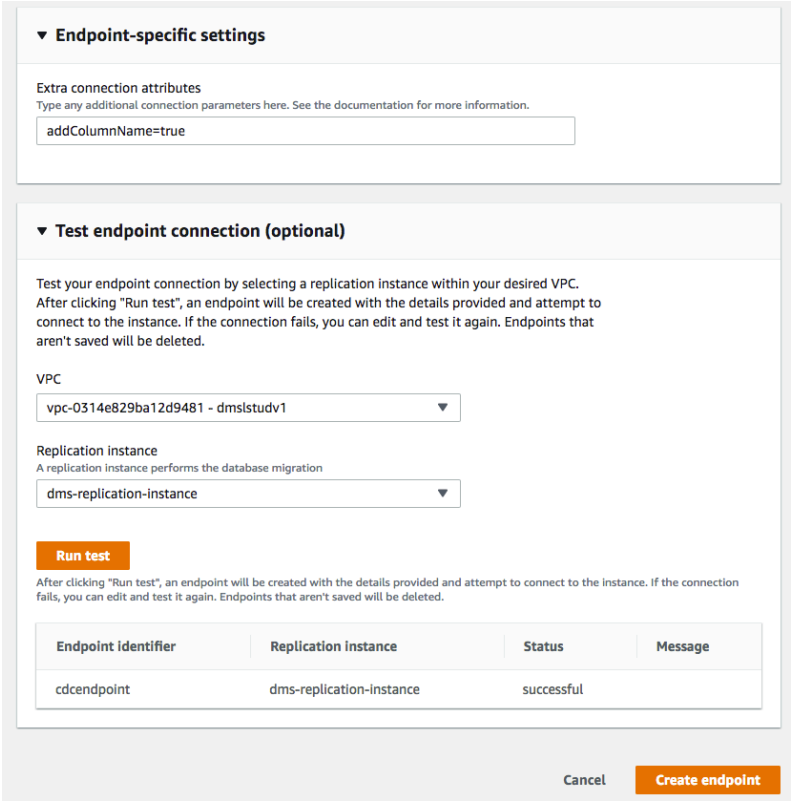
1. On the DMS console, select **Endpoints**.



2. Click **Create endpoint**.
3. For Endpoint type, select **Target**.
4. For Endpoint identifier, type an easily recognized name that includes “cdc”.
5. For Target engine, choose **Amazon S3**.
6. For Service Access Role ARN, paste the ARN value that you copied in the IAM role console group.
   NOTE: The value is similar to the following string, where the account number is specific to your account number: “arn:aws:iam::119911911299:role/data-eng-dms-role”
7. For Bucket name, type the name of the s3 bucket you noted down from pre-lab.
8. For Bucket folder, type **cdc** and For CDC path , leave blank

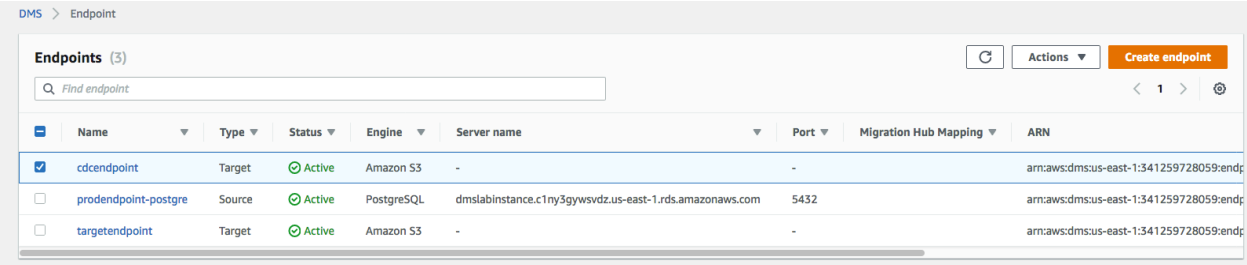


9. Click **Endpoint-specific settings** to expand the section.
10. In the **Extra connection attributes** box, type **addColumnName=true**. This attribute includes the column names in the files in the S3 bucket.
11. Expand the **Test endpoint connection (optional)** section, and choose your **dmslstudv1**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



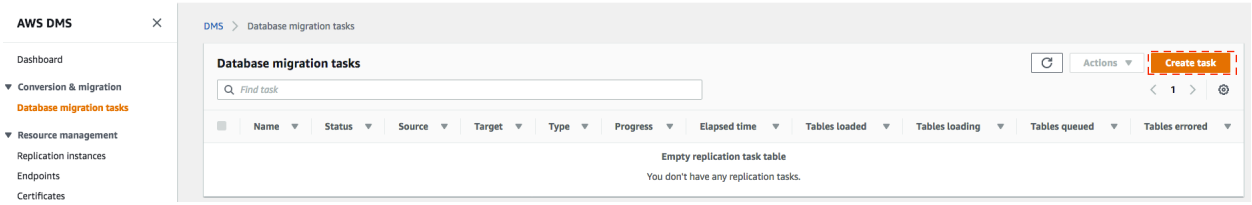
13. Click **Create endpoint**.

14. When available, the endpoint status changes to active.

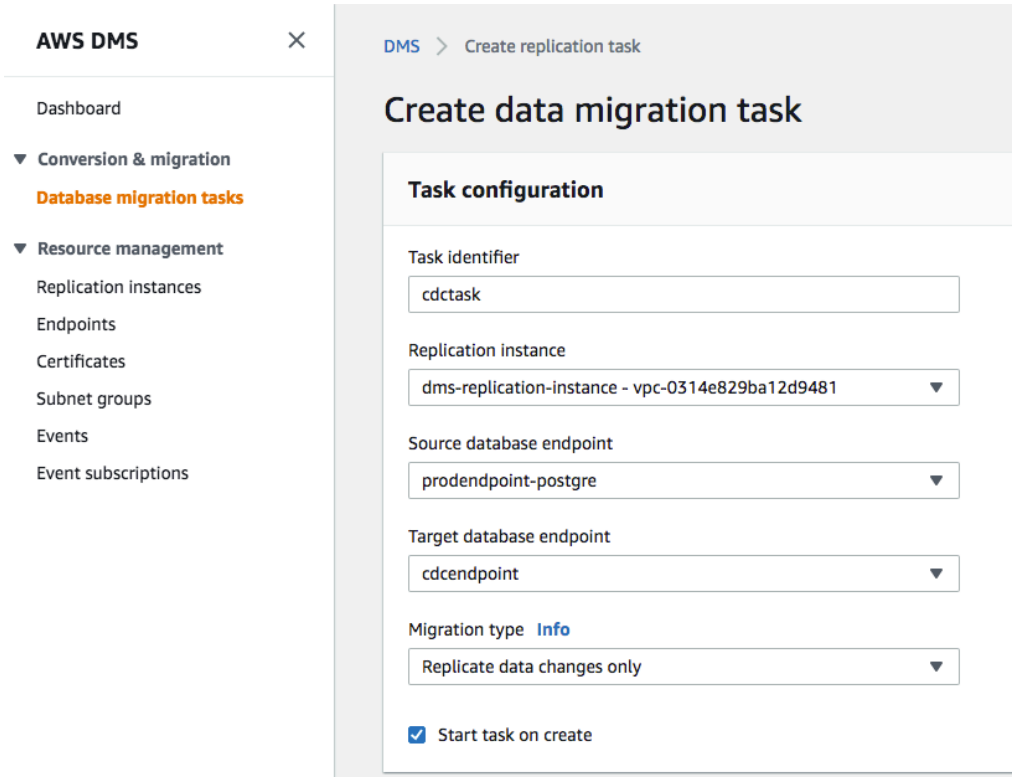


(Optional) Create a task to perform the ongoing replication

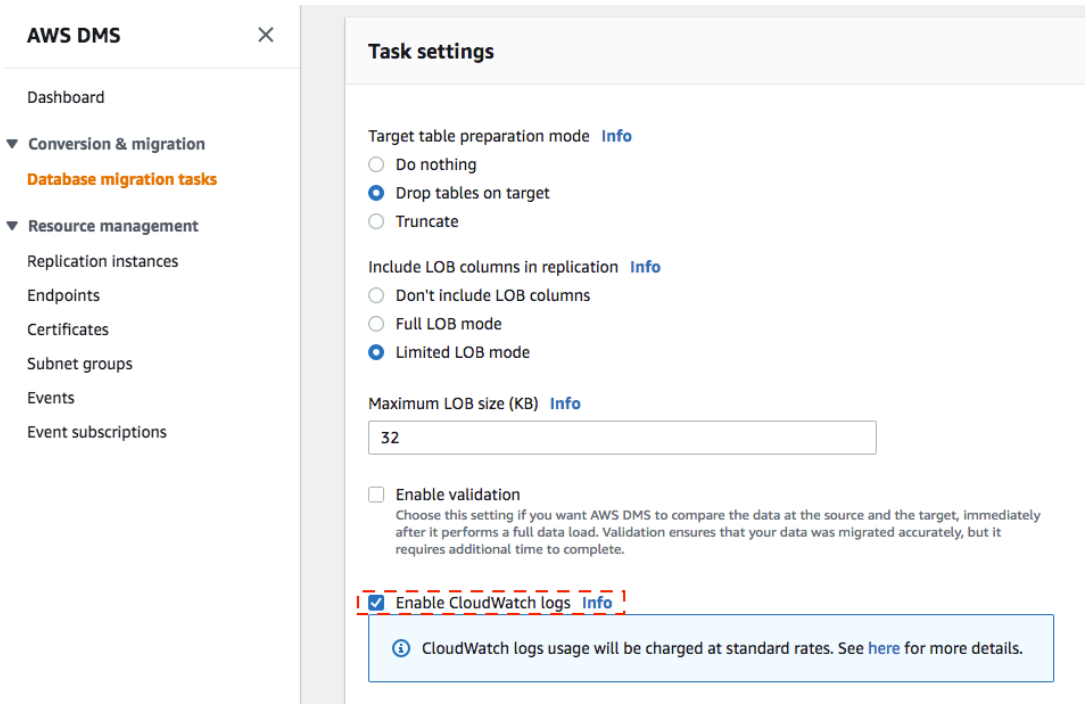
1. On the DMS console, select **Database Migration Tasks**.



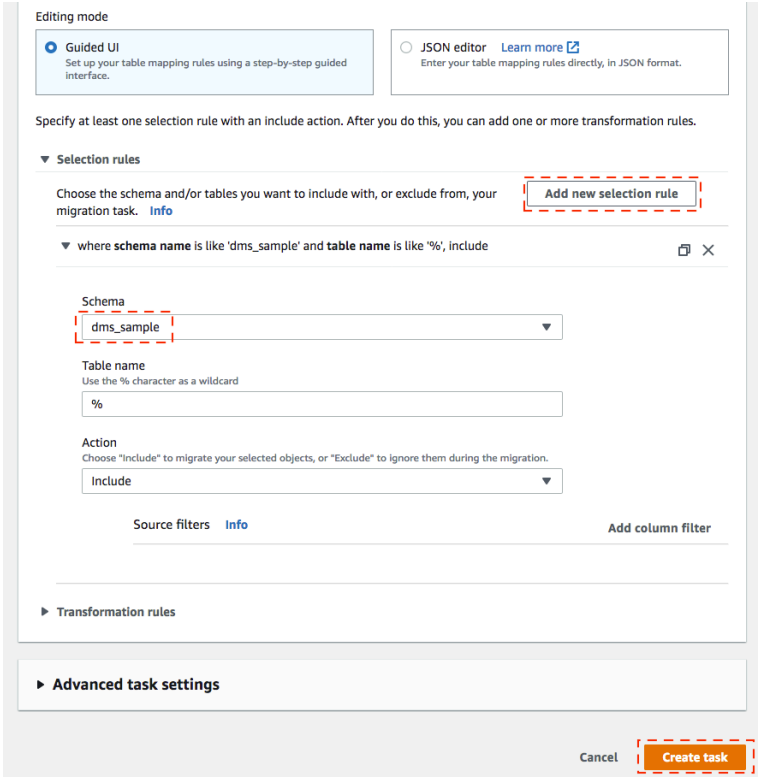
2. Click **Create Task**.
3. Type an easily recognized **Task Identifier**. For example “cdctask”.
4. Select your **Replication instance**.
5. Select your **Source endpoint**.
6. Select your **Target endpoint** as cdc endpoint created in previous section.
7. For **Migration type**, choose **Replicate data changes only**.
8. Select the Start task on create check box.



9. In **Task Settings**, Select the **Enable CloudWatch logs** check box.

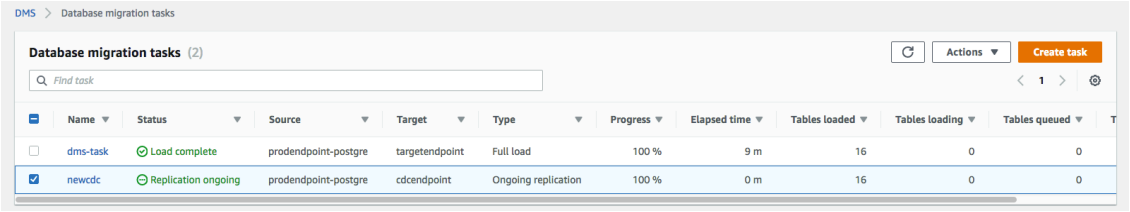


10. Go to **Table Mappings**.
11. Click on **Add new selection rule**
12. For **Schema name**, select **dms_sample** from drop down. Keep the settings for the remaining fields



13. Click **Create task**.

14. Your task is created and starts automatically. You can see status as ongoing replication, after couple of minutes.

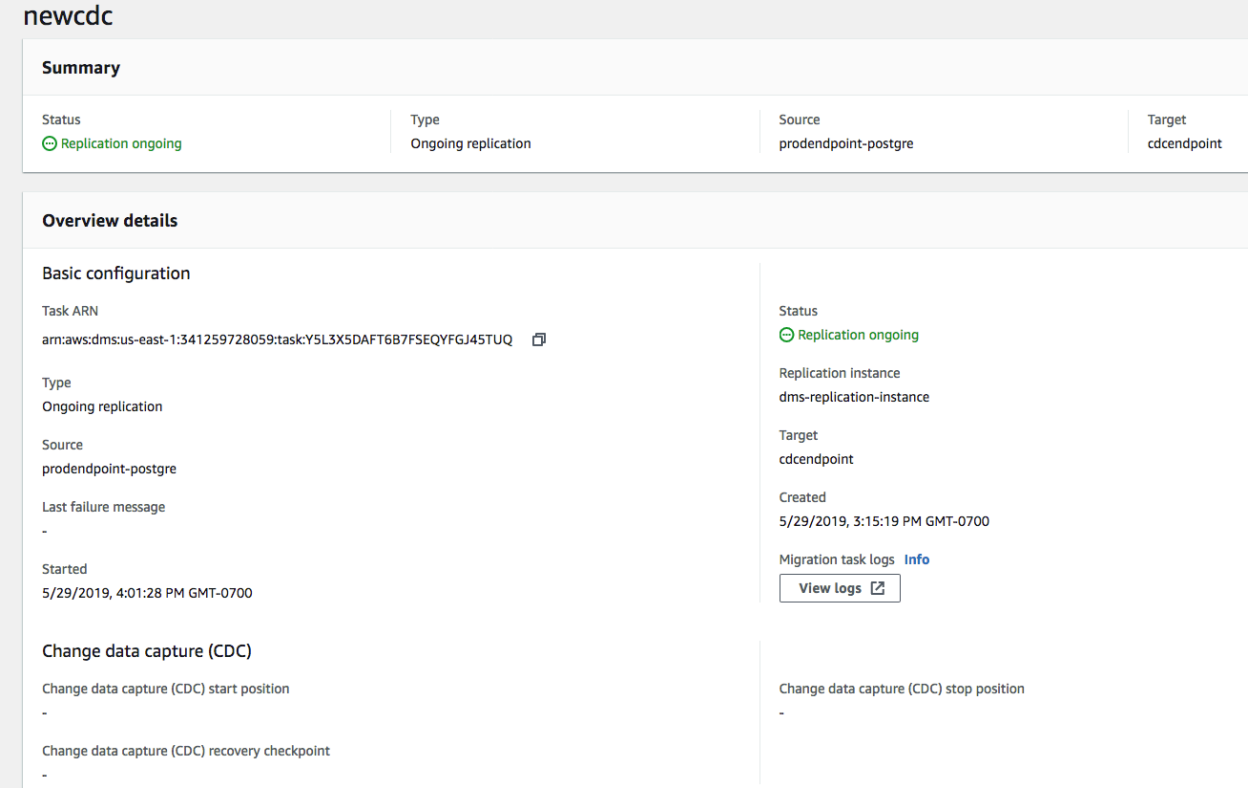


Once complete, the console displays 100% complete.

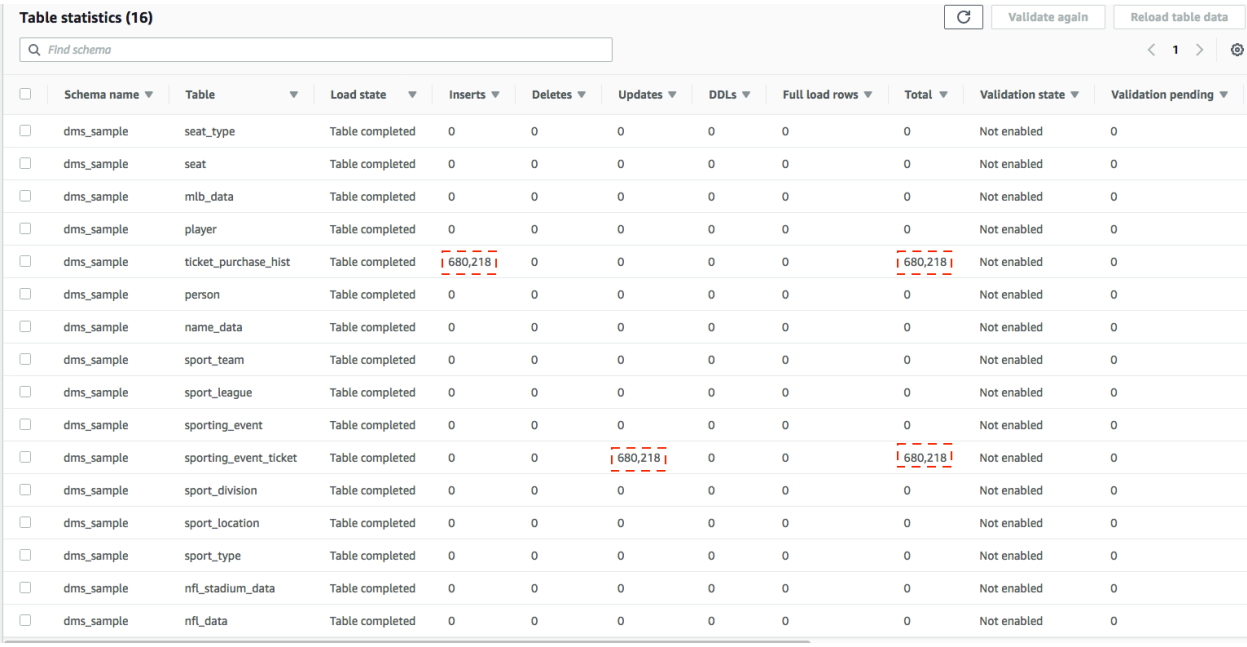
15. Your instructor will generate CDC activity which above migration task will capture, if you ran instructor setup by own, then make sure to follow "**Generate the CDC Data"** section from instructor lab.

    You may need to wait 5 to 10 minutes for CDC data to first reflect in your RDS postgre database and then picked up by DMS CDC migration task.

16. Select your CDC task and explore the summary:



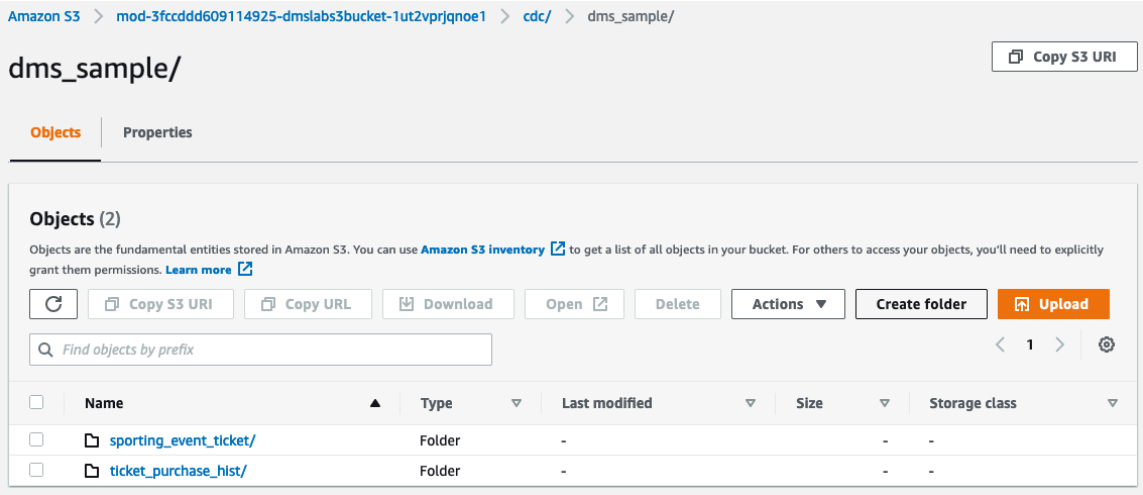
Scroll down and you will see all table changes impacted by CDC:



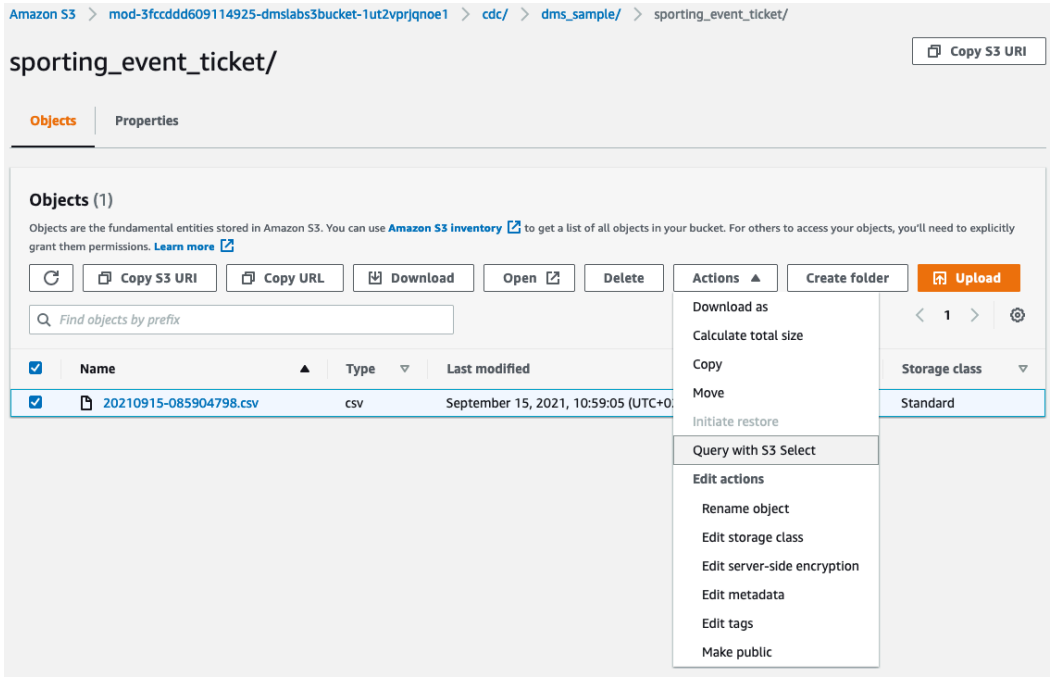
17. Open the S3 console and view the CDC data that was copied by DMS.

Your S3 bucket name will look like below :
BucketName/bucket_folder_name/schema_name/table_name/objects/

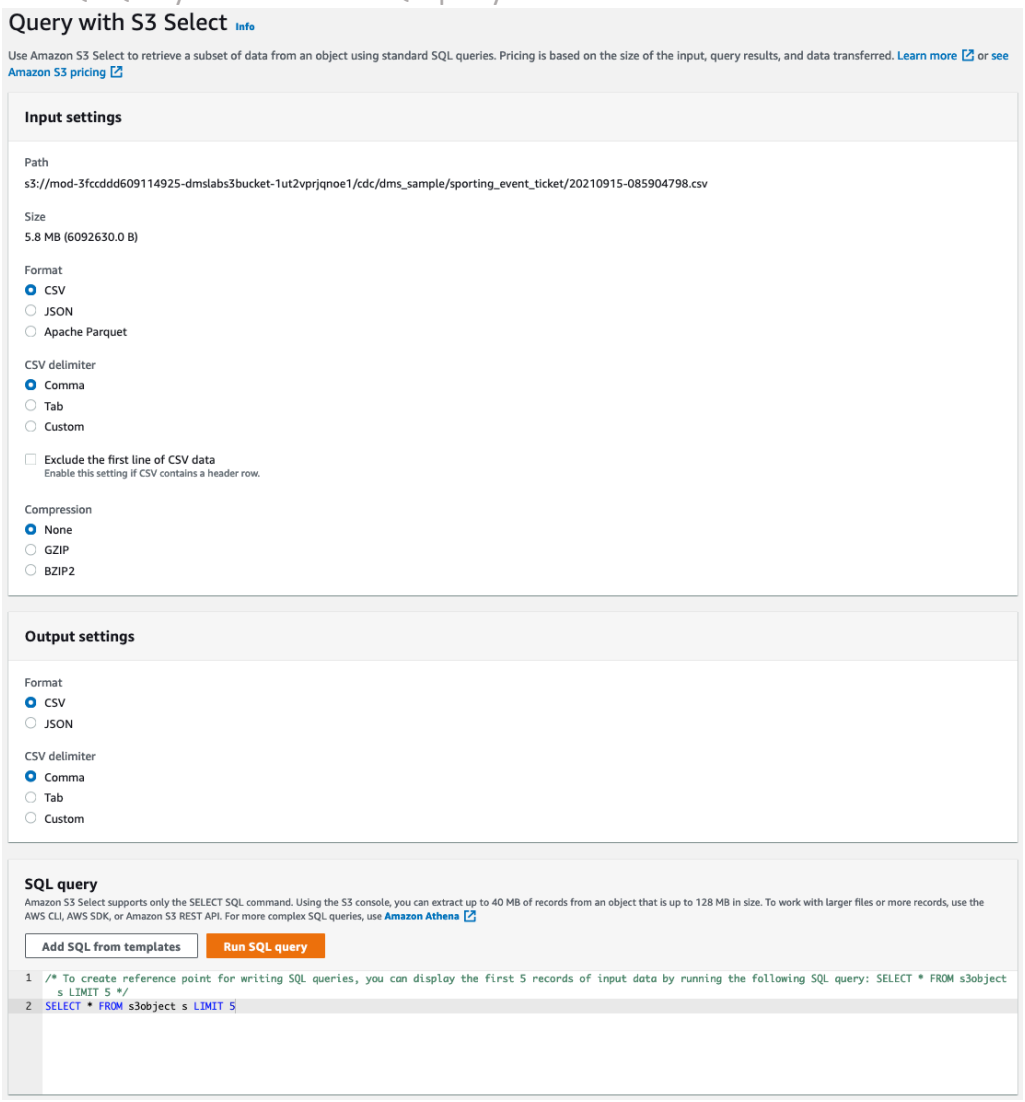
In our lab example this becomes:
"/dmslab-student-dmslabs3bucket-woti4bf73cw3/cdc/dms_sample" with a separate path for each table_name)



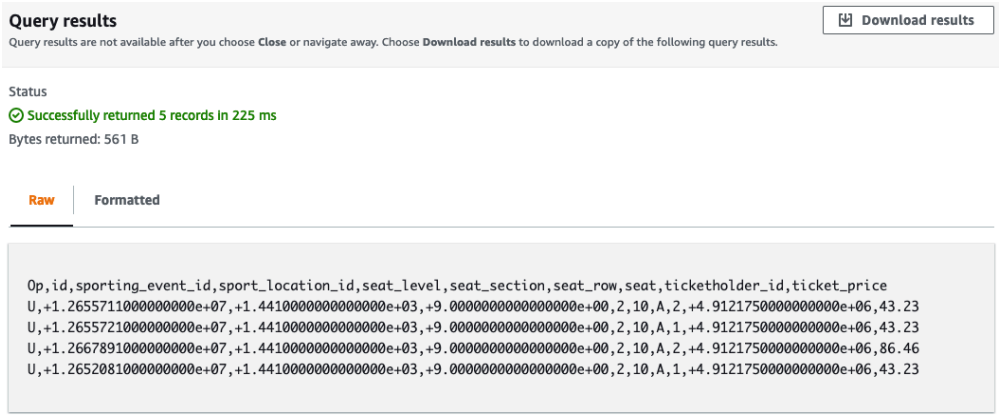
17. Navigate to one of the files and review it using S3 Select:
    a. Navigate in to the directory named **player** and select the check box next to the file name.
    b. Click the **Actions** dropdown button and choose **Query with S3 Select**.



c. In the Query with S3 Select page, leave the default value for Input Settings and SQL Query and click Run SQL query.



d. It will execute the specified SQL query and return the first 5 lines from the CSV 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 file name has date time - 20210915-085904798.csv

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.

Explore the objects in the S3 directory further.