# Amazon Web Services

Lab1. Copy RDS Source Data - Prelab

***March 2021***

## Table of Contents



# Lab2. Copy RDS Source Data - Prelab

## About the lab setup:


RDS Postgres Database is used as a source of ticket sales system for sporting events. It stores transaction information about ticket sales price to selected people and ticket ownership transfer with additional tables for event details. AWS Database Migration Service (DMS) is used for a full data load from the Amazon RDS source to Amazon S3 bucket.

Before the Glue lab starts, you might choose to skip the DMS data migration, instead copy the source data to your S3 bucket directly.

In today's lab, you will copy the data from a centralized S3 bucket to your AWS account, crawl the dataset with AWS Glue crawler for metadata creation and transform the data with AWS Glue to Query data and create a View with Athena and Build a dashboard with Amazon QuickSight.

*****Make sure you are in the us-east-1 (Virginia) region*****

## Setup Cloud9 IDE for Data Copy

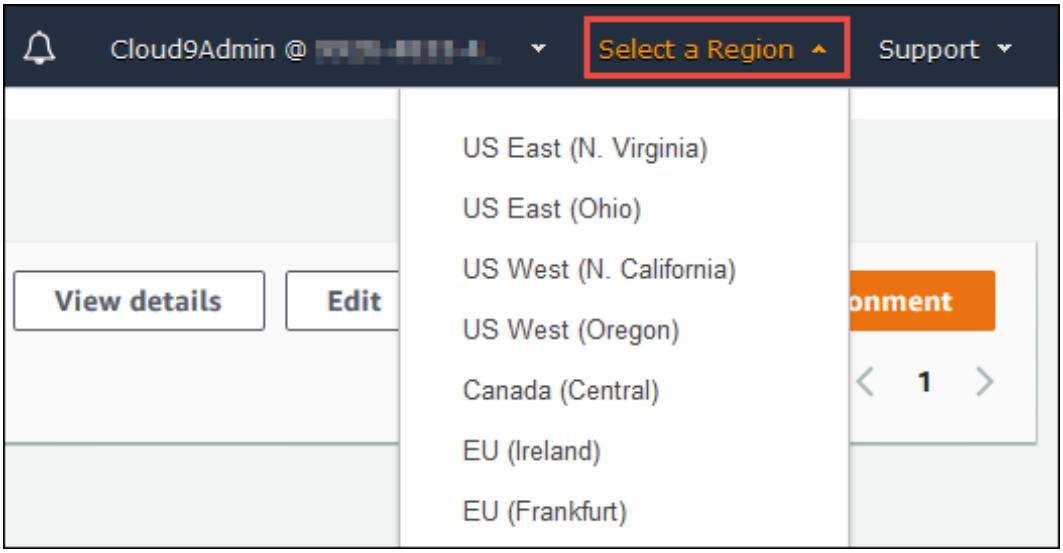
In AWS Cloud9, a development environment (or just environment) is a place where you store your development project's files and where you run the tools to develop your applications. In this tutorial, you create a special kind of environment called an EC2 environment and then work with the files and tools in that environment.

### Create an EC2 Environment with the Console

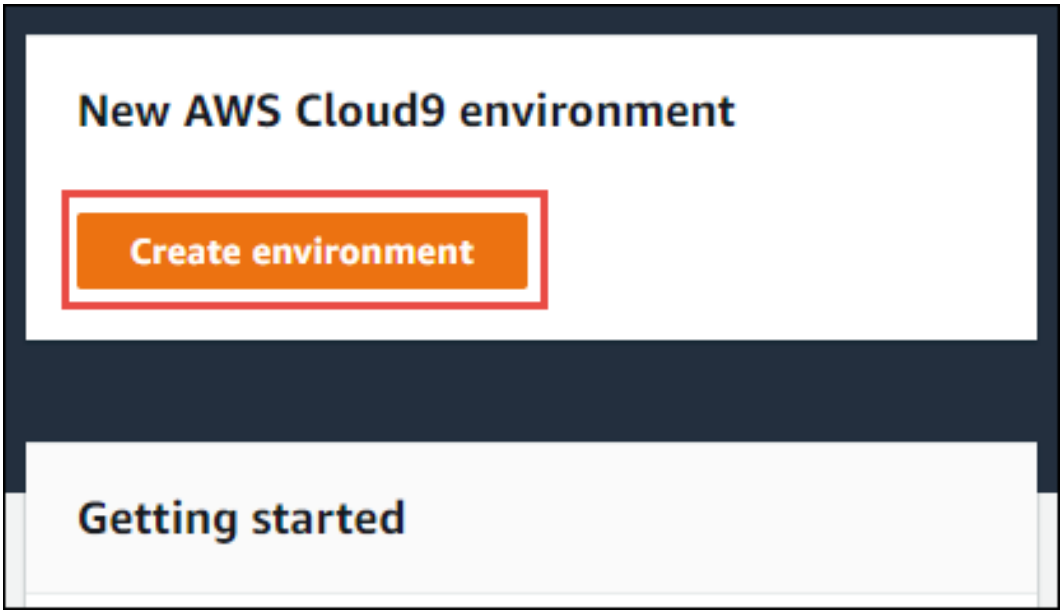
1. Sign in to the AWS Cloud9 console as follows:
    - If you're the only individual using your AWS account or you are an IAM user in a single AWS account, go to https://console.aws.amazon.com/cloud9/.

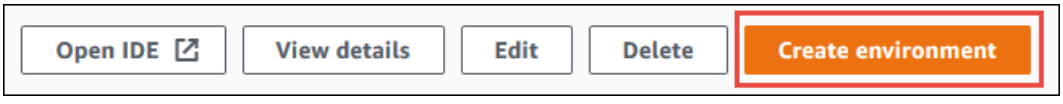
2. After you sign in to the AWS Cloud9 console, in the top navigation bar, choose **US East (N. Virginia)** AWS Region to create the environment in. For a list of available AWS Regions, see AWS Cloud9 in the *AWS General Reference*.




3. If a welcome page is displayed, for **New AWS Cloud9 environment**, choose **Create environment**. Otherwise, choose **Create environment**.




Or:



4. On the **Name environment** page, for **Name**, type a name for your environment. For this tutorial, use **my-demo-environment**.

5. For **Description**, type something about your environment. For this tutorial, use This environment is for the AWS Cloud9 tutorial.

6. Choose **Next step**.

7. On the **Configure settings** page, for **Environment type**, choose **Create a new instance for environment (EC2)**.

8. For **Instance type**, leave the default choice. This choice has relatively low RAM and vCPUs, which is sufficient for this tutorial.

9. For **Platform**, choose the type of Amazon EC2 instance that AWS Cloud9 will create and then connect to this environment: **Amazon Linux** or **Ubuntu**.

10. For **Cost-saving setting**, choose the amount of time until AWS Cloud9 shuts down the Amazon EC2 instance for the environment after all web browser instances that are connected to the IDE for the environment have been closed. Or **leave** the default choice.

11. Leave the default settings for **Network settings (advanced)**.

12. Choose **Next step**.

13. On the **Review** page, choose **Create environment**. Wait while AWS Cloud9 creates your environment. This can take several minutes.

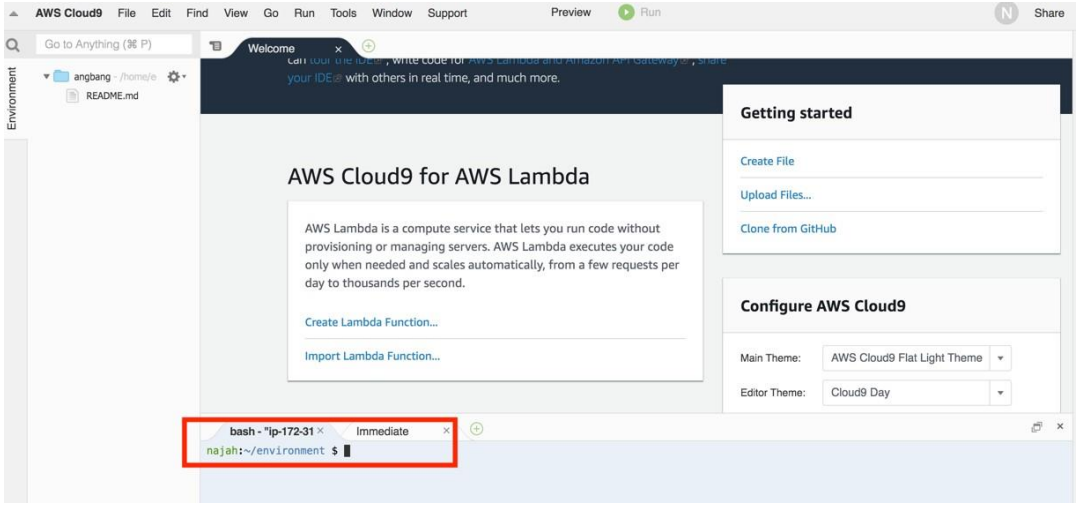
After AWS Cloud9 creates your environment, it displays the AWS Cloud9 IDE for the environment.

If AWS Cloud9 doesn't display the IDE after at least five minutes, there might be a problem with your web browser, your AWS access permissions, the instance, or the associated virtual private cloud (VPC). For possible fixes, see Cannot Open an Environment in *Troubleshooting*.

## Copy Data across from staging Amazon S3 bucket to your S3 bucket

Open Cloud9 Console from AWS and you will see the terminal screen in the bottom:




a) Generate a key pair by issuing the following command

```
ssh-keygen
```

b) Press enter3 times to take the default choices

c) Upload the public key to your EC2 region:

```
aws ec2 import-key-pair --key-name "lfworkshop" --public-key-material
file://~/.ssh/id_rsa.pub
```

d) Issue the following command in the terminal, and replace the bucket name with your existing or a new S3 bucket dedicated for the lab.

NOTE: if you are in an AWS hosted event, the destination S3 bucket is created on your behalf. Go to S3 console, search for a keyword **dmslabs3bucket**. The bucket name looks like "xxx-dmslabs3bucket-xxxx"

```
aws s3 cp --recursive s3://consuming-datalake-staging/tickets/
s3://<BucketName>/tickets/
```

The data will be copied to your S3 Bucket and you will see the following:

```
-1d3xq6mdkqx1n/tickets/dms_sample/nfl_data/LOAD00000001.csv
copy: s3://consuming-datalake-staging/tickets/dms_sample/nfl_stadium_data/LOAD00000001.csv to s3://lake-formation-wf-dmslab
s3bucket-1d3xq6mdkqx1n/tickets/dms_sample/nfl_stadium_data/LOAD00000001.csv
copy: s3://consuming-datalake-staging/tickets/dms_sample/name_data/LOAD00000001.csv to s3://lake-formation-wf-dmslabs3bucke
t-1d3xq6mdkqx1n/tickets/dms_sample/name_data/LOAD00000001.csv
copy: s3://consuming-datalake-staging/tickets/dms_sample/sporting_event_ticket/LOAD00000002.csv to s3://lake-formation-wf-d
mslabs3bucket-1d3xq6mdkqx1n/tickets/dms_sample/sporting_event_ticket/LOAD00000002.csv
copy: s3://consuming-datalake-staging/tickets/dms_sample/person/LOAD00000001.csv to s3://lake-formation-wf-dmslabs3bucket-1
d3xq6mdkqx1n/tickets/dms_sample/person/LOAD00000001.csv
copy: s3://consuming-datalake-staging/tickets/dms_sample/ticket_purchase_hist/LOAD00000001.csv to s3://lake-formation-wf-dm
slabs3bucket-1d3xq6mdkqx1n/tickets/dms_sample/ticket_purchase_hist/LOAD00000001.csv
copy: s3://consuming-datalake-staging/tickets/dms_sample/sporting_event_ticket/LOAD00000001.csv to s3://lake-formation-wf-d
mslabs3bucket-1d3xq6mdkqx1n/tickets/dms_sample/sporting_event_ticket/LOAD00000001.csv
AdministratorAccess:~/environment $
```

## Verify the Data

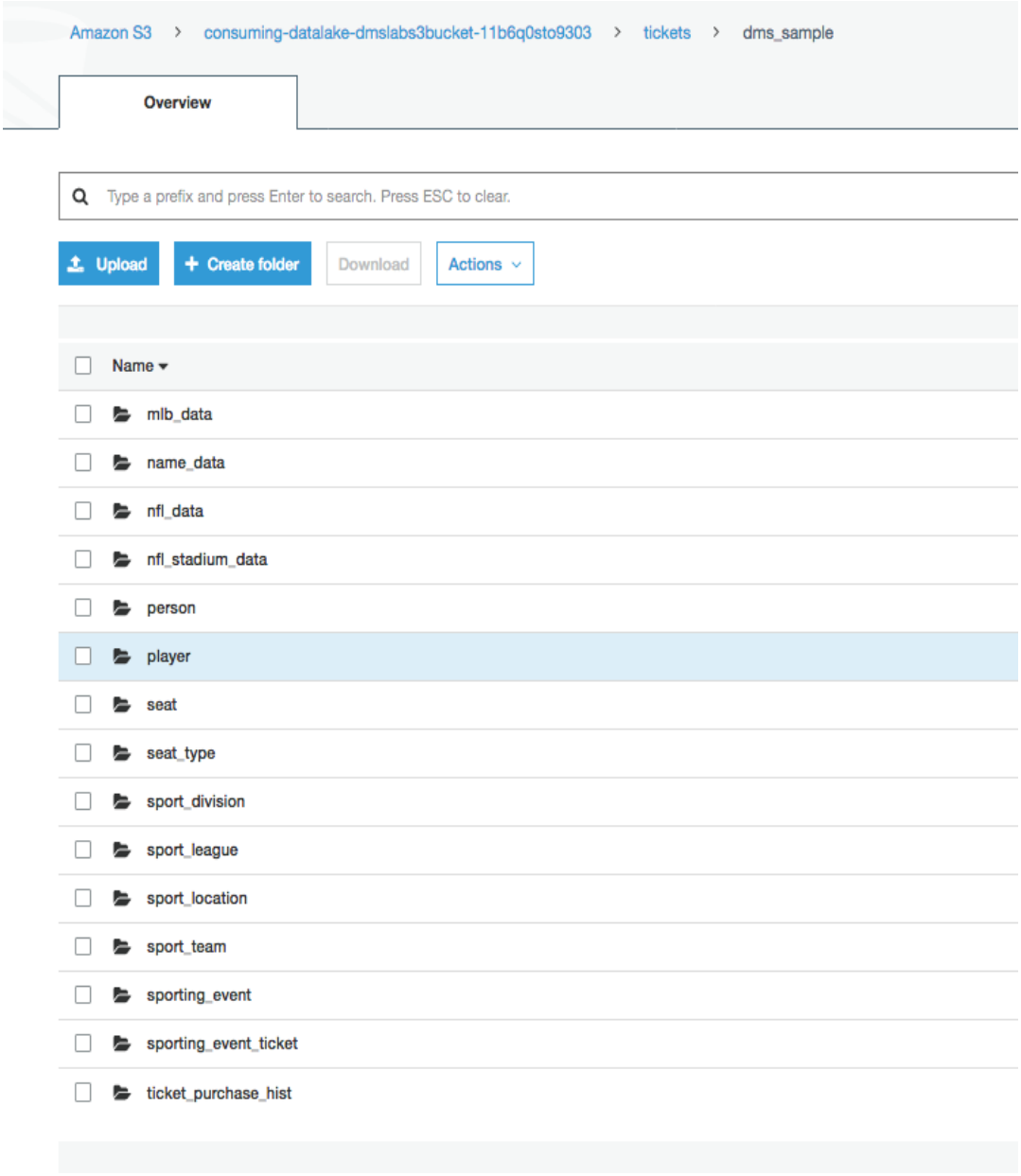
 

# Lab2. Copy RDS Source Data - Prelab

1. Open the S3 console and view the data that was copied from Cloud9 terminal.

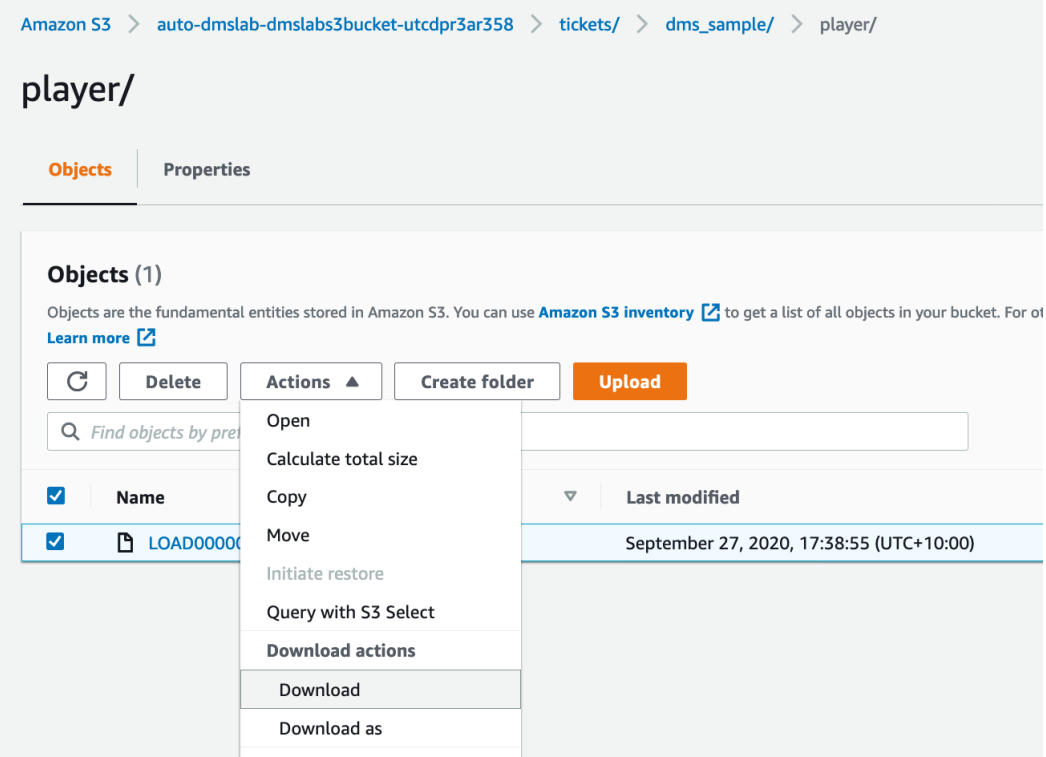
2. Your data files in a S3 bucket will look like this :
BucketName/bucket_folder_name/schema_name/table_name/csv files/

3. In our lab example, this becomes: "/<**BucketName**>/tickets/dms_sample" with a separate path for each table_name




4. Download one of the files:
    a. Select a table/folder name, tick the check box next to a CSV file name, and choose **Download** option from the **Actions** dropdown list.
    b. Click **Save File**.
    c. Open the file.




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 |

Explore the objects in the S3 directory further.

## Next Steps

In the next part of this lab, we will complete the following tasks:

- Extract, Transform and Load Data Lake with AWS Glue

## Appendix A: Self-Paced Data Lake Lab

If you If want to re-run the lab by yourself, please follow the lab instruction published in the GitHub:

https://github.com/aws-samples/data-engineering-for-aws-immersion-day