

Amazon Web Services

Lab1. Copy RDS Source Data - Prelab September 2021

Table of Contents

About the lab setup:	3
Setup Cloud9 IDE for Data CopyErro	r! Bookmark not defined.
Create an EC2 Environment with the Console	Error! Bookmark not defined.
Copy Data across from staging Amazon S3 bucket to your S3 bucketErro	r! Bookmark not defined.
Verify the DataErro	r! Bookmark not defined.
Next Steps	7
Appendix A: Self-Paced Data Lake LabErro	r! Bookmark not defined.

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.

Open AWS CloudShell

Open <u>AWS CloudShell</u> in us-east-1 (N. Virginia) region. It will open a terminal window in the browser. (If there is a pop-up, close it)

We will be launching CloudShell in us-east-1 (N. Virginia) region irrespective of where you are running this whole workshop. By executing the following command you will be copying the data to the correct S3 bucket in whatever region it belongs (It can also be across region).

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

1. Issue the following command in the terminal, and replace the bucket name with your own one.

aws s3 cp --recursive --copy-props none s3://aws-dataengineering-day.workshop.aws/data/ s3://<YourBucketName>/tickets/

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

AWS CloudShell	Actions v	۲
us-east-1		
Preparing your terminal [cloudshell-user@ip-10-1-58-162 ~]\$ Try these commands to get started: ans help or aws «command» help or aws «command»cli-auto-prompt [cloudshell-user@ip-10-1-58-162 ~]\$ aws s3 cprecursivecopy-props none s3://aws-dataengineering-day.workshop.aws/data mslob3bucket-3or58-162 ~]\$ aws s3 cprecursivecopy-props none s3://aws-dataengineering-day.workshop.aws/data/ those is the sample of the sa	a/ s3://dmslab-st slabs3bucket-3or5 islabs3bucket-3or5 islabs3bucket-3or53urfr dmslabs3bucket-3o ibucket-3or53urfru ibucket-3or53urfru ibucket-3or53urfru ibucket-3or53urfru ibucket-3or53urfru islabs3bucket-3or53urfru isla	tudent- 53urfru 53urfru fru9/tic or53urf u9/tick /ticket u9/tick ru9/tic urfru9/ 53urfru ickets/ ket-3or
<pre>copy: s3://aws-dateengineering-day.workshop.aws/data/dms_sample/person/LOAD00000001.csv to s3://dmslab-student-dmslabs3bu s/dms_sample/person/LOAD00000001.csv</pre>	icket-3or53urfru9/	/ticket
<pre>copy: s3://aws-dataengineering-day.workshop.aws/data/dms_sample/sporting_event_ticket/LOAD00000001.csv to s3://dmslab-stu 53urfru9/tickets/dms_sample/sporting_event_ticket/LOAD00000001.csv [cloudshell-user@tp-10-1-58-162 ~]\$</pre>	dent-dmslabs3buck	ket-3or

Verify the Data

- 1. Open the S3 console and view the data that was copied through CloudShell terminal.
- 2. Your S3 bucket name will look like below :

BucketName/bucket_folder_name/schema_name/table_name/objects/

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

Amazon S	3 > mod-3fccddd609114925-dmslabs3bucker	-1ut2vprjqnoe1 >	tickets/ > dms_sample/			
dms_	_sample/					🗇 Copy S3 URI
Object	ts Properties					
Obje	cts (15)					
Objects grant th	are the fundamental entities stored in Amazon S3. You c nem permissions. Learn more	an use Amazon S3 inve	ntory 🗹 to get a list of all objects i	in your bucket. For others t	o access your objects, y	you'll need to explicitly
C	🗇 Copy S3 URI 🗇 Copy URL	🕑 Download	Open 🖸 Delete	Actions 🔻	Create folder	\Lambda Upload
Q F	ind objects by prefix					< 1 > 💿
-						
	Name 🔺	Туре		⊽ Size		e class 🗢
	🗅 mlb_data/	Folder	-			
	🗅 name_data/	Folder	-			
	🗅 nfl_data/	Folder	-			
	nfl_stadium_data/	Folder	-			
	🗅 person/	Folder	-			
	🗅 player/	Folder	-			
	seat_type/	Folder	-			
	🗅 seat/	Folder	-			
	Sport_division/	Folder	-			
	sport_league/	Folder	-			
	sport_location/	Folder	-			
	Sport_team/	Folder	-			
	sporting_event_ticket/	Folder	-			
	sporting_event/	Folder	-			
	ticket_purchase_hist/	Folder	-			

Copyright 2021, Amazon Web Services, All Rights Reserved Page 4

- 4. Navigate to one of the files and review it using <u>S3 Select</u>:
 - 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.

layer/		🗇 Copy S3 UR
Objects Properties		
Objects (1) Objects are the fundamental entities stored in Amazon 53. You can use Amazon 53 inventory 12 to get a list of all objects in y grant them permissions. Learn more 12 O O O O	our bucket. For others to access your obje	cts, you'll need to explicitly
C Di Copy S3 URI Di Copy URL Di Download Open Di Delete Q Find objects by prefix	Actions A Create fold Download as Calculate total size	er [r] Upload
✓ Name Type Last modified	Сору	Storage class
C LOAD00000001.csv csv September 15, 2021, 10:05:44 (UTC+02:00)	Move	Standard
	Initiate restore	
	Query with 55 Select	
	Banama abject	
	Edit storage class	
	Edit server-side encryption	
	Edit metadata	
	Edit tags	

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

Input settings	
Path	
s3://mod-3fccddd609114925-dm	:labs3bucket-1ut2vprjqnoe1/tickets/dms_sample/player/LOAD00000001.csv
Size	
393.3 KB (402738.0 B)	
Format	
Apache Parquet	
CSV delimiter	
Caston	
Exclude the first line of CSV da Enable this setting if CSV contains a	ta header row.
Compression	
None	
GZIP	
O BZIP2	
Output settings	
Format	
O CSV	
O JSON	
CSV delimiter	
Comma	
O Tab	
Custom	
SOL querv	
Amazon S3 Select supports only the SEL	ECT SQL command. Using the S3 console, you can extract up to 40 MB of records from an object that is up to 128 MB in size. To work with larger files or more records, use the AW
indeon by beleet supports only the see	or more complex SQL queries, use Amazon Athena 🔀
CLI, AWS SDK, or Amazon S3 REST API. F	
Add SQL from templates	Run SQL query
CLI, AWS SDK, or Amazon S3 REST API. F Add SQL from templates	Run SQL query t for writing SQL queries, you can display the first 5 records of input data by running the following SQL query: SELECT * FROM s3objec

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

Query results Query results are not available after you choose Close or navigate away. Choose Download results to download a copy of the following query results.	U Download results
Status	
Successfully returned 5 records in 208 ms Bytes returned: 352 B	
Raw Formatted	
id,sport_team_id,last_name,first_name,full_name	
+1.000000000000000e+00,+1.3100000000000000e+02,Adam Loewen,Adam , Loewen +1.10000000000000e+01,+1.31000000000000e+02,A.J. Pollock,A.J. , Pollock	
+2.1000000000000000+01,+1.3100000000000000+02,Alex Sanabia,Alex , Sanabia	
+3.1000000000000000000000000000000000000	

Note that column names are included in the file in the first row

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

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