

# COVID-19 PROJECT

## DATA ENGINEERING MASTERCLASS

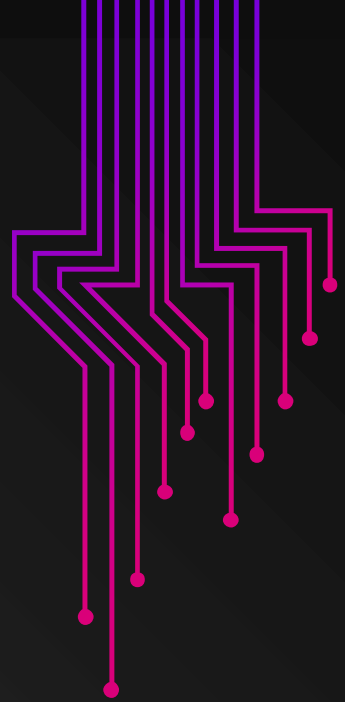
PRESENTED BY:

KIROLOS SAMIR YOSSIF GIRGIS

PRESENTED TO:

ENG. AMR SALEH

ENG. AHMED REDA



# THE ORIGINAL PLAN

## CREATE HDFS DIRECTORY

Create HDFS directory /ds containing /COVID\_HDFS\_LZ as a sub-directory.



## USING HIVE FOR ANALYSIS

Use Hive editor in Hue to create tables, external tables after cleaning the data.



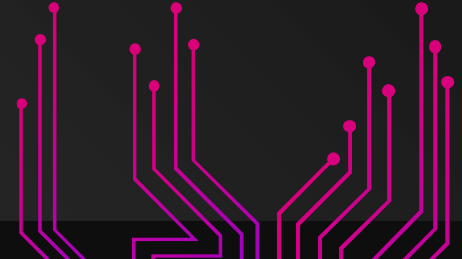
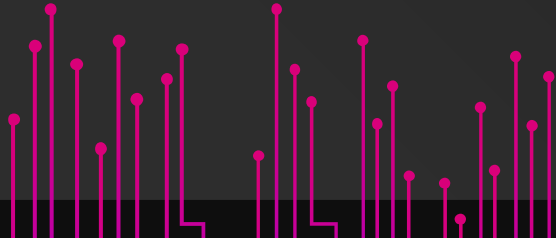
## LOAD THE DATA

Load dataset to Cloudera QuickStart VM using WinSCP to create two folders landing\_zone and scripts.



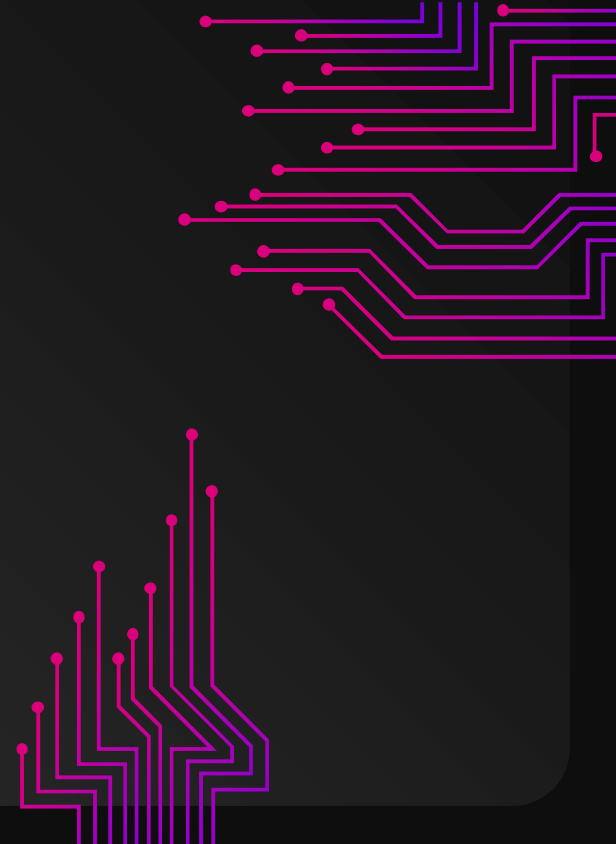
## LOAD DATA TO HDFS

Load the data from landing\_zone to HDFS directory /ds/COVID\_HDFS\_LZ using Load\_COVID\_TO\_HDFS.sh script.



# ISSUES PREVENTED THE **CONTINUITY** OF THE PLAN

- Following the original plan, many configuration issues appeared prevented the continuity of the plan.
- After loading the data to HDFS directories, Hive should be used for the analysis, but Hive crashed many times when creating the external and final output tables.
- The only solution for these issues is to use another way to perform these operations.
- **AWS was the solution used.**



# TABLE OF CONTENTS

01

DATA STORAGE



03

DATA ANALYSIS



02

DATA INGESTION



04

DATA VISUALIZATION



01

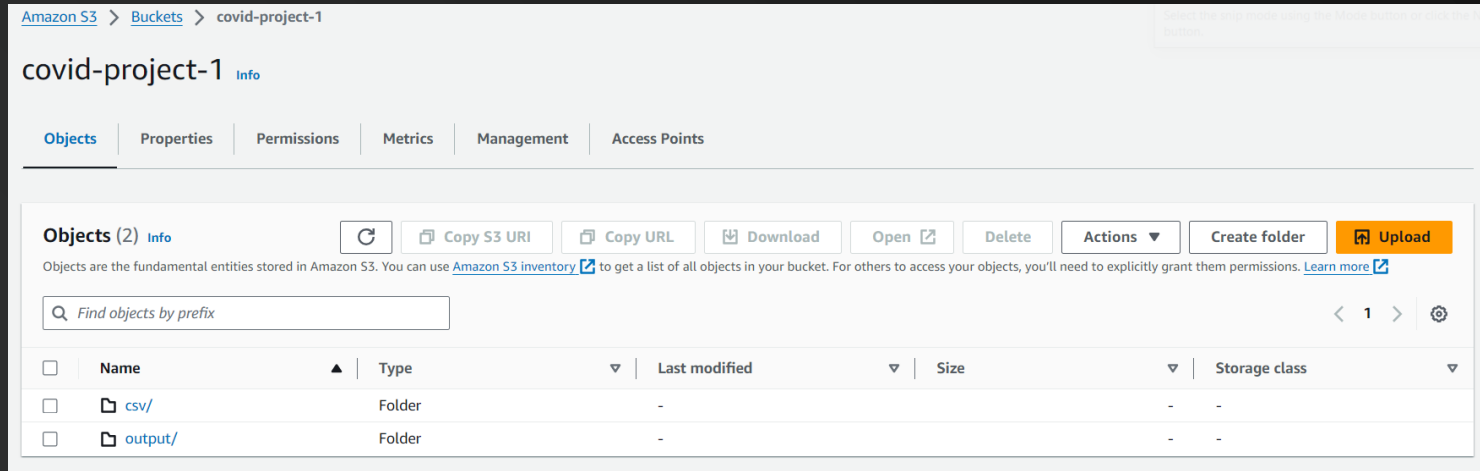
# DATA STORAGE

Using AWS S3



# 1.1 Creating S3 Bucket and Folder Structure

- Starting with creating a S3 bucket (covid-project-1) that contains two subfolders as follows:
  - “csv” folder that will contain the source data covid-19.csv.
  - “output” folder that will contain the final output csv files.

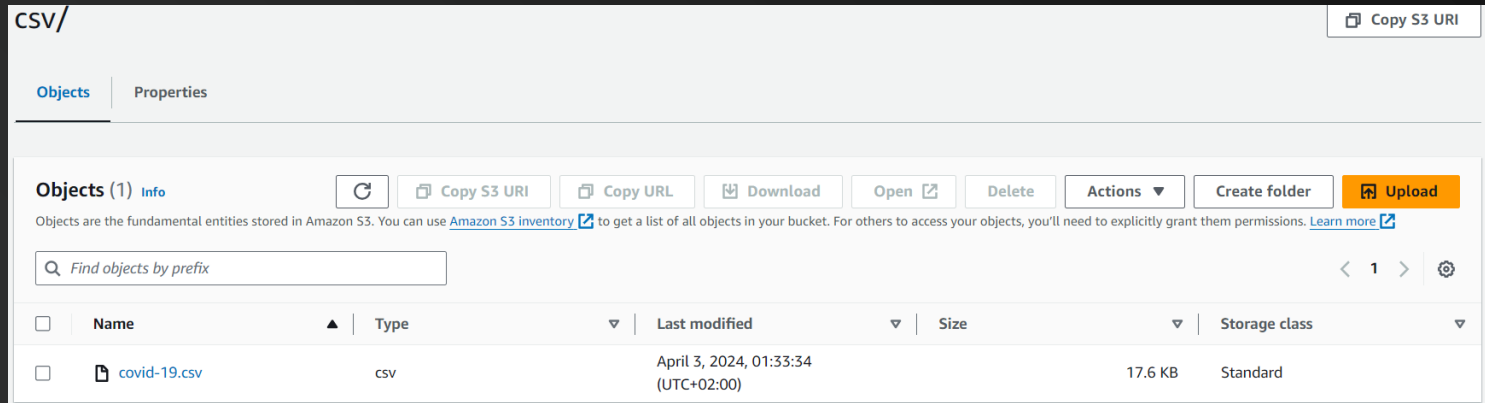


The screenshot displays the Amazon S3 console interface for a bucket named 'covid-project-1'. The breadcrumb navigation at the top shows 'Amazon S3 > Buckets > covid-project-1'. Below the bucket name, there are tabs for 'Objects', 'Properties', 'Permissions', 'Metrics', 'Management', and 'Access Points'. The 'Objects' tab is active, showing a list of objects. Above the list, there are buttons for 'Copy S3 URI', 'Copy URL', 'Download', 'Open', 'Delete', 'Actions', 'Create folder', and 'Upload'. A search bar with the placeholder 'Find objects by prefix' is also present. The object list table has columns for 'Name', 'Type', 'Last modified', 'Size', and 'Storage class'. Two folders are listed: 'csv/' and 'output/'.

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	csv/	Folder	-	-	-
<input type="checkbox"/>	output/	Folder	-	-	-

# 1.2 Uploading data source to /CSV

- Uploading the source data named “covid-19.csv” to the /CSV folder in S3 bucket (covid-project-1).



The screenshot displays the Amazon S3 console interface for a bucket named 'covid-project-1'. The path '/CSV/' is shown at the top. The 'Objects' tab is selected, showing a list of objects. A single object, 'covid-19.csv', is listed with a size of 17.6 KB and a storage class of 'Standard'. The object was last modified on April 3, 2024, at 01:33:34 (UTC+02:00). The console includes various action buttons like 'Copy S3 URI', 'Copy URL', 'Download', 'Open', 'Delete', 'Actions', 'Create folder', and 'Upload'.

CSV/ Copy S3 URI

Objects Properties

Objects (1) [Info](#) Refresh Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	covid-19.csv	csv	April 3, 2024, 01:33:34 (UTC+02:00)	17.6 KB	Standard

02

# DATA INGESTION

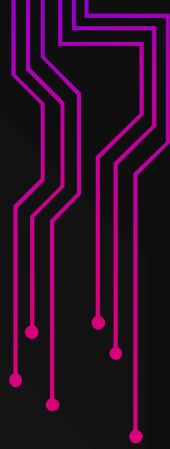
Using AWS **GLUE**





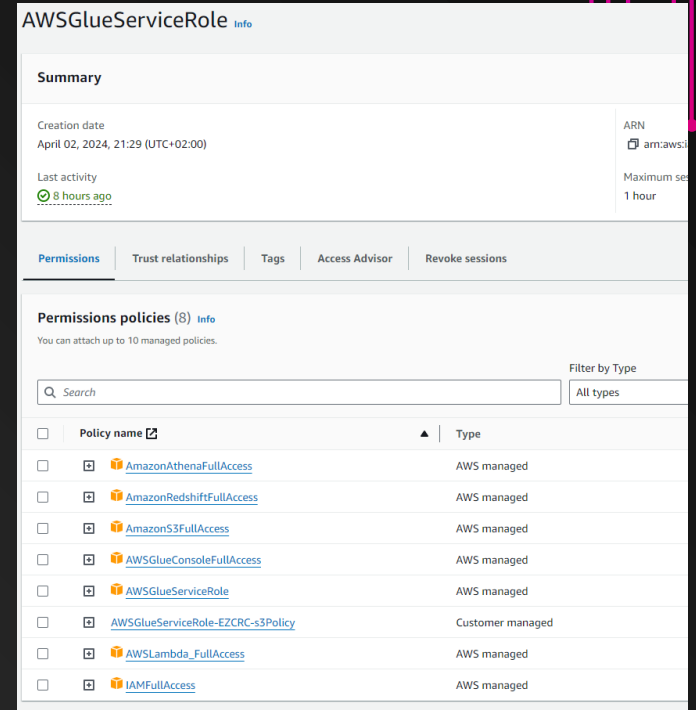
## 2.1 Creating Glue Crawler

- Starting with create crawler named “S3-To-Glue” and choosing the data source from S3.
- Creating IAM Role that will help Glue to read the data from s3 and use COPY command to load the data into AWS Redshift.
- Creating database in which we will store the metadata and schema.
- Running the crawler created to confirm that it is working and all configuration are done properly.



## 2.2 Creating Redshift serverless Workgroup

- Creating new workgroup named “covid-project”.
- Configuring a Virtual Private Cloud (VPC), VPC security group, and subnets.
- Creating a namespace named “covid-19” that contains public database and dev schema.
- Associating IAM role in the permissions that allow redshift to have full access to S3 and Glue.



The screenshot displays the AWS IAM console page for the **AWSGlueServiceRole**. The page is divided into several sections:

- Summary:** Shows the role's creation date as April 02, 2024, 21:29 (UTC+02:00), its ARN as `arn:aws:iam::111111111111:role/AWSGlueServiceRole`, and its last activity as 8 hours ago.
- Permissions:** A tabbed interface showing the role's permissions. It lists 8 policies attached to the role.
- Permissions policies (8):** A table listing the attached policies, including AWS managed policies like `AmazonAthenaFullAccess`, `AmazonRedshiftFullAccess`, `AmazonS3FullAccess`, `AWSGlueConsoleFullAccess`, `AWSGlueServiceRole`, `AWSLambda_FullAccess`, and `IAMFullAccess`, as well as a customer-managed policy `AWSGlueServiceRole-EZCRC-s3Policy`.

Policy name	Type
<code>AmazonAthenaFullAccess</code>	AWS managed
<code>AmazonRedshiftFullAccess</code>	AWS managed
<code>AmazonS3FullAccess</code>	AWS managed
<code>AWSGlueConsoleFullAccess</code>	AWS managed
<code>AWSGlueServiceRole</code>	AWS managed
<code>AWSGlueServiceRole-EZCRC-s3Policy</code>	Customer managed
<code>AWSLambda_FullAccess</code>	AWS managed
<code>IAMFullAccess</code>	AWS managed

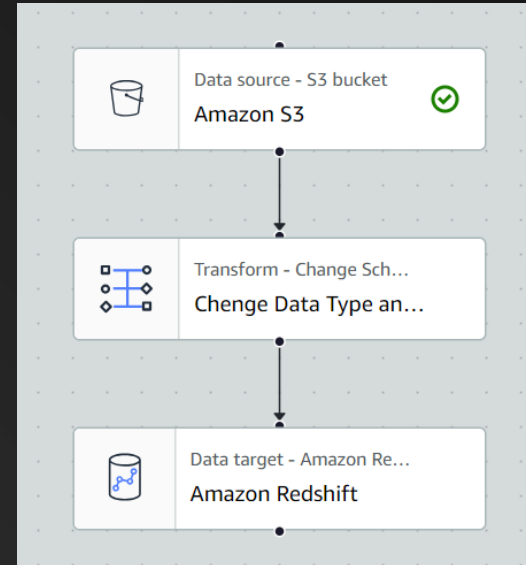
## 2.2 Creating Redshift serverless Workgroup ..cont.

- Navigating to Redshift Query editor to create a new table in public database and dev schema named “covid\_staging” with the following code:

```
▶ Run Limit 100 Explain Isolated session ⓘ  
1 CREATE TABLE IF NOT EXISTS covid_staging  
2 (  
3     Country VARCHAR,  
4     Total_Cases DOUBLE PRECISION,  
5     New_Cases DOUBLE PRECISION,  
6     Total_Deaths DOUBLE PRECISION,  
7     New_Deaths DOUBLE PRECISION,  
8     Total_Recovered DOUBLE PRECISION,  
9     Active_Cases DOUBLE PRECISION,  
10    Serious DOUBLE PRECISION,  
11    Tot_Cases DOUBLE PRECISION,  
12    Deaths DOUBLE PRECISION,  
13    Total_Tests DOUBLE PRECISION,  
14    Tests DOUBLE PRECISION,  
15    CASES_per_Test DOUBLE PRECISION,  
16    Death_in_Closed_Cases DOUBLE PRECISION,  
17    Rank_by_Testing_rate DOUBLE PRECISION,  
18    Rank_by_Death_rate DOUBLE PRECISION,  
19    Rank_by_Cases_rate DOUBLE PRECISION,  
20    Rank_by_Death_of_Closed_Cases DOUBLE PRECISION  
21 );
```

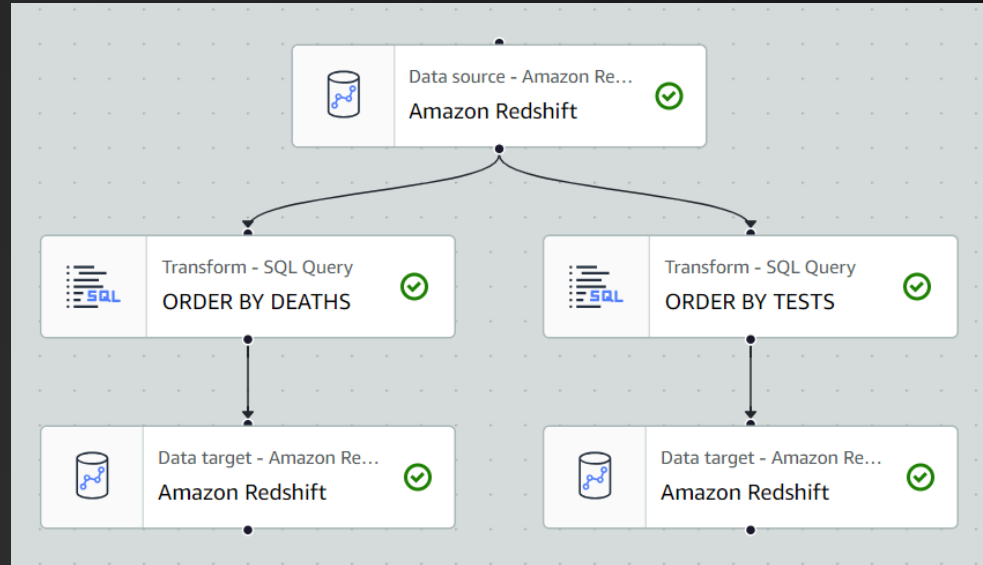
## 2.3 Creating Glue ETL Jobs

- Creating Glue two ETL Jobs as follows:
  1. The first one to ingest the data from the S3 bucket /CSV as a data source, change the data types and column names to match the “covid\_staging table created, and load the data to Redshift.



## 2.3 Creating Glue ETL Jobs ..Cont.

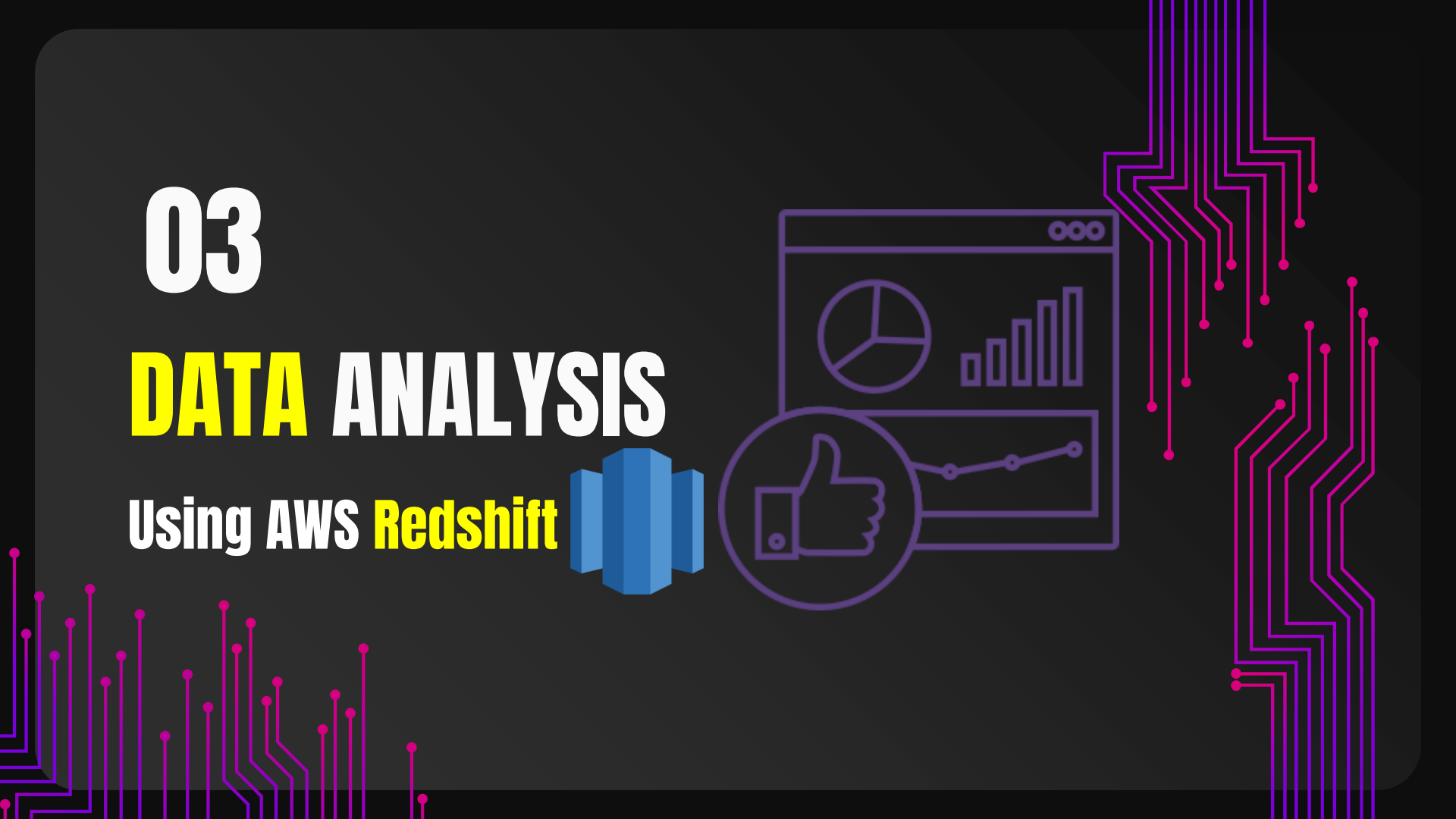
- Creating Glue two ETL Jobs as follows:
  1. The First uses redshift table and sorted and order it by deaths and tests and then load it to new external tables in redshift.
  2. The Second uses redshift table and sorted and order it by deaths and tests and then load it to new external tables in redshift.



03

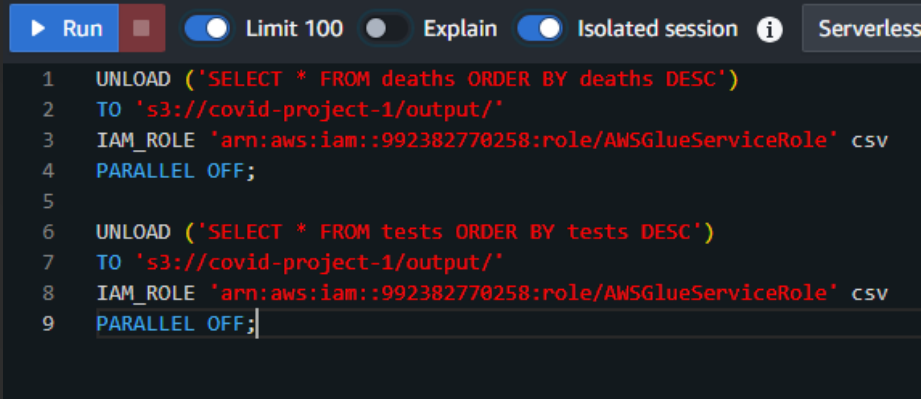
# DATA ANALYSIS

Using AWS Redshift



## 3.1 Analyzing the data in Redshift

- Making sure that the data is loaded to “covid\_staging” table from AWS Glue.
- Making sure that the data is loaded to deaths and tests and cleaning and preparing these data to be sent to S3 bucket “covid-project-1” /OUTPUT.
- Unload the data to the designated bucket as a .csv file using the following code:

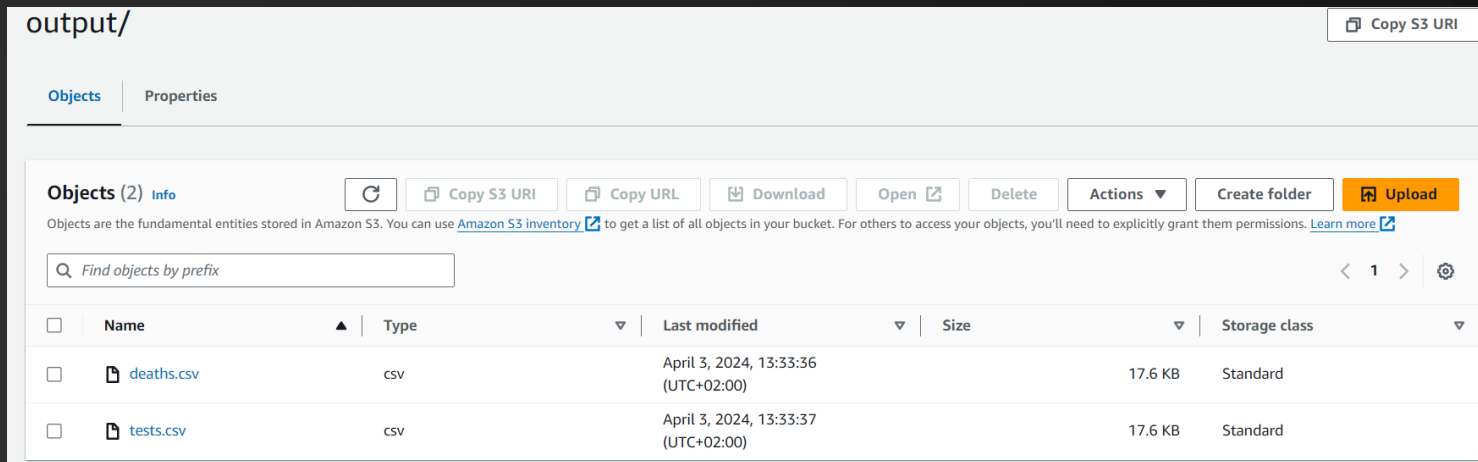


The screenshot shows the AWS Redshift console query editor interface. At the top, there is a toolbar with buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and 'Serverless'. Below the toolbar, the query editor contains two SQL statements. The first statement unloads data from the 'deaths' table into an S3 bucket, and the second statement unloads data from the 'tests' table into the same S3 bucket. Both statements use the 'IAM\_ROLE' parameter to specify the role used for the unload operation.

```
1 UNLOAD ('SELECT * FROM deaths ORDER BY deaths DESC')
2 TO 's3://covid-project-1/output/'
3 IAM_ROLE 'arn:aws:iam::992382770258:role/AWSGlueServiceRole' csv
4 PARALLEL OFF;
5
6 UNLOAD ('SELECT * FROM tests ORDER BY tests DESC')
7 TO 's3://covid-project-1/output/'
8 IAM_ROLE 'arn:aws:iam::992382770258:role/AWSGlueServiceRole' csv
9 PARALLEL OFF;
```

## 3.2 Downloading data from /OUTPUT

- Downloading the data sent to /OUTPUT folder in S3 bucket (covid-project-1) to start the visualization phase.



output/ Copy S3 URI

**Objects** | Properties

**Objects (2)** [Info](#) Refresh Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	<a href="#">deaths.csv</a>	csv	April 3, 2024, 13:33:36 (UTC+02:00)	17.6 KB	Standard
<input type="checkbox"/>	<a href="#">tests.csv</a>	csv	April 3, 2024, 13:33:37 (UTC+02:00)	17.6 KB	Standard



04

# DATA VISUALIZATION

Using MS POWER BI



# 4.1 Creating Power BI Dashboard

## COVID-19 VISUALIZATION

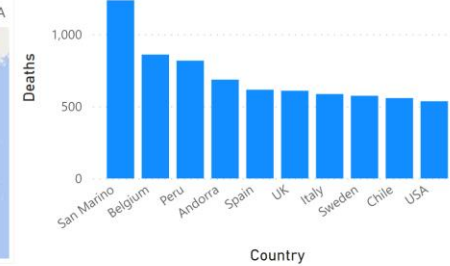
#	Country	Deaths
1	San Marino	1,237.00
2	Belgium	860.00
3	Peru	818.00
4	Andorra	686.00
5	Spain	616.00
6	UK	609.00
7	Italy	586.00
8	Sweden	574.00
9	Chile	558.00
10	USA	536.00
<b>Total</b>		<b>7,080.00</b>

Top 10 Deaths Rate by Country and Country

Country ● San Mari... ● Belgium ● Peru ● Andorra ● Spain ● UK ● Italy ● Sweden ● Chile ● USA



Deaths by Country



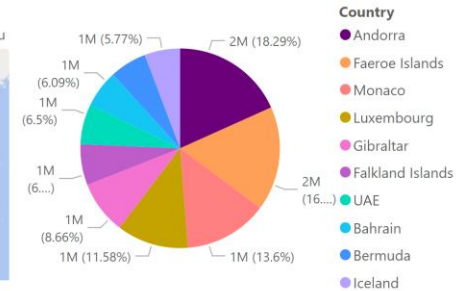
#	Country	Tests
1	Andorra	1778642
2	Faeroe Islands	1642742
3	Monaco	1322632
4	Luxembourg	1126386
5	Gibraltar	841971
6	Falkland Islands	645863
7	UAE	632496
8	Bahrain	592064
9	Bermuda	581621
10	Iceland	561236
<b>Total</b>		<b>9725653</b>

Top 10 Test Rate by Country and Country

Country ● Andorra ● UK ● USA ● San Mari... ● Belgium ● Spain ● Italy ● Chile ● Sweden ● Peru



Tests by Country



## 4.2 Project **GitHub** Link

<https://github.com/kirolosgirgis/Data-Engineering-MC/tree/main/O5-Kirollos%20Graduation%20Project>

The image features a dark gray background with a large, rounded rectangular area in the center. This central area is framed by decorative circuit-like lines in shades of purple and pink. These lines are arranged in a way that suggests a network or data flow, with some lines ending in small dots. The overall aesthetic is modern and tech-oriented.

# Thanks!