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

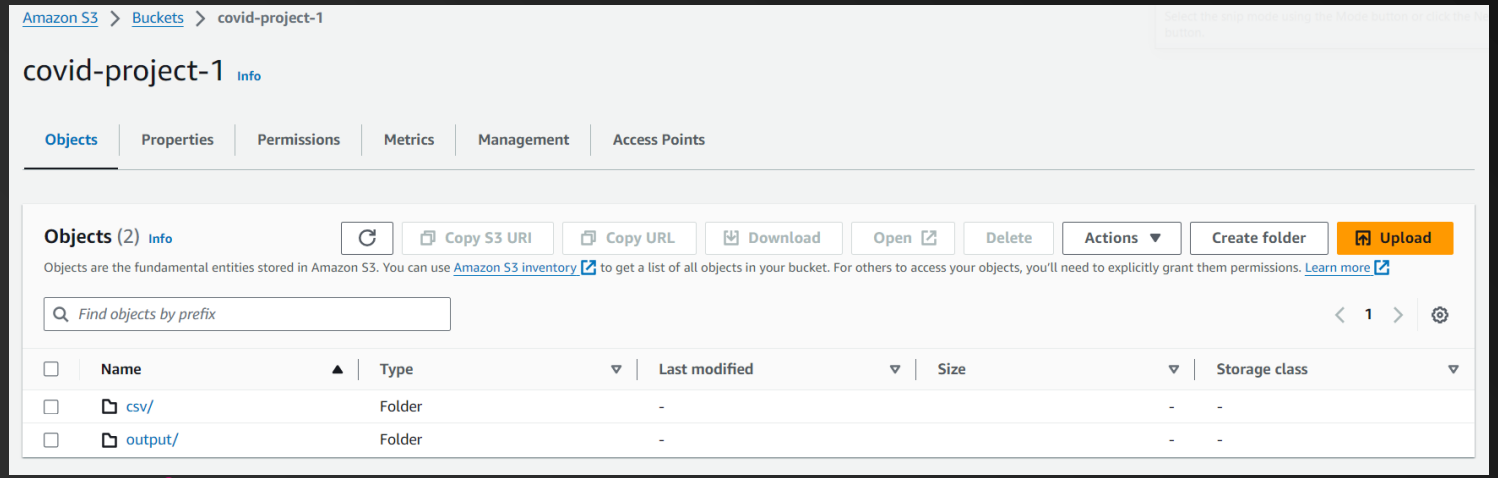
01

DATA STORAGE

Using AWS S3

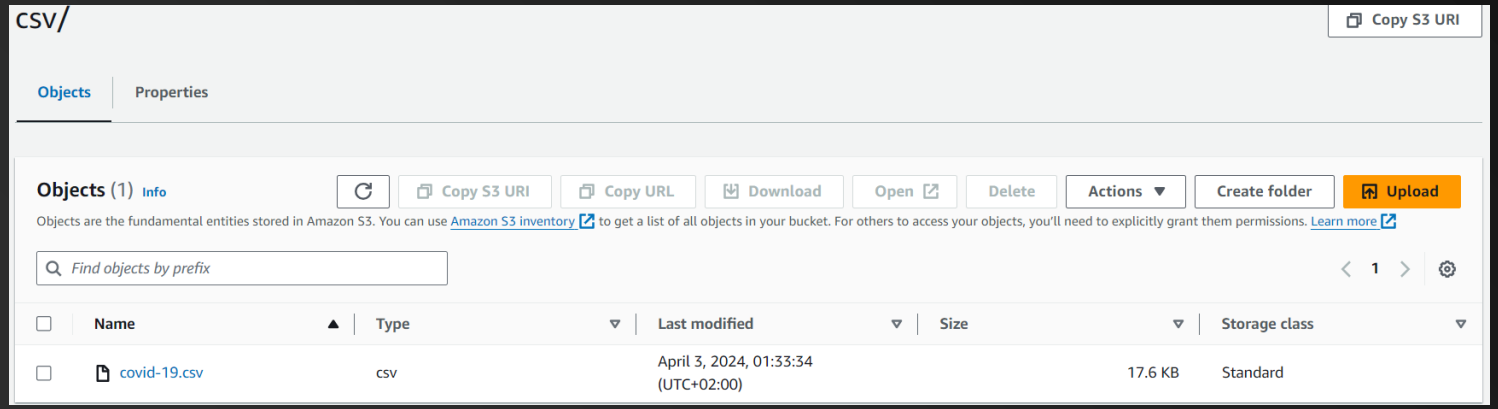
1.1 Creating S3 Bucket and Folder Structure

- Staring 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.



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).



02

DATA INGESTION

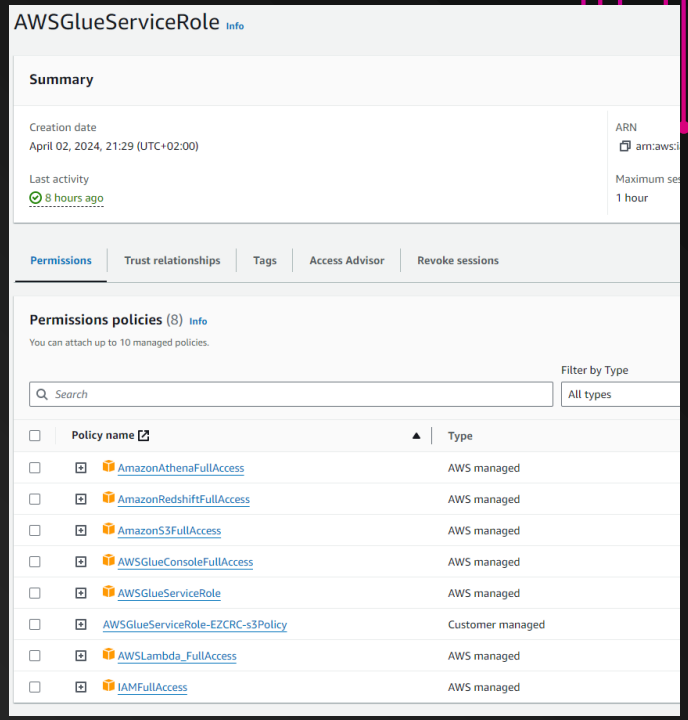
Using AWS GLUE

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.



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:

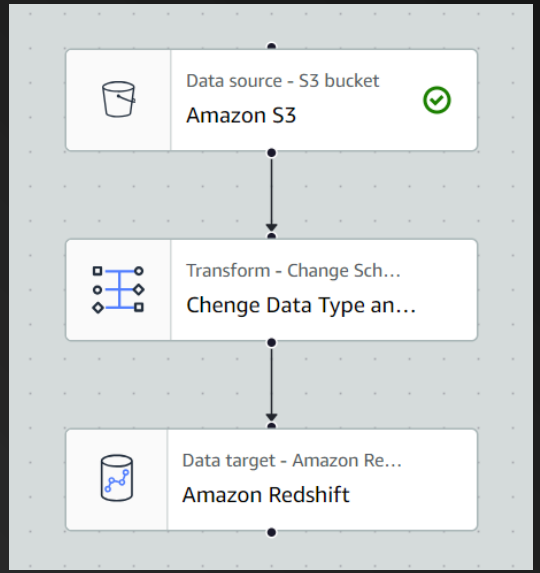
Run | Limit 100 | Explain | Isolated session

```
CREATE TABLE IF NOT EXISTS covid_staging
(
 Country                        VARCHAR,
 Total_Cases                    DOUBLE PRECISION,
 New_Cases                      DOUBLE PRECISION,
 Total_Deaths                   DOUBLE PRECISION,
 New_Deaths                     DOUBLE PRECISION,
 Total_Recovered                DOUBLE PRECISION,
 Active_Cases                   DOUBLE PRECISION,
 Serious                        DOUBLE PRECISION,
 Tot_Cases                      DOUBLE PRECISION,
 Deaths                         DOUBLE PRECISION,
 Total_Tests                    DOUBLE PRECISION,
 Tests                          DOUBLE PRECISION,
 CASES_per_Test                 DOUBLE PRECISION,
 Death_in_Closed_Cases          DOUBLE PRECISION,
 Rank_by_Testing_rate           DOUBLE PRECISION,
 Rank_by_Death_rate             DOUBLE PRECISION,
 Rank_by_Cases_rate             DOUBLE PRECISION,
 Rank_by_Death_of_Closed_Cases  DOUBLE PRECISION
);
```

2.3 Creating Glue ETL Jobs

- Creating Glue two ETL Jobs as follows:

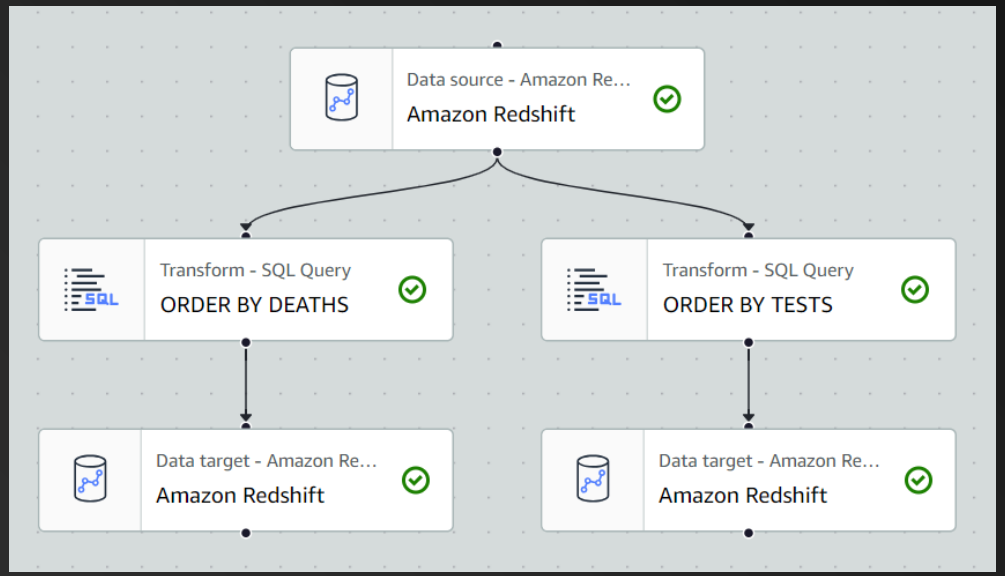
  1. The first one to ingest the data from the S3 bucket /CSV as a data source, change the data types and column names to match the “covid_staging table created, and load the data to Redshift.



2.3 Creating Glue ETL Jobs ..Cont.

- Creating Glue two ETL Jobs as follows:

  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 in 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:

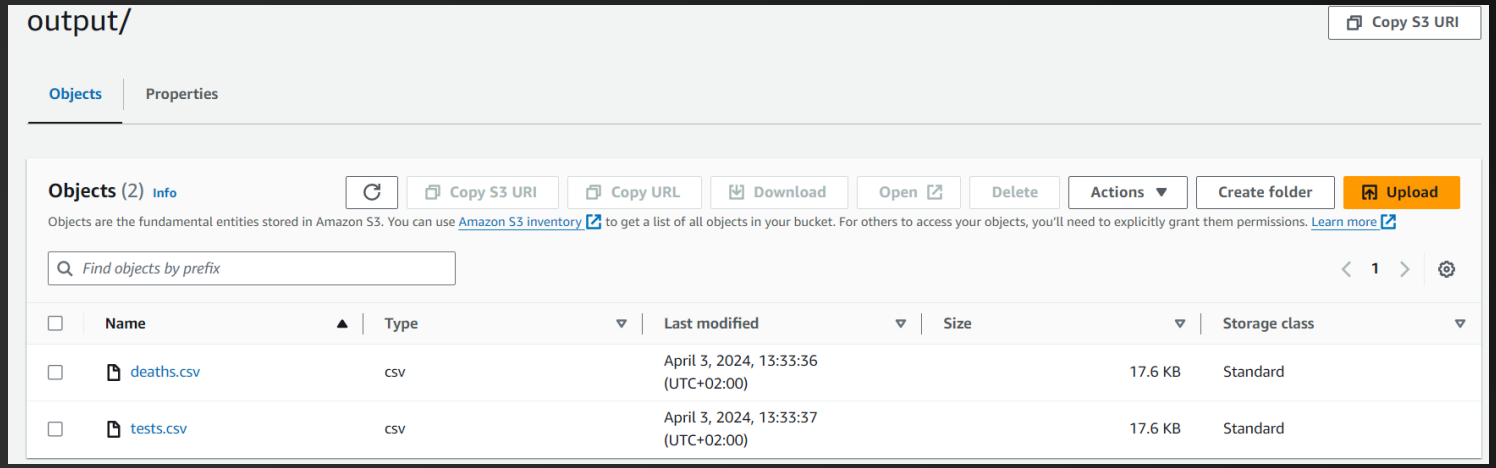
Run | Limit 100 | Explain | Isolated session | Serverless

```
UNLOAD ('SELECT * FROM deaths ORDER BY deaths DESC')
TO 's3://covid-project-1/output/'
IAM_ROLE 'arn:aws:iam::992382770258:role/AWSGlueServiceRole' csv
PARALLEL OFF;

UNLOAD ('SELECT * FROM tests ORDER BY tests DESC')
TO 's3://covid-project-1/output/'
IAM_ROLE 'arn:aws:iam::992382770258:role/AWSGlueServiceRole' csv
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.



04

DATA VISUALIZATION

Using MS POWER PI

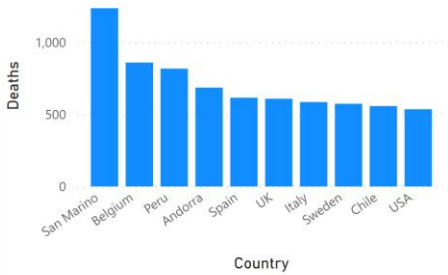
4.1 Creating Power BI Dashboard

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
Total		**7,080.00**

Top 10 Deaths Rate by Country and Country

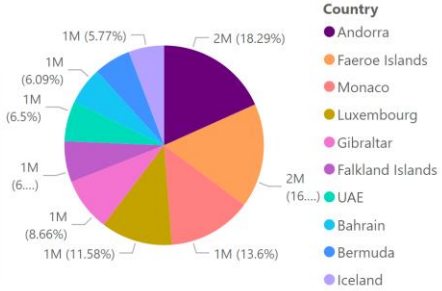
Deaths by Country



#	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
Total		**9725653**

Top 10 Test Rate by Country and Country

Tests by Country



4.2 Project GitHub Link

https://github.com/kirolosgirgis/Data-Engineering-MC/tree/main/05-Kirolos%20Graduation%20Project

Thanks!