**AWS Lake Formation Homework**

This report summarizes the AWS Lake Formation data analysis tasks completed using Lake Formation and COVID-19 database. It includes step-by-step processes, Lake Formation security configuration, code snippets and analytical insights.

**Lake Formation**

**Step-by-step Process and Code Snippets**

1. **AWS Lake Formation CloudFormation Template**

First, create three users: DEOnlineTraining, Covid-Analyst-1, Covid-Analyst-2 with DEOnlineTraining as the administrator and Covid-Analyst-1, Covid-Analyst-2 as data analyst. Acquire the user ARN after creating the users for the lake permission to grant data location permissions in the next step.

Then, Create a CloudFormation template HW19-LF.yaml that:

* Creates an S3 bucket.
* Creates a Glue Data Catalog Database for Athena SQL queries.
* Creates a Glue Crawler that crawls the database in the bucket.
* Creates an Athena WorkGroup with result output to the created S3 bucket.
* Registers the S3 path as a Lake Formation data location.
* Grants Lake Formation permissions:
  + Data location access to the Glue Crawler role
  + Database to designated principals.
* Sets Lake Formation admins and default permissions.
* Grant data location permission to the analyst.

After provisioning the AWS Lake Formation template, deploy it in AWS and wait until all the resources are correctly created.

A screenshot of a computer

AI-generated content may be incorrect.

1. **AWS Lake Formation Data Ingestion**

After the successful deployment of the CloudFormation template, direct to the S3 bucket, created a file name data, and upload the COVID-19 database (country\_wise\_latest.csv, link: https://www.kaggle.com/datasets/imdevskp/corona-virus-report/data) to the created file.

A screenshot of a computer

AI-generated content may be incorrect.

Then, direct to the AWS Glue Console to start the crawler to get the dataset schema. Once the crawler successfully completes the task, the dataset schema can be found in the DataCatalog Tables in the AWS Glue Console.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. **AWS Lake Formation Security Configuration**

In this step, we will grant data permissions to the three users in Lake Formation using Lake Formation Data Filters. In this scenario, “Country/Region” and “WHO Region” are sensitive data columns.

For the Administrator (User: DEOnlineTraining), we grant all permissions to this user to ensure it can access all the rows and columns. This ensures that the administrator can access all rows and all columns without restrictions, including sensitive fields such as “Country/Region” and “WHO Region”.

For the data analysts, we demonstrate fine-grained access control using Data Filters in Lake Formation:

Covid-Analyst-1:

Column-level restriction: Cannot access the “Country/Region”, “Confirmed last week”, “1 week change”, “1 week % increase” column.

Row-level restriction: Data is filtered to allow access only to specific regions.

A screenshot of a computer

AI-generated content may be incorrect.

Covid-Analyst-2

Column-level restriction: Cannot access the “WHO Region” column.

Row-level restriction: Data is filtered to allow access only to specific countries

A screenshot of a computer

AI-generated content may be incorrect.

After applying the filters to set up the security configuration, we will test the availability by trying to query the columns / rows using Athena.

As for the Administrator, the results contain all the columns and rows.

A screenshot of a computer

AI-generated content may be incorrect.

As for Covid-Analyst-1, the “Country/Region” column cannot be accessed. Meanwhile, the WHO Region Americas cannot be found, indicating that the filter works for this analyst.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

As for Covid-Analyst-2, the “WHO Region” column cannot be accessed. Meanwhile, only the specific countries can be accessed, indicating that the filter works for this analyst.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Query Results Example**

For the given Covid-19 datasets, the following analysis will be conducted, and the results will be shown to demonstrate how security settings affect the data analysis. The analysis are conducted in the same Athena WorkGroup.

**The first analysis will be based on the total confirmed cases and total death cases:**

The Administrator can analysis the total confirmed cases and total death cases of each country and the WHO region it belongs to.

**Code Snippet and result:**

select sum("confirmed") as "total\_confirmed\_cases",

sum("deaths") as "total\_death\_cases",

"who region",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region", "who region"

order by "total\_confirmed\_cases" desc

limit 10

A screenshot of a computer

AI-generated content may be incorrect.

Covid-Analyst-1 can only find the total confirmed and death cases in the specific WHO region.

**Code Snippet and Result:**

select sum("confirmed") as "total\_confirmed\_cases", sum("deaths") as "total\_death\_cases", "who region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "who region"

A screenshot of a computer

AI-generated content may be incorrect.

Covid-Analyst-2 can only find the total confirmed and death cases in the specific country/region.

**Code Snippet and Result:**

select sum("confirmed") as "total\_confirmed\_cases",

sum("deaths") as "total\_death\_cases",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "total\_confirmed\_cases" desc

A screenshot of a computer

AI-generated content may be incorrect.

**The second analysis is based on total new cases, including new Covid-19 cases, new death cases, and new recovered cases.**

The Administrator can analysis the total new cases of each country.

**Code Snippet and Result:**

select sum("new cases") as "total\_new\_cases",

sum("new deaths") as "total\_new\_death\_cases",

sum("new recovered") as "total\_new\_recovered\_cases",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "total\_new\_cases" desc

limit 10

A screenshot of a computer

AI-generated content may be incorrect.

Covid-Analyst-1 can only find the total new cases in the specific WHO region.

**Code Snippet and Result:**

select sum("new cases") as "total\_new\_cases",

sum("new deaths") as "total\_new\_death\_cases",

sum("new recovered") as "total\_new\_recovered\_cases",

"who region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "who region"

A screenshot of a computer

AI-generated content may be incorrect.

Covid-Analyst-2 can only find the total new cases in the specific country/region.

**Code Snippet and Result:**

select sum("new cases") as "total\_new\_cases",

sum("new deaths") as "total\_new\_death\_cases",

sum("new recovered") as "total\_new\_recovered\_cases",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "total\_new\_cases" desc

A screenshot of a computer

AI-generated content may be incorrect.

**The third analysis is based on survival rates, which is calculated by dividing the recovered cases by the confirmed cases.**

The Administrator can analysis the survival rate of each country.

**Code Snippet and Result:**

select round(sum(coalesce("recovered", 0.0)) / sum(coalesce("confirmed", 0.0)), 4) as "survival\_rate",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "survival\_rate" desc

limit 10

A screenshot of a computer

AI-generated content may be incorrect.

Covid Analyst-1 can only find the survival rate in the specific WHO region.

**Code Snippet and Result:**

select round(sum(coalesce("recovered", 0.0)) / sum(coalesce("confirmed", 0.0)), 4) as "survival\_rate",

"who region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "who region"

order by survival\_rate desc

A screenshot of a computer

AI-generated content may be incorrect.

Covid-Analyst-2 can only find the survival rate in the specific country/region.

**Code Snippet and Result:**

select round(sum(coalesce("recovered", 0.0)) / sum(coalesce("confirmed", 0.0)), 4) as "survival\_rate",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "survival\_rate" desc

A screenshot of a computer

AI-generated content may be incorrect.

**The fourth analysis is based on death rate and survival rate, which is calculated by averaging the data from death/100 cases and recovered/100 cases columns.**

The Administrator can analysis the recovery and death rate of each country.

**Code Snippet and Result:**

select round(avg(coalesce("Deaths / 100 Cases", 0.0)), 4) as "death\_rate",

round(avg(coalesce("Recovered / 100 Cases", 0.0)), 4) as "recovery\_rate",

"country/region",

"who region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region", "who region"

order by "death\_rate" desc

limit 10

A screenshot of a computer

AI-generated content may be incorrect.

Covid Analyst-1 can only find the death rate and recovery rate in the specific WHO region.

**Code Snippet and Result:**

select round(avg(coalesce("Deaths / 100 Cases", 0.0)), 4) as "death\_rate",

round(avg(coalesce("Recovered / 100 Cases", 0.0)), 4) as "recovery\_rate",

"who region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "who region"

A screenshot of a computer

AI-generated content may be incorrect.

Covid Analyst-2 can only find the death rate and recovery rate in the specific country/region.

**Code Snippet and Result:**

select round(avg(coalesce("Deaths / 100 Cases", 0.0)), 4) as "death\_rate",

round(avg(coalesce("Recovered / 100 Cases", 0.0)), 4) as "recovery\_rate",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "death\_rate" desc

A screenshot of a computer

AI-generated content may be incorrect.

**The fourth analysis is based on increase rate, which the data is selected from the column ”1 week % increase”.**

The Administrator can analysis the survival rate of each country and WHO region.

**Code Snippet and Result:**

select sum("1 week % increase") as "increase\_rate",

"country/region", "who region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region", "who region"

order by "increase\_rate" desc

limit 10

**A screenshot of a computer

AI-generated content may be incorrect.**

Covid Analyst-1 cannot give out results as the access to the column is blocked.

A close-up of a computer screen

AI-generated content may be incorrect.

Covid Analyst-2 can only find the increase rate in the specific country/region.

**Code Snippet and Result:**

select sum("1 week % increase") as "increase\_rate",

"country/region"

from "AwsDataCatalog"."hw19-covid"."data"

group by "country/region"

order by "increase\_rate" desc

A screenshot of a computer

AI-generated content may be incorrect.

1. **How AWS Lake Formation Enhances Data Security**

AWS Lake Formation enhances data security in a data lake environment by introducing a fine-grained access control model on top of traditional S3 and IAM permissions. Lake Formation enables fine-grained access control, which supports column-level and row-level filtering to ensure the least privilege of access is granted to certain users. Lake Formation integrates with AWS Glue Catalog to manage all permissions in one place. In addition, Lake Formation can provide data security by registering the specific S3 bucket location.