Pre-requisitos

1. Para esta sesión va a requerir una cuenta de Amazon Web Services - AWS. Para esto hay varias opciones:

   a) Utilizar una cuenta propia ya creada.

   b) Crear una cuenta nueva. En este caso recuerde que requiere ingresar datos de una tarjeta de crédito. Usaremos recursos disponibles en la capa gratuita (https://aws.amazon.com/free/), pero es posible que se generen algunos costos menores.

   c) **Recomendado:** usar la cuenta de **AWS Academy** enviada a su correo por el instructor. En este caso emplee el **Learner Lab**.

2. Ingrese a su cuenta y familiarícese con la consola.

3. Asegúrese de que en la esquina superior derecha aparezca la región *N. Virginia*.

4. **Nota:** defina un nombre para su equipo, defínalo claramente en **reporte** y use este nombre como parte inicial de **todos los recursos que cree**.

5. **Nota 2:** los dos miembros del equipo deben realizar todos los pasos de taller e incluir los soportes solicitados.

6. **Nota 3:** la entrega de este taller consiste en un **reporte** y unos **archivos de soporte**. Cree el archivo de su **reporte** como un documento de texto en el que pueda fácilmente incorporar capturas de pantalla, textos y similares. Su reporte debe estar en formato PDF. **Las capturas en AWS deben mostrar claramente que se trata de su cuenta.**

1. Descargue los datos que utilizaremos como fuente

1. Para este taller usaremos datos de Covid publicados por la OMS.

2. Visite la página `https://data.who.int/dashboards/covid19/data`. Navegue a la sección *COVID-19 downloadable statistical releases*.

3. Allí encontrará los datos *Daily frequency reporting of new COVID-19 cases and deaths by date reported to WHO*, los cuales puede descargar en formato CSV.

4. En la página encontrará una descripción breve de los datos. Provea una descripción de los campos de la tabla en su **reporte**.

5. Note que los códigos de los países están en formato ISO Alpha-2.

6. Para mapear estos códigos de país a otros códigos, y obtener otra información de cada país usaremos la tabla en formato CSV en el siguiente enlace `https://gist.github.com/tadast/8827699#file-countries_codes_and_coordinates-csv`.

7. Para descargar la tabla en formato CSV, de click en `Raw` y luego click derecho y seleccione `Guardar como`. El formato por defecto debe ser CSV.

8. Abra estos dos archivos en Excel o en un editor de texto. En su **reporte**, describa los campos de cada uno de los archivos (a manera de un diccionario de datos) y en el archivo de WHO identifique la última fecha de registro.

2. Configure buckets en S3

1. En la consola de Amazon AWS, vaya al servicio S3.

2. Cree un bucket que servirá para los datos de entrada de su Data lake. Como referencia aquí lo llamaremos bucket de entrada.

3. En este bucket cree una carpeta en la raiz del bucket para los datos de covid, puede llamarse `covid`. Suba allí el archivo `WHO-COVID-19-global-table-data.csv`.

4. Cree otra carpeta en la raíz del bucket para los datos de codigos, puede llamarse `paises`. Suba allí el archivo `countries_codes_and_coordinates.csv`.

5. Cree otro bucket para almacenar los resultados de las consultas. Como referencia aquí lo llamaremos bucket de salida.

6. Tome un pantallazo de sus buckets e inclúyalo en su **reporte**. Su usario de AWS debe ser visible en el pantallazo.

3. Conecte los datos con Athena usando Glue

1. En la consola de Amazon AWS, vaya al servicio Athena.

2. Athena es un servicio que permite consultar datos almacenamos en S3 usando el lenguaje SQL.

3. En el panel izquierdo seleccione `Workgroups`.

4. Click en `Create workgroup`. Como nombre del Workgroup use su (primer) nombre (sin repetir existentes).

5. Como motor de analítica (analytics engine) seleccione Athena SQL. Note que la alternativa es Apache Spark.

6. En la sección `Query result configuration`, diligencie el primer campo `Location of the query result` usando el botón `Browse S3` y seleccione el bucket de salida que creó.

7. Deje los demás campos en sus valores por defecto. Click en `Create workgroup` en la parte inferior.

8. De regreso en la consola de Athena, en el panel izquierdo seleccione el ítem `Data Sources`.

9. Click en `Create Data Source` para crear una nueva fuente de datos para consultar.

10. En su **reporte** incluya un pantallazo con algunas de las opciones de fuentes de datos que puede incluir.

11. Seleccione `S3 - AWS Glue Data catalog` como fuente de datos. Click en `Next`.

12. Note que no solamente se selecciona S3 como fuente de los datos crudos, sino que también se selecciona Glue como el servicio para crear un catálogo de datos sobre los datos crudos.

13. Glue apoya los procesos de ETL y ELT (extracción, transformación, carga) al construir y mantener un catálogo de datos inicialmente crudos (en formato CSV). Este catálogo puede ser luego usado por Athena para realizar las consultas usando SQL sobre los datos en formato CSV.

14. Seleccione `AWS Glue Data catalog in this account` ya que el catálogo se construirá en la misma cuenta.

15. En el método para crear tablas seleccione `Create a crawler in AWS Glue`.

16. El Crawler se encargará de recorrer los datos e identificar los campos, tipos, tamaño, etc, para construir el catálogo de datos. Note que también se puede crear una tabla manualmente. En este caso lo haremos con el Crawler.

17. Click en `Create in AWS Glue`.

18. Esto lo llevará a la consola de AWS Glue en una nueva pestaña del navegador.

19. En el panel izquierdo seleccione `Crawlers` (Rastreadores).

20. Click en `Create crawler`.

21. Ingrese un nombre para el crawler, use su primer nombre-covid. Click en `Siguiente`.

22. Selecciones `Not yet` para indicar que los datos no han sido mapeados a tablas de Glue.

23. Click en `Add a data source`. Como fuente seleccione S3, con la opción `in this account` pues es un bucket S3 en la misma cuenta.

24. Use el botón `Browse S3` para seleccionar el bucket de entrada y la carpeta covid. Click en Choose.

25. Mantenga las demás opciones en sus valores por defecto y click en `Add an S3 data source`.

26. Click en Next.

27. En `Choose an IAM role/Elija un rol de IAM`, seleccione el `LabRole`.
28. Click en `Next/Siguiente`.
29. En `Target database` seleccione `Add database/Añadir una base de datos` para crear una nueva base de datos (se abre una nueva pestaña). Defina un nombre, usando su primer nombre-covid-bd, sin tildes, deje los demás campos vacíos y click en `Create database/Crear`.
30. Regrese a la configuración del Crawler, refreque la lista de bases de datos y sleccione la que acaba de crear.
31. En `Crawler schedule - Frequency` seleccione `On demand`. En su **reporte** incluya algunas de las demás opciones (en un pantallazo). Click en `Next/Siguiente`.
32. Revise la configuración del crawler, tome un pantallazo, inclúyalo en su **reporte**, y click en `Create Crawler`.
33. Repita los pasos anteriores para agregar un nuevo crawler para el archivo de países. Como nombre utilice `sunombre-paises`, sin tilde. Para el **output**, seleccione **la misma base de datos** del primer crawler, i.e., `sunombre-covid-db`.
34. Una vez tenga los dos crawlers creados, vuelva a la consola de AWS Glue, seleccione cada uno y click en `Run Crawler` para ejecutarlos. Esta operación tardará un poco.
35. Una vez los crawlers hayan terminado de correr podrá ver las tablas obtenidas en el ítem `Tables` del panel izquierdo.
36. Tome un pantallazo de los *esquemas* de las tablas creadas e inclúyalo en su **reporte**.
37. Note que las tablas creadas tienen los nombres de las carpetas que creó en S3, mientras la base de datos tiene el nombre especificado en el Crawler.

4. Consulte los datos usando SQL desde Athena

1. Regrese a la consola de Athena.
2. En el `Editor query` seleccione su Workgroup (con el nombre definido arriba), como Data source un `AwsDataCatalog` y como Database la base de datos creada `sunombre-covid-db`.
3. En el panel de edición de consultas escriba consultas (**una por una**) como

```
SELECT * FROM "covid" limit 10;


SELECT * FROM "paises" limit 10;
```

```
SELECT * FROM "covid" where country='Colombia';
```

```
SELECT * FROM "paises" where country='"Colombia"';
```

y ejecútelas con `Run`.

4. A cada una de las consultas anteriores sobre la tabla covid agregue una instrucción que permita ordenarlas de acuerdo con la fecha, del más reciente al menos reciente. Ejecute nuevamente, revise los resultados y guarde un pantallazo de cada en su **reporte**.

5. En la tabla **paises** los nombres de los campos son un poco largos. Para modificarlos, vaya a AWS Glue, seleccione la tabla (click en el nombre de la tabla) y en `Actions` click en `Edit schema`. Allí sobre cada nombre de campo de doble click y haga los cambios apropiados. Por ejemplo quite «code» de «alpha-2 code» o «(average)» de «latitude (average)». Click en `Save as new table version`.

6. Note que en la tabla **paises**, los campos vienen con comillas dobles. Para quitarlos e interpretarlos como parte del formato, vaya a AWS Glue, seleccione la tabla y en `Actions` click en `Edit table`.

7. Cambie el Serde serialization lib por `org.apache.hadoop.hive.serde2.OpenCSVSerde`. En los campos `Serde parameters` incluya la llave `escapeChar` con valor \, y la llave `quoteChar` con valor “.

8. Click en `Apply/Save`.

9. En Athena actualice los datos y vuelva a ejecutar la consulta

```
SELECT * FROM "paises" limit 10;
```

¿Qué cambios observa? Incluya la respuesta en su **reporte**.

10. Note que ahora puede ejecutar la consulta con una ligera modificación

```
SELECT * FROM "paises" where country='Colombia';
```

11. Realice un join entre las dos tablas con el siguiente comando

```
SELECT * FROM "covid" AS cvd JOIN "paises" AS ps
ON cvd.country=ps.country
LIMIT 5;
```

¿Qué hace esta consulta? Incluya la respuesta en su **reporte**. Incluya un snapshot de soporte.